Linguistic Knowledge in Data-Driven Natural Language Processing

Yulia Tsvetkov

CMU-LTI-16-017

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213

Thesis Committee:

Chris Dyer (chair), Carnegie Mellon University
Alan Black, Carnegie Mellon University
Noah Smith, Carnegie Mellon University
Jacob Eisenstein, Georgia Institute of Technology

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Language and Information Technologies

to my parents

Abstract

The central goal of this thesis is to bridge the divide between theoretical linguistics—the scientific inquiry of language—and applied data-driven statistical language processing, to provide deeper insight into data and to build more powerful, robust models. To corroborate the practical importance of the synergy between linguistics and NLP, I present model-based approaches that incorporate linguistic knowledge in novel ways and improve over strong linguistically-uninformed statistical baselines. In the first part of the thesis, I show how linguistic knowledge comes to the rescue in processing languages which lack large data resources. I introduce two new approaches to cross-lingual knowledge transfer from resource-rich to resource-constrained, typologically diverse languages: (i) morpho-phonological knowledge transfer that models the historical process of lexical borrowing between languages, and demonstrates its utility in improving machine translation, and (ii) semantic transfer that allows to identify metaphors—ultimately, a language- and culture-salient phenomenon—across languages. In the second part, I argue that integrating explicit linguistic knowledge and guiding models towards linguistically-informed generalizations help improve learning also in resource-rich conditions. I present first steps towards characterizing and integrating linguistic knowledge in neural NLP (i) through structuring training data using knowledge from language acquisition and thereby learning the curriculum for better task-specific distributed representations, (ii) for using linguistic typology in training hybrid linguistically-informed multilingual deep learning models, and (iii) in leveraging linguistic knowledge for evaluation and interpretation of learned distributed representations. The scientific contributions of this thesis include a range of answers to new research questions and new statistical models; the practical contributions are new tools and data resources, and several quantitatively and qualitatively improved NLP applications.

Acknowledgement

I am an incredibly lucky person. At the most important landmarks in my life I met remarkable people, many of whom became my lifetime friends. Meeting my advisor Chris Dyer was one of such pivotal events. Chris, you were an amazing advisor in every possible way. Somehow, you always managed to provide what I needed most at that specific moment. Whether it was a big-picture perspective, or help with a particular technical problem, or emotional support—you were always there for me at the right time with the most pertinent help, provided in the most genuine way. I am constantly inspired by how bright and knowledgeable you are, equally deeply and broadly, by your insights and vision, your genuine excitement about research, and also by your incredible kindness and openness. Chris, I believe that you are among the few brilliant people of our time that were born to change the world, and I am extremely grateful that our paths crossed.

Shuly Wintner, I am worried that nothing I write here can express how grateful I am to you. Thank you for your endless support and for believing in me more than I did. Thank you for your guidance, your friendship, and for showing me what passion for research and for languages is. I am forever indebted to you.

Thank you to my thesis committee members Alan Black, Jacob Eisenstein, and Noah Smith. Your advice, thought provoking questions, and support were indispensable. Noah, I have benefitted immensely from our short interactions, even just from watching how you manage your collaborations, how you teach, how you write. For me you are a role model researcher and advisor. Thank you for sharing your wisdom and experience.

Thank you to Wang Ling, Manaal Faruqui, Nathan Schneider, Archna Bhatia, my friends and paper buddies who made these years so colorful, looking forward to drinking with you at the next conference! Thank you to the clab members, collaborators, friends, and colleagues—Waleed, Guillaume, Sunayana, Dani, Austin, Kartik, Swabha, Anatole, Lori, Alon, Leo, Florian, David, Brian—I learned a lot from each of you. And a big thanks to Jaime Carbonell for building such an incredible institute.

Most especially, my parents, thank you for your unconditional love and friendship; Eden and Gahl thank you for making my life worthy to live, and Igor—my biggest friend and ally, my grounding and wings, I will always love you.

Contents

Chapter 1

Introduction

The first few decades of research in computational linguistics were dedicated to the development of computational (and mathematical) models of various levels of linguistic representation (phonology, morphology, syntax, semantics, and pragmatics). The main goal of such investigations was to obtain a sufficiently formal, accurate and effective model of linguistic knowledge, that will facilitate the construction of natural language processing (NLP) applications. Unfortunately, by the early 1990s it became clear that such approaches, commonly known as *logic-* and *rule-based*, hit a glass ceiling. While these linguistically-underpinned models were extremely expressive, and could be used to drive toy example applications, they did not, in general, scale up to real-world systems. This became strikingly evident for machine translation already in the 1950s and 1960s (Bar-Hillel, 1960), and for other, less ambitious applications such as parsing, later on.

The 1990s saw a complete paradigm shift (Hall *et al.*, 2008; Church, 2007). Driven in part by the failure of rule-based systems to scale up, and in part by the increasing availability of massive quantities of data, *data-driven* approaches emerged. They replaced the reliance on formalized linguistic knowledge by the extraction of *implicit* linguistic patterns from textual (and, later, also spoken) data. Statistical methods grounded in this paradigm turned out to be extremely powerful, and significantly reduced the amount of manual knowledge engineering. In machine translation, the statistical approach (Brown *et al.*, 1990) has become the de-facto standard, with revolutionary achievements. Other application areas have also seen great improvements in accuracy, but more importantly, robustness in the face of real-world data.

In the mid 2010s, two and half decades later, the (now traditional) statistical methods are superseded by statistical *neural* approaches in the deep learning paradigm shift (Manning, 2016). Again, this shift is happening because the new approaches are even more powerful than the old ones, and are easier to engineer. These developments, however, come with a price. While contemporary statistical approaches are efficient and accurate (with sufficient amounts of training data), and can support a vast array of NLP applications that are more and more ubiquitous, the original motivation driving research in computational linguistics, namely the formalization of linguistic knowledge, was lost. Many data-driven statistical models are opaque: they represent implicit patterns of language use, but these patterns are invisible. In neural network approaches, virtually all building blocks of NLP models are replaced by continuous dense representations: words and phrases are replaced by their real-valued vector representations (Mikolov *et al.*, 2013b; Pennington *et al.*, 2014); linguistic features can be replaced by unsupervisedly learned word vectors in most NLP tasks, from core NLP (Collobert *et al.*, 2011; Turian *et al.*, 2010a) to end-user applications (Cho *et al.*, 2014; Sutskever *et al.*, 2014; Bordes *et al.*, 2014; Rush *et al.*, 2015); explicit *n*-grams are replaced by distributed representations of longer contexts (Bengio *et al.*, 2003; Mikolov *et al.*, 2010); lexicalized concepts and relations are now encoded in a hierarchy of abstract representations in a neural network. The new building blocks of the data-driven NLP are linguistically opaque, and it is not clear (yet) how to characterize (or interpret) linguistic knowledge learned by neural models or how to build hybrid models, integrating explicit linguistic knowledge into neural architectures.

The motivation driving this thesis work is a realization—which is self-evident to some researchers (e.g., Wintner, 2009; Manning, 2016), but questionable to many others—that despite the growing divide between Linguistics and NLP in the era of big data, Linguistics, as the scientific inquiry of language, has yet much to offer to data-driven language processing, and, in fact, is essential for building intelligent, language-aware applications. While hand-crafted rules alone are not sufficient to model the full complexity of linguistic knowledge, what Linguistics has to offer to contemporary data driven approaches is (i) understanding what *research questions* need to be asked for a meaningful automation of languages, (ii) a better-informed *construction and exploration of data*, and (iii) a thorough analysis and understanding of *linguistic generalization principles* to guide statistical learning towards automatically constructing better generalizations with less data.

There is an independent value to the formal and computational representation and modeling of linguistic knowledge that is worth retaining (we should not shut this branch of science just because we failed to study it!). Furthermore, despite recent advances, current models are still insufficient for building robust applications. Purely data-driven models underperform in low-resource settings: they are inadequate, for example, to translate African languages, to detect metaphors in Russian and Persian, to grammatically parse Cantonese, to model Latin or Hebrew morphology, to build dialog systems for indigenous languages. The same models produce generalizations that are sensitive to noise, domain variations, and frequency effects even in resource-rich scenarios. It turns out that linguistic modeling is beneficial for such NLP applications, even in an age of primarily statistical approaches. Several works demonstrate that *hybrid* methods that can combine linguistic knowledge with otherwise data-driven approaches in sufficiently sophisticated ways yield better, more accurate and yet scalable systems.

One of the main vehicles of linguistic investigation has to do with the notion of *generalization*. For example, the notion of *part of speech* generalizes the various syntactic contexts that a word is expected to occur in. Parts of speech (or syntactic *categories*) are very rough approximations. A better generalization would associate a verb, for example, not only with the category 'V' but also with a sub-catgory that represents the number and types of arguments that the verb is expected to occur with. An even finer generalization will take into account not just the syntactic category of some verb's arguments, but also facets of their semantics, in the form of *selectional restrictions*.

Statistical models also provide generalization, induced from large data sets. For example, word embeddings associate words with a numerical representation of the context this word is expected to occur in. However, these generalizations are different from the ones common in linguistics. The underlying assumption of my research is that augmentation of statistical generalization, such as the ones created by word embeddings, with linguistically-motivated generalizations of the form exemplified above, can result in a significant improvement in the quality of NLP applications that are built on top of these models.

Beyond generalization, linguistic knowledge can be extremely useful to alleviate the problem of missing or skewed data. Although we are living in the era of big data when from the web it is easy to obtain billions or trillions of words, there are several scenarios in which we cannot be too data hungry. First, the vast majority of the world's languages barely exist on the web at all. Second, even in a billion-word corpus, there is a long tail of rare and out-of-vocabulary words (Zipf, 1949). Finally, language is not always paired with correlated events: corpora contain what people said, but not what they meant, or how they understood things, or what they did in response to the language. To construct robust language understanding applications, explicit linguistic- and world-knowledge must be used to augment existing corpora and models with implied knowledge that is not expressed in training texts. The goal of this thesis is thus to investigate how linguistic knowledge can be operationalized in contemporary statistical approaches to overcome the data- and generalization-bottleneck.

1.1 Thesis Statement

The central goal of this thesis is to bridge the divide between theoretical linguistics—the scientific inquiry of language—and applied data-driven statistical language processing, to provide deeper insight into data and to build more powerful, robust models. To corroborate the practical importance of the synergy between linguistics and NLP, I present model-based approaches that incorporate linguistic knowledge in novel ways and improve over strong linguistically-uninformed statistical baselines. In the first part of the thesis, I show how linguistic knowledge comes to the rescue in processing languages which lack large data resources. I introduce two new approaches to cross-lingual knowledge transfer from resource-rich to resource-constrained, typologically diverse languages: (i) morpho-phonological knowledge transfer that models the historical process of lexical borrowing between languages, and demonstrates its utility in improving machine translation, and (ii) semantic transfer that allows to identify metaphors—ultimately, a language- and culture-salient phenomenon—across languages. In the second part, I argue that integrating explicit linguistic knowledge and guiding models towards linguistically-informed generalizations help improve learning also in resource-rich conditions. I present first steps towards characterizing and integrating linguistic knowledge in neural NLP (i) through structuring training data using knowledge from language acquisition and thereby learning the curriculum for better task-specific distributed representations, (ii) for using linguistic typology in training hybrid linguistically-informed multilingual deep learning models, and (iii) in leveraging linguistic knowledge for evaluation and interpretation of learned distributed representations. The scientific contributions of this thesis include a range of answers to new research questions and new statistical models; the practical contributions are new tools and data resources, and several quantitatively and qualitatively improved NLP applications.

1.2 Thesis Overview

This thesis is a collection of diverse case studies with a unifying goal of integrating explicit, rich linguistic knowledge into machine learning methods. It is presented in two parts. The first part, comprising the next two chapters, focuses on computational implementation of linguistic theories and their integration in cross-lingual statistical models that are used to bridge between resource-rich and resource-constrained languages. The second part, comprising Chapter 4 through Chapter 6, discusses characterizing and integrating linguistic knowledge in neural models in resource-rich conditions. Due to diversity of research questions, linguistic representations, and methodologies, the chapters are self-contained: each chapter includes its own background and motivation, methodology, experiments, and prior work. I conclude with a summary of contributions and a detailed discussion of future work, that considers more practical applications of linguistic knowledge, as well as how statistical models of languages could benefit linguistics. The chapters are organized as follows:

Part I: Linguistic Knowledge Transfer for Resource-Constrained NLP

Chapter 2 describes a new approach to cross-lingual knowledge transfer that models the historical process of lexical borrowing between languages. The chapter first presents a morpho-phonological model with features based on universal constraints from Optimality Theory (OT), and it shows that compared to several standard—but linguistically more naïve—baselines, the OT-inspired model obtains good performance at predicting donor forms from borrowed

forms with only a few dozen training examples, making this a cost-effective strategy for sharing lexical information across languages. Then, this chapter demonstrates applications of the lexical borrowing model in machine translation, using resource-rich donor language to obtain translations of out-of-vocabulary loanwords in a lower resource language.
Linguistic representation: Phonology and morphology.
Linguistic platform: Optimality Theory (Prince and Smolensky, 2008).
NLP applications: Lexical borrowing, machine translation.
Relevant publications: *JAIR* (Tsvetkov and Dyer, 2016), *NAACL 2015* (Tsvetkov *et al.*, 2015a), *ACL 2015* (Tsvetkov and Dyer, 2015).

Chapter 3 shows how the theory of conceptual metaphor (TCM) can be operationalized to transfer semantic knowledge and to identify metaphors—ultimately, a language- and culture-salient phenomenon—across languages. The chapter devises a set of conceptual semantic features inspired by TCM, such as a degree of abstractness and semantic supersenses. These features survive in translation, which ensures a shared feature representation across languages and thus enables model transfer. A metaphor detection model is constructed using English resources, and it obtains state-of-the-art performance relative to previous work in this language. Using a model transfer approach by pivoting through a bilingual dictionary, the model can identify metaphoric expressions in other languages. It obtains comparable results when tested on Spanish, Farsi, and Russian, supporting the hypothesis that metaphor is conceptual, rather than lexical, in nature. A secondary contribution of this chapter is a collection of publicly released resources: metaphor datasets in English and Russian, and an automatically constructed supersense taxonomy for English adjectives.
Linguistic representation: Semantics.
Linguistic platform: Conceptual Metaphor Theory (Lakoff and Johnson, 1980).
NLP applications: Metaphor detection.
Relevant publications: *ACL 2014* (Tsvetkov *et al.*, 2014c).

Part II: Linguistic Knowledge in Neural NLP

Chapter 4 begins a sequence of three chapters on characterizing and integrating knowledge in distributed representations and neural network models. The chapter argues that insight into linguistic coherence, prototypicality, simplicity, and diversity of data helps improve learning. It presents a novel method that optimizes—using Bayesian optimization—linguistic content and structure of training data to find a better curriculum for learning distributed representations of words. The chapter shows that learning the curriculum improves performance on a variety of downstream tasks over random orders and in comparison to the natural corpus order. In addition, the analysis of learned curricula sheds interesting light on desired properties

of training data for different tasks.
Linguistic representation: Syntax and semantics.
Linguistic platform: Prototype Theory (Rosch, 1978); syntactic annotations.
NLP applications: Part-of-speech tagging, parsing, named entity recognition, sentiment analysis.
Relevant publications: *ACL 2016* (Tsvetkov *et al.*, 2016b).

Chapter 5 introduces "polyglot" language models, recurrent neural network models trained to predict symbol sequences in many different languages using shared representations of symbols and conditioning on typological information about the language to be predicted. These are applied to the problem of modeling phone sequences—a domain in which universal symbol inventories and cross-linguistically shared feature representations are a natural fit. Intrinsic evaluation on held-out perplexity, qualitative analysis of the learned representations, and extrinsic evaluation in two downstream applications that make use of phonetic features show that polyglot models better generalize to held-out data than comparable monolingual models, and that polyglot phonetic feature representations are of higher quality than those learned monolingually.
Linguistic representation: Phonetics and phonology.
Linguistic platform: Language universals and linguistic typology.
NLP applications: Lexical borrowing, speech synthesis.
Relevant publications: *NAACL 2016* (Tsvetkov *et al.*, 2016c).

Chapter 6 presents QVEC—a computationally inexpensive intrinsic evaluation measure of the quality of word embeddings based on alignment to a matrix of features extracted from manually crafted lexical resources—that obtains strong correlation with performance of the vectors in a battery of downstream semantic evaluation tasks. In addition, its computation induces an alignment of vector dimensions to linguistic concepts. This facilitates qualitative evaluation of dimensions in vector space representations.
Linguistic representation: Syntax and semantics.
Linguistic platform: Coarse semantic and syntactic annotations.
NLP applications: Metaphor detection, text classification, sentiment analysis, part-of-speech tagging, parsing.
Relevant publications: *EMNLP 2015* (Tsvetkov *et al.*, 2015b), *RepEval 2016* (Tsvetkov *et al.*, 2016a).

Part I

Linguistic Knowledge in Resource-Constrained NLP: Cross-Lingual Knowledge Transfer from Resource-Rich Languages

Chapter 2

Morpho-Phonological Knowledge Transfer

This chapter shows how to overcome a severe lack of training data in resource-poor languages by effectively operationalizing linguistic research and linguistic analyses of just a few dozen examples. The chapter addresses modeling lexical borrowing, the process by which a recipient language incorporates lexical material from a donor language. This is a ubiquitous phenomenon in the world's languages, and many languages have lexicons consisting of large numbers of borrowed words. The chapter shows how to incorporate borrowed lexical material as another source of evidence for cross-lingual lexical correspondence and use this in machine translation, focusing on both improving medium-resource performance and bootstrapping applications in languages that lack parallel data. Research described in this chapter was conducted in collaboration with Chris Dyer and in part with Waleed Ammar. It is based on the *NAACL 2015* publication (Tsvetkov *et al.*, 2015a), the *ACL 2015* publication (Tsvetkov and Dyer, 2015), and an extension of these papers in the *JAIR* article (Tsvetkov and Dyer, 2016).

2.1 Background and Motivation

State-of-the-art natural language processing (NLP) tools, such as text parsing, speech recognition and synthesis, text and speech translation, semantic analysis and inference, rely on availability of language-specific data resources that exist only for a few resource-rich languages. To make NLP tools available in more languages, techniques have been developed for projecting such resources from resource-rich languages using parallel (translated) data as a bridge for cross-lingual part-of-speech tagging (Yarowsky *et al.*, 2001; Das and Petrov, 2011; Li *et al.*, 2012; Täckström *et al.*, 2013), syntactic parsing (Wu, 1997; Kuhn, 2004; Smith and Smith, 2004; Hwa *et al.*, 2005; Xi and Hwa, 2005; Burkett and Klein, 2008; Snyder *et al.*, 2009;

Ganchev *et al.*, 2009; Tiedemann, 2014), word sense tagging (Diab and Resnik, 2002), semantic role labeling (Padó and Lapata, 2009; Kozhevnikov and Titov, 2013), metaphor identification (Tsvetkov *et al.*, 2014c), and others. The limiting reagent in these methods is parallel data. While small parallel corpora do exist for many languages (Smith *et al.*, 2013), suitably large parallel corpora are expensive, and these typically exist only for English and a few other geopolitically or economically important language pairs. Furthermore, while English is a high-resource language, it is linguistically a typological outlier in a number of respects (e.g., relatively simple morphology, complex system of verbal auxiliaries, large lexicon, etc.), and the assumption of construction-level parallelism that projection techniques depend on is thus questionable. Given this state of affairs, there is an urgent need for methods for establishing lexical links across languages that do not rely on large-scale parallel corpora. Without new strategies, most of the 7,000+ languages in the world—many with millions of speakers—will remain resource-poor from the standpoint of NLP.

We advocate a novel approach to automatically constructing language-specific resources, even in languages with no resources other than raw text corpora. Our main motivation is research in **linguistic borrowing**—the phenomenon of transferring linguistic constructions (lexical, phonological, morphological, and syntactic) from a "donor" language to a "recipient" language as a result of contacts between communities speaking different languages (Thomason and Kaufman, 2001). Borrowed words (also called loanwords, e.g., in figure 2.1) are lexical items adopted from another language and integrated (nativized) in the recipient language. Borrowing occurs typically on the part of minority language speakers, from the language of wider communication into the minority language (Sankoff, 2002); that is one reason why donor languages often bridge between resource-rich and resource-limited languages. Borrowing is a distinctive and pervasive phenomenon: *all* languages borrowed from other languages at some point in their lifetime, and borrowed words constitute a large fraction (10–70%) of most language lexicons (Haspelmath, 2009).

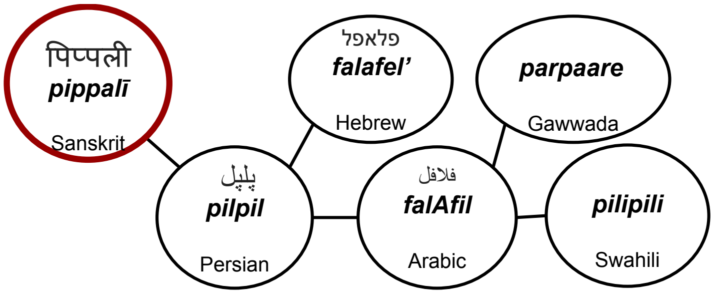


Figure 2.1: An example of the multilingual borrowing from Sanskrit into typologically diverse, low- and high-resource languages (Haspelmath and Tadmor, 2009).

Loanword nativization is primarily a phonological process. Donor words undergo phonological repairs to adapt a foreign word to the segmental, phonotactic, suprasegmental and morpho-phonological constraints of the recipient language (Holden, 1976; Van Coetsem, 1988; Ahn and Iverson, 2004; Kawahara, 2008; Hock and Joseph, 2009; Calabrese and Wetzels, 2009; Kang, 2011, *inter alia*). Common phonological repair strategies include feature/phoneme epenthesis, elision, degemination, and assimilation. When speakers encounter a foreign word (either a lemma or an inflected form), they analyze it morphologically as a stem, and morphological loanword integration thus amounts to selecting an appropriate donor surface form (out of existing inflections of the same lemma), and applying the recipient language morphology (Repetti, 2006). Adapted loanwords can freely undergo recipient language inflectional and derivational processes. Nouns are borrowed preferentially, then other parts of speech, then affixes, inflections, and phonemes (Whitney, 1881; Moravcsik, 1978; Myers-Scotton, 2002, p. 240).

Although borrowing is pervasive and a topic of enduring interest for historical and theoretical linguists (Haugen, 1950; Weinreich, 1979), only limited work in computational modeling has addressed this phenomenon. However, it is a topic well-suited to computational models (e.g., the systematic phonological changes that occur during borrowing can be modeled using established computational primitives such as finite state transducers), and models of borrowing have useful applications. Our work can be summarized as the development of a computational model of lexical borrowing and an exploration of its applications to augment language resources and computational approaches to NLP in resource-limited languages. Specifically, we demonstrate how multilingual dictionaries extracted using models of borrowing improve resource-limited statistical machine translation (MT), using a pivoting paradigm where the borrowing pair and the translation pair have only a single language in common.

The problem we address is the identification of plausible donor words (in the donor language) given a loanword (in the recipient language), and vice versa. For example, given a Swahili loanword *safari* 'journey', our model identifies its Arabic donor سفريه (*sfryh*)[1] 'journey' (§2.2). Although at a high level, this is an instance of the well-known problem of modeling string transductions, our interest is being able to identify correspondences across languages with minimal supervision, so as to make the technique applicable in low-resource settings. To reduce the supervision burden, our model includes awareness of the morpho-phonological repair strategies that native speakers of a language subconsciously employ to adapt a loanword to phonological constraints of the recipient language (§2.2.3). To this end, we use constraint-based theories of phonology, as exemplified by Optimality Theory (OT) (Prince and Smolensky, 2008; McCarthy, 2009), which non-computational linguistic work has demonstrated to be particularly well suited to account for phonologically complex borrowing processes (Kang, 2011). We operationalize OT constraints as features in our borrowing model (§2.2.4). We conduct a case study on Arabic and Swahili, two phylogenetically unrelated languages with a long history of contact; we then apply the model to additional language pairs (§2.2.5). We then

[1] We use Buckwalter notation to write Arabic glosses.

employ models of lexical borrowing to obtain cross-lingual bridges from loanwords in a low-resource language to their donors in a resource-rich language. The donor language is used as pivot to obtain translations via triangulation of out-of-vocabulary loanwords (§2.3). We conduct translation experiments with three resource-poor setups: Swahili–English pivoting via Arabic, Maltese–English pivoting via Italic, and Romanian–English[2] pivoting via French. In intrinsic evaluation, Arabic–Swahili, Italian–Maltese, and French–Romanian borrowing models significantly outperform transliteration and cognate discovery models (§2.4.1). We then provide a systematic quantitative and qualitative analysis of contribution of integrated translations, relative to baselines and oracles, and on corpora of varying sizes (§2.4.2). The proposed pivoting approach yields substantial improvements (up to +1.6 BLEU) in Swahili–Arabic–English translation, moderate improvement (up to +0.8 BLEU) in Maltese–Italian–English translation, and small (+0.2 BLEU) but statistically significant improvements in Romanian–French–English.

Our contributions are twofold. While there have been software implementations of OT (Hayes *et al.*, 2013), they have been used chiefly to facilitate linguistic analysis; we show how to use OT to formulate a model that can be learned with less supervision than linguistically naïve models. To the best of our knowledge, this is the first computational model of lexical borrowing used in a downstream NLP task. Second, we show that lexical correspondences induced using this model can project resources—namely, translations—leading to improved performance in a downstream translation system.

The task of modeling borrowing is under-explored in computational linguistics, although it has both important practical applications and lends itself to modeling with a variety of established computational techniques. In this section we first situate the task with respect to two most closely related research directions: modeling transliteration and modeling cognate forms. We then motivate the new line of research proposed in this work: modeling borrowing.

Borrowing vs. transliteration Borrowing is not transliteration. Transliteration refers to writing in a different *orthography*, whereas borrowing refers to *expanding a language* to include words adapted from another language. Unlike borrowing, transliteration is more amenable to orthographic—rather than morpho-phonological—features, although transliteration can also be prone to phonetic adaptation (Knight and Graehl, 1998). Borrowed words might have begun as transliterations, but a characteristic of borrowed words is that they become assimilated in the linguistic system of the recipient language, and became regular content words, for example, 'orange' and 'sugar' are English words borrowed from Arabic نارنج (*nArnj*) and السكر (*Alskr*), respectively. Whatever their historical origins, synchronically, these words are indistinguishable to most speakers from words that have native ancestral forms in the language. Thus, the morpho-phonological processes that must be accounted for in borrowing models

[2] Romanian is not resource-poor from MT perspective, but in this work we simulate a resource-poor scenario.

are more complex than is required by transliteration models.

Borrowing vs. inheritance Cognates are words in related languages that are inherited from a single word in a common ancestral language (the proto-language). Loanwords, on the other hand, can occur between any languages, either related or not, that historically came into contact. From a modeling perspective, cognates and borrowed words require separate investigation as loanwords are more likely to display marginal phonotactic (and other phonological) patterns than inherited lexical items. Theoretical analysis of cognates has tended to be concerned with a diachronic point of view, that is modeling word changes across time. While of immense scientific interest, language processing applications are arguably better served by models of synchronic processes, peculiar to loanword analysis.

Why borrowing? Borrowing is a distinctive and pervasive phenomenon: *all* languages borrowed from other languages at some point in their lifetime, and borrowed words constitute a large fraction of most language lexicons. Another important property of borrowing is that in adaptation of borrowed items, changes in words are systematic, and knowledge of morphological and phonological patterns in a language can be used to predict how borrowings will be realized in that language, without having to list them all. Therefore, modeling of borrowing is a task well-suited for computational approaches.

Our suggestion in this work is that we can identify borrowing relations between resource-rich donor languages (such as English, French, Spanish, Arabic, Chinese, or Russian) and resource-limited recipient languages. For example, 30–70% of the vocabulary in Vietnamese, Cantonese, and Thai—relatively resource-limited languages spoken by hundreds of millions of people—are borrowed from Chinese and English, languages for which numerous data resources have been created. Similarly, African languages have been greatly influenced by Arabic, Spanish, English, and French—widely spoken languages such as Swahili, Zulu, Malagasy, Hausa, Tarifit, Yoruba contain up to 40% of loanwords. Indo-Iranian languages—Hindustani, Hindi, Urdu, Bengali, Persian, Pashto—spoken by 860 million, also extensively borrowed from Arabic and English (Haspelmath and Tadmor, 2009). In short, at least a billion people are speaking resource-scarce languages whose lexicons are heavily borrowed from resource-rich languages.

Why is this important? Lexical translations or alignments extracted from large parallel corpora have been widely used to project annotations from high- to low-resource languages (Hwa *et al.*, 2005; Täckström *et al.*, 2013; Ganchev *et al.*, 2009, *inter alia*). Unfortunately, large-scale parallel resources are unavailable for the majority of resource-limited languages. Loanwords can be used as a source of cross-lingual links complementary to lexical alignments obtained from parallel data or bilingual lexicons. This holds promise for applying existing cross-lingual methods and bootstrapping linguistic resources in languages where no parallel data is available.

2.2 Constraint-based Models of Lexical Borrowing

Our task is to identify plausible donor–loan word pairs in a language pair. While modeling string transductions is a well-studied problem in NLP, we wish to be able to learn the cross-lingual patterns from minimal training data. We therefore propose a model whose features are motivated by linguistic knowledge—rather than overparameterized with numerous weakly correlated features which are more practical when large amounts of training data is available. The features in our scoring model are inspired by Optimality Theory (OT; §2.2.1), in which borrowing candidates are ranked by universal constraints posited to underly the human faculty of language, and the candidates are determined by transduction processes articulated in prior studies of contact linguistics.

As illustrated in figure 2.2, our model is conceptually divided into three main parts: (1) a mapping of orthographic word forms in two languages into a common phonetic space; (2) generation of loanword pronunciation candidates from a donor word; and (3) ranking of generated loanword candidates, based on linguistic constraints of the donor and recipient languages. In our proposed system, parts (1) and (2) are rule-based; whereas (3) is learned. Each component of the model is discussed in detail in the rest of this section.

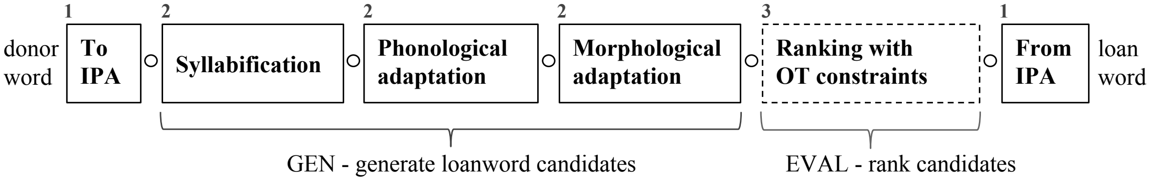


Figure 2.2: Our morpho-phonological borrowing model conceptually has three main parts: (1) conversion of orthographic word forms to pronunciations in International Phonetic Alphabet format; (2) generation of loanword pronunciation candidates; (3) ranking of generated candidates using Optimality-Theoretic constraints. Part (1) and (2) are rule-based, (1) uses pronunciation dictionaries, (2) is based on prior linguistic studies; part (3) is learned. In (3) we learn OT constraint weights from a few dozen automatically extracted training examples.

The model is implemented as a cascade of finite-state transducers. Parts (1) and (2) amount to unweighted string transformation operations. In (1), we convert orthographic word forms to their pronunciations in the International Phonetic Alphabet (IPA), these are pronunciation transducers. In (2) we syllabify donor pronunciations, then perform insertion, deletion, and substitution of phonemes and morphemes (affixes), to generate multiple loanword candidates from a donor word. Although string transformation transducers in (2) can generate loanword candidates that are not found in a recipient language vocabulary, such candidates are filtered out due to composition with the recipient language lexicon acceptor.

Our model performs string transformations from donor to recipient (recapitulating the historical process). However, the resulting relation (i.e., the final composed transducer) is a bidirectional model which can just as well be used to reason about underlying donor forms given recipient forms. In a probabilistic cascade, Bayes' rule could be used to reverse the direction and infer underlying donor forms given a loanword. However, we instead opt to train the model discriminatively to find the most likely underlying form, given a loanword. In part (3), candidates are "evaluated" (i.e., scored) with a weighted sum of universal constraint violations. The non-negative weights, which we call the "cost vector", constitute our model parameters and are learned using a small training set of donor–recipient pairs. We use a shortest path algorithm to find the path with the minimal cost.

2.2.1 OT: constraint-based evaluation

Borrowing relations may be the result of quite complex transformations on the surface. Our decision to evaluate borrowing candidates by weighting counts of "constraint violations" is based on Optimality Theory, which has shown that complex surface phenomena can be well-explained as the interaction of constraints on the form of outputs and the relationships of inputs and outputs (Kager, 1999).

OT posits that surface phonetic words of a language emerge from *underlying phonological forms* according to a two-stage process: first, various candidates for the surface form are enumerated for consideration (the 'generation' or GEN phase); then, these candidates are weighed against one another to see which most closely conforms to—or equivalently, least egregiously violates—the phonological preferences of the language. If the preferences are correctly characterized, then the actual surface form should be selected as the optimal realization of the underlying form. Such preferences are expressed as violable constraints ('violable' because in many cases there may be no candidate that satisfies all of them).

There are two types of OT constraints: *markedness* and *faithfulness* constraints. Markedness constraints (McCarthy and Prince, 1995) describe unnatural (dispreferred) patterns in the language. Faithfulness constraints (Prince and Smolensky, 2008) reward correspondences between the underlying form and the surface candidates. To clarify the distinction between faithfulness and markedness constraint groups to the NLP readership, we can draw the following analogy to the components of machine translation or speech recognition: faithfulness constraints are analogical to the translation model or acoustic model (reflecting how well an output candidate is appropriate to the input), while markedness constraints are analogical to the language model (requiring well-formedness of the output candidate). Without faithfulness constraints, the optimal surface form could differ arbitrarily from the underlying form. As originally proposed, OT holds that the set of constraints is universal, but their ranking is language-specific.

/ɛg/	DEP-IO	MAX-IO	ONSET	NO-CODA
a. ☞ ɛg			*	*
b. ɛgə	*!		*	
c. ɛ		*!	*	
d. ʔɛg	*!			*

Table 2.1: A constraint tableau. DEP-IO » MAX-IO » ONSET » NO-CODA are ranked OT constraints according to the phonological system of English. /ɛg/ is the underlying phonological form, and (a), (b) (c), and (d) are the output candidates under consideration. The actual surface form is (a), as it incurs lower ranked violations than other candidates.

[ʃarr]	*COMPLEX	NO-CODA	MAX-IO	DEP-IO
a. ʃarr	*!	*		
b. ʃar.ri		*!		*
c. ☞ ʃa.ri			*	*
d. ʃa.rri	*!			*

Table 2.2: An example OT analysis adapted to account for borrowing. OT constraints are ranked according to the phonological system of the recipient language (here, Swahili). The donor (Arabic) word شرّ (*$r~*) 'evil' is considered as the underlying form. The winning surface form (c) is the Swahili loanword *shari* 'evil'.

In OT, then, the "grammar" is the set of universal constraints and their language-specific ranking, and a "derivation" for a surface form consists of its underlying form, surface candidates, and constraint violations by those candidates (under which the surface form is correctly chosen). An example of OT analysis is shown in table 2.1; OT constraints will be explained later, in Tables 2.3 and 2.4.

OT has been adapted to account for borrowing by treating the donor language word as the underlying form for the recipient language; that is, the phonological system of the recipient language is encoded as a system of constraints, and these constraints account for how the donor word is adapted when borrowed. We show an example in table 2.2. There has been substantial prior work in linguistics on borrowing in the OT paradigm (Yip, 1993; Davidson and Noyer, 1997; Jacobs and Gussenhoven, 2000; Kang, 2003; Broselow, 2004; Adler, 2006; Rose and Demuth, 2006; Kenstowicz and Suchato, 2006; Kenstowicz, 2007; Mwita, 2009), but none of it has led to computational realizations.

OT assumes an ordinal constraint ranking and strict dominance rather than constraint "weighting". In that, our OT-inspired model departs from OT's standard evaluation assumptions: following Goldwater and Johnson (2003), we use a linear scoring scheme.

2.2.2 Case study: Arabic–Swahili borrowing

In this section, we use the Arabic–Swahili[3] language-pair to describe the prototypical linguistic adaptation processes that words undergo when borrowed. Then, we describe how we model these processes in more general terms.

The Swahili lexicon has been influenced by Arabic as a result of a prolonged period of language contact due to Indian Ocean trading (800 CE–1920), as well as the influence of Islam (Rothman, 2002). According to several independent studies, Arabic loanwords constitute from 18% (Hurskainen, 2004b) to 40% (Johnson, 1939) of Swahili word types. Despite a strong susceptibility of Swahili to borrowing and a large fraction of Swahili words originating from Arabic, the two languages are typologically distinct with profoundly dissimilar phonological and morpho-syntactic systems. We survey these systems briefly since they illustrate how Arabic loanwords have been substantially adapted to conform to Swahili phonotactics. First, Arabic has five syllable patterns:[4] CV, CVV, CVC, CVCC, and CVVC (McCarthy, 1985, pp. 23–28), whereas Swahili (like other Bantu languages) has only open syllables of the form CV or V. At the segment level, Swahili loanword adaptation thus involves extensive vowel epenthesis in consonant clusters and at a syllable final position if the syllable ends with a consonant, for example, كتاب (*ktAb*) → *kitabu* 'book' (Polomé, 1967; Schadeberg, 2009; Mwita, 2009). Second, phonological adaptation in Swahili loanwords includes shortening of vowels (unlike Arabic, Swahili does not have phonemic length); substitution of consonants that are found in Arabic but not in Swahili (e.g., emphatic (pharyngealized) /tˤ/→/t/, voiceless velar fricative /x/→/k/, dental fricatives /θ/→/s/, /ð/→/z/, and the voiced velar fricative /ɣ/→/g/); adoption of Arabic phonemes that were not originally present in Swahili /θ/, /ð/, /ɣ/ (e.g., تحذير (*tH*yr*)→ *tahadhari* 'warning'); degemination of Arabic geminate consonants (e.g., شرّ (*$r~*)→ *shari* 'evil'). Finally, adapted loanwords can freely undergo Swahili inflectional and derivational processes, for example, الوزير (*Alwzyr*) → *waziri* 'minister', *mawaziri* 'ministers', *kiuwaziri* 'ministerial' (Zawawi, 1979; Schadeberg, 2009).

2.2.3 Arabic–Swahili borrowing transducers

We use unweighted transducers for pronunciation, syllabification, and morphological and phonological adaptation and describe these below. An example that illustrates some of the possible string transformations by individual components of the model is shown in figure 2.3. The goal of these transducers is to *minimally* overgenerate Swahili adapted forms of Arabic words, based on the adaptations described above.

[3] For simplicity, we subsume Omani Arabic and other historical dialects of Arabic under the label "Arabic"; our data and examples are in Modern Standard Arabic. Similarly, we subsume Swahili, its dialects and protolanguages under "Swahili".

[4] C stands for consonant, and V for vowel.

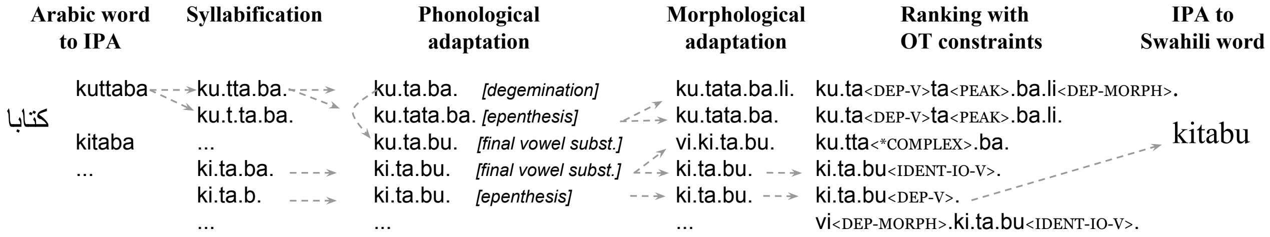


Figure 2.3: An example of an Arabic word كتابا (*ktAbA*) 'book.sg.indef' transformed by our model into a Swahili loanword *kitabu*.

Pronunciation

Based on the IPA, we assign shared symbols to sounds that exist in both sound systems of Arabic and Swahili (e.g., nasals /n/, /m/; voiced stops /b/, /d/), and language-specific unique symbols to sounds that are unique to the phonemic inventory of Arabic (e.g., pharyngeal voiced and voiceless fricatives /ħ/, /ʕ/) or Swahili (e.g., velar nasal /ŋ/). For Swahili, we construct a pronunciation dictionary based on the Omniglot grapheme-to-IPA mapping.[5] In Arabic, we use the CMU Arabic vowelized pronunciation dictionary containing about 700K types which has the average of four pronunciations per unvowelized input word type (Metze *et al.*, 2010).[6] We then design four transducers—Arabic and Swahili word-to-IPA and IPA-to-word transducers—each as a union of linear chain transducers, as well as one acceptor per pronunciation dictionary listing.

Syllabification

Arabic words borrowed into Swahili undergo a repair of violations of the Swahili segmental and phonotactic constraints, for example via vowel epenthesis in a consonant cluster. Importantly, *repair depends upon syllabification*. To simulate plausible phonological repair processes, we generate multiple syllabification variants for input pronunciations. The syllabification transducer optionally inserts syllable separators between phones. For example, for an input phonetic sequence /kuttɑbɑ/, the output strings include /ku.t.tɑ.bɑ/, /kut.tɑ.bɑ/, and /ku.ttɑ.bɑ/ as syllabification variants; each variant violates different constraints and consequently triggers different phonological adaptations.

[5] www.omniglot.com

[6] Since we are working at the level of word types which have no context, we cannot disambiguate the intended form, so we include all options. For example, for the input word كتابا (*ktAbA*) 'book.sg.indef', we use both pronunciations /kitɑbɑ/ and /kuttɑbɑ/.

Phonological adaptation

Phonological adaptation of syllabified phone sequences is the crux of the loanword adaptation process. We implement phonological adaptation transducers as a composition of plausible context-dependent insertions, deletions, and substitutions of phone subsets, based on prior studies summarized in §2.2.2. In what follows, we list phonological adaptation components in the order of transducer composition in the borrowing model. The **vowel deletion** transducer shortens Arabic long vowels and vowel clusters. The **consonant degemination** transducer shortens Arabic geminate consonants, for example, it degeminates /tt/ in /ku.ttɑ.bɑ/, outputting /ku.tɑ.bɑ/. The **substitution of similar phonemes** transducer substitutes similar phonemes and phonemes that are found in Arabic but not in Swahili (Polomé, 1967, p. 45). For example, the emphatic /t^{ʕ}/, /d^{ʕ}/, /s^{ʕ}/ are replaced by the corresponding non-emphatic segments [t], [d], [s]. The **vowel epenthesis** transducer inserts a vowel between pairs of consonants (/ku.ttɑ.bɑ/ → /ku.tatɑ.bɑ/), and at the end of a syllable, if the syllable ends with a consonant (/ku.t.tɑ.bɑ/ → /ku.ta.tɑ.bɑ/). Sometimes it is possible to predict the final vowel of a word, depending on the word-final coda consonant of its Arabic counterpart: /u/ or /o/ added if an Arabic donor ends with a labial, and /i/ or /e/ added after coronals and dorsals (Mwita, 2009). Following these rules, the **final vowel substitution** transducer complements the inventory of final vowels in loanword candidates.

Morphological adaptation

Both Arabic and Swahili have significant morphological processes that alter the appearance of lemmas. To deal with morphological variants, we construct morphological adaptation transducers that optionally strip Arabic concatenative affixes and clitics, and then optionally append Swahili affixes, generating a superset of all possible loanword hypotheses. We obtain the list of Arabic affixes from the Arabic morphological analyzer SAMA (Maamouri *et al.*, 2010); the Swahili affixes are taken from a hand-crafted Swahili morphological analyzer (Littell *et al.*, 2014). For the sake of simplicity in implementation, we strip no more than one Arabic prefix and no more than one suffix per word; and in Swahili – we concatenate at most two Swahili prefixes and at most one suffix.

2.2.4 Learning constraint weights

Due to the computational problems of working with OT (Eisner, 1997, 2002), we make simplifying assumptions by (1) bounding the theoretically infinite set of underlying forms with a small linguistically-motivated subset of allowed transformations on donor pronunciations, as described in §2.2.3; (2) imposing a priori restrictions on the set of the surface realizations by intersecting the candidate set with the recipient pronunciation lexicon; (3) assuming that the set of constraints is finite and regular (Ellison, 1994); and (4) assigning linear weights to

Faithfulness constraints	
MAX-IO-MORPH	no (donor) affix deletion
MAX-IO-C	no consonant deletion
MAX-IO-V	no vowel deletion
DEP-IO-MORPH	no (recipient) affix epenthesis
DEP-IO-V	no vowel epenthesis
IDENT-IO-C	no consonant substitution
IDENT-IO-C-M	no substitution in manner of pronunciation
IDENT-IO-C-A	no substitution in place of articulation
IDENT-IO-C-S	no substitution in sonority
IDENT-IO-C-P	no pharyngeal consonant substitution
IDENT-IO-C-G	no glottal consonant substitution
IDENT-IO-C-E	no emphatic consonant substitution
IDENT-IO-V	no vowel substitution
IDENT-IO-V-O	no substitution in vowel openness
IDENT-IO-V-R	no substitution in vowel roundness
IDENT-IO-V-F	no substitution in vowel frontness
IDENT-IO-V-FIN	no final vowel substitution

Table 2.3: Faithfulness constraints prefer pronounced realizations completely congruent with their underlying forms.

constraints, rather than learning an ordinal constraint ranking with strict dominance (Boersma and Hayes, 2001; Goldwater and Johnson, 2003).

As discussed in §2.2.1, OT distinguishes markedness constraints which detect dispreferred phonetic patterns in the language, and faithfulness constraints, which ensure correspondences between the underlying form and the surface candidates. The implemented constraints are listed in Tables 2.3 and 2.4. Faithfulness constraints are integrated in phonological transformation components as transitions following each insertion, deletion, or substitution. Markedness constraints are implemented as standalone identity transducers: inputs are equal outputs, but path weights representing candidate evaluation with respect to violated constraints are different.

The final "loanword transducer" is the composition of all transducers described in §2.2.3 and OT constraint transducers. A path in the transducer represents a syllabified phonemic sequence along with (weighted) OT constraints it violates, and shortest path outputs are those, whose cumulative weight of violated constraints is minimal.

OT constraints are realized as features in our linear model, and feature weights are learned in a discriminative training to maximize the accuracy obtained by the loanword transducer on a small development set of donor–recipient pairs. For parameter estimation, we employ the Nelder–Mead algorithm (Nelder and Mead, 1965), a heuristic derivative-free method that iteratively optimizes, based on an objective function evaluation, the convex hull of $n+1$ simplex

Markedness constraints	
NO-CODA	syllables must not have a coda
ONSET	syllables must have onsets
PEAK	there is only one syllabic peak
SSP	complex onsets rise in sonority, complex codas fall in sonority
*COMPLEX-S	no consonant clusters on syllable margins
*COMPLEX-C	no consonant clusters within a syllable
*COMPLEX-V	no vowel clusters

Table 2.4: Markedness constraints impose language-specific structural well-formedness of surface realizations.

vertices.[7] The objective function used in this work is the soft accuracy of the development set, defined as the proportion of correctly identified donor words in the total set of 1-best outputs.

2.2.5 Adapting the model to a new language

The Arabic–Swahili case study shows that, in principle, a borrowing model can be constructed. But a reasonable question to ask is: how much work is required to build a similar system for a new language pair? We claim that our design permits rapid development in new language pairs. First, string transformation operations, as well as OT constraints, are language-universal. The only adaptation required is a linguistic analysis to identify plausible morpho-phonological repair strategies for the new language pair (i.e., a subset of allowed insertions, deletions, and substitutions of phonemes and morphemes). Since we need only to overgenerate candidates (the OT constraints will filter bad outputs), the effort is minimal relative to many other grammar engineering exercises. The second language-specific component is the grapheme-to-IPA converter. While this can be a non-trivial problem in some cases, the problem is well studied, and many under-resourced languages (e.g., Swahili), have "phonographic" systems where orthography corresponds to phonology. This tendency can be explained by the fact that, in many cases, lower-resource languages have developed orthography relatively recently, rather than having organically evolved written forms that preserve archaic or idiosyncratic spellings that are more distantly related to the current phonology of the language such as we see in English.

To illustrate the ease with which a language pair can be engineered, we applied our borrowing model to the Italian–Maltese and French–Romanian language pairs. Maltese and Romanian, like Swahili, have a large number of borrowed words in their lexicons (Tadmor, 2009). Maltese (a phylogenetically Semitic language) has 30.3%–35.1% loanwords of Romance (Ital-

[7] The decision to use Nelder–Mead rather than more conventional gradient-based optimization algorithms was motivated purely by practical limitations of the finite-state toolkit we used which made computing derivatives with latent structure impractical from an engineering standpoint.

ian/Sicilian) origin (Comrie and Spagnol, 2015). Although French and Romanian are sister languages (both descending from Latin), about 12% of Romanian types are French borrowings that came into the language in the past few centuries (Schulte, 2009). For both language pairs we manually define a set of allowed insertions, deletions, and substitutions of phonemes and morphemes, based on the training sets. A set of Maltese affixes was defined based on the linguistic survey by Fabri *et al.* (2014). We employ the GlobalPhone pronunciation dictionary for French (Schultz and Schlippe, 2014), converted to IPA, and automatically constructed Italian, Romanian, and Maltese pronunciation dictionaries using the Omniglot grapheme-to-IPA conversion rules for those languages.

2.3 Models of Lexical Borrowing in Statistical Machine Translation

Before turning to an experimental verification and analysis of the borrowing model, we introduce an external application where the borrowing model will be used as a component—machine translation. We rely on the borrowing model to project translation information from a high-resource donor language into a low-resource recipient language, thus mitigating the deleterious effects of out-of-vocabulary (OOV) words.

OOVs are a ubiquitous and difficult problem in MT. When a translation system encounters an OOV—a word that was not observed in the training data, and the trained system thus lacks its translation variants—it usually outputs the word just as it is in the source language, producing erroneous and disfluent translations. All MT systems, even when trained on billion-sentence-size parallel corpora, will encounter OOVs at test time. Often, these are named entities and neologisms. However, the OOV problem is much more acute in morphologically-rich and low-resource scenarios: there, OOVs are primarily not lexicon-peripheral items such as names and specialized/technical terms, but also regular content words. Since borrowed words are a component of the regular lexical content of a language, projecting translations onto the recipient language by identifying borrowed lexical material is a plausible strategy for solving this problem.

Procuring translations for OOVs has been a subject of active research for decades. Translation of named entities is usually generated using transliteration techniques (Al-Onaizan and Knight, 2002; Hermjakob *et al.*, 2008; Habash, 2008). Extracting a translation lexicon for recovering OOV content words and phrases is done by mining bi-lingual and monolingual resources (Rapp, 1995; Callison-Burch *et al.*, 2006; Haghighi *et al.*, 2008; Marton *et al.*, 2009; Razmara *et al.*, 2013; Saluja *et al.*, 2014). In addition, OOV content words can be recovered by exploiting cognates, by transliterating and then "pivoting" via a closely-related resource-richer language, when such a language exists (Hajič *et al.*, 2000; Mann and Yarowsky, 2001; Kondrak *et al.*, 2003; De Gispert and Marino, 2006; Habash and Hu, 2009; Durrani *et al.*, 2010; Wang *et al.*, 2012; Nakov and Ng, 2012; Dholakia and Sarkar, 2014). Our work is similar in

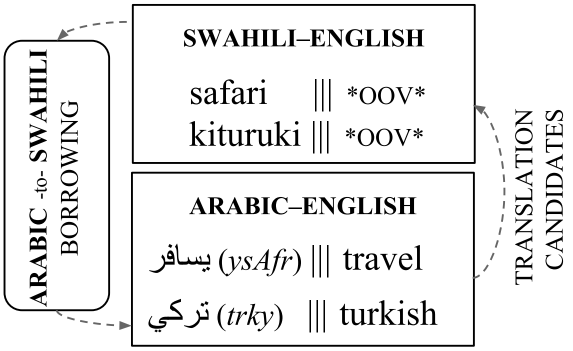


Figure 2.4: To improve a resource-poor Swahili–English MT system, we extract translation candidates for OOV Swahili words borrowed from Arabic using the Swahili-to-Arabic borrowing system and Arabic–English resource-rich MT.

spirit to the latter pivoting approach, but we show how to obtain translations for OOV content words by pivoting via an unrelated, often typologically distant resource-rich language.

Our solution is depicted, at a high level, in figure 2.4. Given an OOV word in resource-poor MT, we use our borrowing system to identify list of likely donor words from the donor language. Then, using the MT system in the resource-rich language, we translate the donor words to the same target language as in the resource-poor MT (here, English). Finally, we integrate translation candidates in the resource-poor system.

We now discuss integrating translation candidates acquired via borrowing plus resource-rich translation.

Briefly, phrase-based translation works as follows. A set of candidate translations for an input sentence is created by matching contiguous spans of the input against an inventory of phrasal translations, reordering them into a target-language appropriate order, and choosing the best one according to a model that combines features of the phrases used, reordering patterns, and target language model (Koehn *et al.*, 2003). A limitation of this approach is that it can only generate input/output phrase pairs that were directly observed in the training corpus. In resource-limited languages, the standard phrasal inventory will generally be incomplete due to limited parallel data. Thus, the decoder's only hope for producing a good output is to find a fluent, meaning-preserving translation using incomplete translation lexicons. "Synthetic phrases" is a strategy of integrating translated phrases directly in the MT *translation model*, rather than via pre- or post-processing MT inputs and outputs (Tsvetkov *et al.*, 2013b; Chahuneau *et al.*, 2013; Schlinger *et al.*, 2013; Ammar *et al.*, 2013; Tsvetkov *et al.*, 2014b; Tsvetkov and Dyer, 2015). Synthetic phrases are phrasal translations that are not directly extractable from the training data, generated by auxiliary translation and postediting processes (for example, extracted from a borrowing model). An important advantage of synthetic phrases is that the process often benefits from phrase synthesizers that have high recall (relative to precision) since the global translation model will still have the final say on whether a synthesized phrase will be used.

For each OOV, the borrowing system produces the n-best list of plausible donors; for each donor we then extract the k-best list of its translations.[8] Then, we pair the OOV with the resulting $n \times k$ translation candidates. The translation candidates are noisy: some of the generated donors may be erroneous, the errors are then propagated in translation.[9] To allow the low-resource translation system to leverage good translations that are missing in the default phrase inventory, while being able to learn how trustworthy they are, we integrate the borrowing-model acquired translation candidates as synthetic phrases.

To let the translation model learn whether to trust these phrases, the translation options obtained from the borrowing model are augmented with an indicator feature indicating that the phrase was generated externally (i.e., rather than being extracted from the parallel data). Additional features assess properties of the donor–loan words' relation; their goal is to provide an indication of plausibility of the pair (to mark possible errors in the outputs of the borrowing system). We employ two types of features: phonetic and semantic. Since borrowing is primarily a phonological phenomenon, phonetic features will provide an indication of how typical (or atypical) pronunciation of the word in a language; loanwords are expected to be less typical than core vocabulary words. The goal of semantic features is to measure semantic similarity between donor and loan words: erroneous candidates and borrowed words that changed meaning over time are expected to have different meaning from the OOV.

2.3.1 Phonetic features

To compute phonetic features we first train a (5-gram) language model (LM) of IPA pronunciations of the donor/recipient language vocabulary (p_ϕ). Then, we re-score pronunciations of the donor and loanword candidates using the LMs. We hypothesize that in donor–loanword pairs both the donor and the loanword phone LM score is high. We capture this intuition in three features: $f_1 = p_\phi(donor)$, $f_2 = p_\phi(loanword)$, and the harmonic mean between the two scores $f_3 = \frac{2f_1f_2}{f_1+f_2}$ (the harmonic mean of a set of values is high only when all of the values are high).

[8] We set n and k to 5; we did not experiment with other values.

[9] We give as input into the borrowing system all OOV words, although, clearly, not all OOVs are loanwords, and not all loanword OOVs are borrowed from the donor language. However, an important property of the borrowing model is that its operations are not general, but specific to the language-pair and reduced only to a small set of plausible changes that the donor word can undergo in the process of assimilation in the recipient language. Thus, the borrowing system only *minimally* overgenerates the set of output candidates given an input. If the borrowing system encounters an input word that was not borrowed from the target donor language, it usually (but not always) produces an empty output.

2.3.2 Semantic features

We compute a semantic similarity feature between the candidate donor and the OOV loanword as follows. We first train, using large monolingual corpora, 100-dimensional word vector representations for donor and recipient language vocabularies.[10] Then, we employ canonical correlation analysis (CCA) with small donor–loanword dictionaries (training sets in the borrowing models) to project the word embeddings into 50-dimensional vectors with maximized correlation between their dimensions. The semantic feature annotating the synthetic translation candidates is cosine distance between the resulting donor and loanword vectors. We use the `word2vec` Skip-gram model (Mikolov *et al.*, 2013a) to train monolingual vectors,[11] and the CCA-based tool (Faruqui and Dyer, 2014b) for projecting word vectors.[12]

2.4 Experiments

We now turn to the problem of empirically validating the model we have proposed. Our evaluation consists of two parts. First, we perform an intrinsic assessment of the model's ability to learn borrowing correspondences and compare these to similar approaches that use less linguistic knowledge but which have been used to solve similar string mapping problems. Second, we show the effect of borrowing-augmented translations in translation systems, exploring the effects of the features proposed above.

2.4.1 Intrinsic evaluation of models of lexical borrowing

Our experimental setup is defined as follows. The input to the borrowing model is a loanword candidate in Swahili/Maltese/Romanian, the outputs are plausible donor words in the Arabic/Italian/French monolingual lexicon (i.e., any word in pronunciation dictionary). We train the borrowing model using a small set of training examples, and then evaluate it using a held-out test set. In the rest of this section we describe in detail our datasets, tools, and experimental results.

Resources

We employ Arabic–English and Swahili–English bitexts to extract a training set (corpora of sizes 5.4M and 14K sentence pairs, respectively), using a cognate discovery technique (Kondrak, 2001). Phonetically and semantically similar strings are classified as cognates; phonetic similarity is the string similarity between phonetic representations, and semantic similarly is

[10] We assume that while parallel data is limited in the recipient language, monolingual data is available.

[11] `code.google.com/p/word2vec`

[12] `github.com/mfaruqui/eacl14-cca`

approximated by translation.[13] We thereby extract Arabic and Swahili pairs $\langle a, s\rangle$ that are phonetically similar ($\frac{\Delta(a,s)}{\min(|a|,|s|)} < 0.5$) where $\Delta(a, s)$ is the Levenshtein distance between a and s and that are aligned to the same English word e. FastAlign (Dyer *et al.*, 2013) is used for word alignments. Given an extracted word pair $\langle a, s\rangle$, we also extract word pairs $\{\langle a', s\rangle\}$ for all proper Arabic words a' which share the same lemma with a producing on average 33 Arabic types per Swahili type. We use MADA (Habash *et al.*, 2009) for Arabic morphological expansion.

From the resulting dataset of 490 extracted Arabic–Swahili borrowing examples,[14] we set aside randomly sampled 73 examples (15%) for evaluation,[15] and use the remaining 417 examples for model parameter optimization. For Italian–Maltese language pair, we use the same technique and extract 425 training and 75 (15%) randomly sampled test examples. For French–Romanian language pair, we use an existing small annotated set of borrowing examples,[16] with 282 training and 50 (15%) randomly sampled test examples.

We use `pyfst`—a Python interface to OpenFst (Allauzen *et al.*, 2007)—for the borrowing model implementation.[17]

Baselines

We compare our model to several baselines. In the Levenshtein distance baselines we chose the closest word (either surface or pronunciation-based). In the cognates baselines, we evaluate a variant of the Levenshtein distance tuned to identify cognates (Mann and Yarowsky, 2001; Kondrak and Sherif, 2006); this method was identified by Kondrak and Sherif (2006) among the top three cognate identification methods. In the transliteration baselines we generate plausible transliterations of the input Swahili (or Romanian) words in the donor lexicon using the model of Ammar *et al.* (2012), with multiple references in a lattice and without reranking. The CRF transliteration model is a linear-chain CRF where we label each source character with a sequence of target characters. The features are label unigrams, label bigrams, and label conjoined with a moving window of source characters. In the OT-uniform baselines, we evaluate the accuracy of the borrowing model with uniform weights, thus the shortest path in the loanwords transducer will be forms that violate the fewest constraints.

[13] This cognate discovery technique is sufficient to extract a small training set, but is not generally applicable, as it requires parallel corpora or manually constructed dictionaries to measure semantic similarity. Large parallel corpora are unavailable for most language pairs, including Swahili–English.

[14] In each training/test example one Swahili word corresponds to all extracted Arabic donor words.

[15] We manually verified that our test set contains clear Arabic–Swahili borrowings. For example, we extract Swahili *kusafiri, safari* and Arabic السفر, يسفر, سفر (*Alsfr, ysAfr, sfr*) all aligned to 'travel'.

[16] `http://wold.clld.org/vocabulary/8`

[17] `https://github.com/vchahun/pyfst`

	AR–SW	IT–MT	FR–RO
Reachability	87.7%	92.7%	82.0%
Ambiguity	857	11	12

Table 2.5: The evaluation of the borrowing model design. Reachability is the percentage of donor–recipient pairs that are reachable from a donor to a recipient language. Ambiguity is the average number of outputs that the model generates per one input.

Evaluation

In addition to predictive accuracy on all models (if a model produces multiple hypotheses with the same 1-best weight, we count the proportion of correct outputs in this set), we evaluate two particular aspects of our proposed model: (1) appropriateness of the model family, and (2) the quality of the learned OT constraint weights. The first aspect is designed to evaluate whether the morpho-phonological transformations implemented in the model are required *and* sufficient to generate loanwords from the donor inputs. We report two evaluation measures: model *reachability* and *ambiguity*. Reachability is the percentage of test samples that are reachable (i.e., there is a path from the input test example to a correct output) in the loanword transducer. A naïve model which generates all possible strings would score 100% reachability; however, inference may be expensive and the discriminative component will have a greater burden. In order to capture this trade-off, we also report the inherent *ambiguity* of our model, which is the average number of outputs potentially generated per input. A generic Arabic–Swahili transducer, for example, has an ambiguity of 786,998—the size of the Arabic pronunciation lexicon.[18]

Results

The reachability and ambiguity of the borrowing model are listed in table 2.5. Briefly, the model obtains high reachability, while significantly reducing the average number of possible outputs per input: in Arabic from 787K to 857 words, in Maltese from 129K to 11, in French from 62K to 12. This result shows that the loanword transducer design, based on the prior linguistic analysis, is a plausible model of word borrowing. Yet, there are on average 33 correct Arabic words out of the possible 857 outputs, thus the second part of the model—OT constraint weights optimization—is crucial.

The accuracy results in table 2.6 show how challenging the task of modeling lexical borrowing between two distinct languages is, and importantly, that orthographic and phonetic baselines including the state-of-the-art generative model of transliteration are not suitable for this task. Phonetic baselines for Arabic–Swahili perform better than orthographic ones, but

[18] Our measure of ambiguity is equivalent to perplexity assuming a uniform distribution over output forms.

		Accuracy (%)		
		AR–SW	IT–MT	FR–RO
Orthographic baselines	Levenshtein-orthographic	8.9	61.5	38.0
	Transliteration	16.4	61.3	36.0
Phonetic baselines	Levenshtein-pronunciation	19.8	64.4	26.3
	Cognates	19.7	63.7	30.7
OT	OT-uniform constraint weights	29.3	65.6	58.5
	OT-learned constraint weights	**48.4**	**83.3**	**75.6**

Table 2.6: The evaluation of the borrowing model accuracy. We compare the following setups: orthographic (surface) and phonetic (based on pronunciation lexicon) Levenshtein distance, a cognate identification model that uses heuristic Levenshtein distance with lower penalty on vowel updates and similar letter/phone substitutions, a CRF transliteration model, and our model with uniform and learned OT constraint weights assignment.

substantially worse than OT-based models, even if OT constraints are not weighted. Crucially, the performance of the borrowing model with the learned OT weights corroborates the assumption made in numerous linguistic accounts that OT is an adequate analysis of the lexical borrowing phenomenon.

Qualitative evaluation

The constraint ranking learned by the borrowing model (constraints are listed in Tables 2.3, 2.4) is in line with prior linguistic analysis. In Swahili no-coda dominates all other markedness constraints. Both *complex-s and *complex-c, restricting consonant clusters, dominate *complex-v, confirming that Swahili is more permissive to vowel clusters. ssp—sonority-based constraint—captures a common pattern of consonant clustering, found across languages, and is also learned by our model as undominated by most competitors in Swahili, and as a dominating markedness constraint in Romanian. Morphologically-motivated constraints also comply with tendencies discussed in linguistic literature: donor words may remain unmodified and are treated as a stem, and then are reinfected according to the recipient morphology, thus dep-io-morph can be dominated more easily than max-io-morph. Finally, vowel epenthesis dep-io-v is the most common strategy in Arabic loanword adaptation, and is ranked lower according to the model; however, it is ranked highly in the French–Romanian model, where vowel insertion is rare.

A second interesting by-product of our model is an inferred syllabification. While we did not conduct a systematic quantitative evaluation, higher-ranked Swahili outputs tend to contain linguistically plausible syllabifications, although the syllabification transducer inserts optional syllable boundaries between every pair of phones. This result further attests to the plausible constraint ranking learned by the model. Example Swahili syllabifications[19] along with the constraint violations produced by the borrowing model are depicted in table 2.7.

en	ar orth.	ar pron.	sw syl.	Violated constraints
book	ktAb	kitAb	ki.ta.bu.	ident-io-c-g$\langle A, a\rangle$, dep-io-v$\langle \epsilon, u\rangle$
palace	AlqSr	AlqaSr	ka.sri	max-io-morph$\langle Al, \epsilon\rangle$, ident-io-c-s$\langle q, k\rangle$, ident-io-c-e$\langle S, s\rangle$, *complex-c$\langle sr\rangle$, dep-io-v$\langle \epsilon, i\rangle$
wage	Ajrh	Aujrah	u.ji.ra.	max-io-v$\langle A, \epsilon\rangle$, onset$\langle u\rangle$, dep-io-v$\langle \epsilon, i\rangle$, max-io-c$\langle h, \epsilon\rangle$

Table 2.7: Examples of inferred syllabification and corresponding constraint violations produced by our borrowing model.

2.4.2 Extrinsic evaluation of pivoting via borrowing in MT

We now turn to an extrinsic evaluation, looking at two low-resource translation tasks: Swahili–English translation (resource-rich donor language: Arabic), Maltese–English translation (resource-rich donor language: Italian), and Romanian–English translation (resource-rich donor language: French). We begin by reviewing the datasets used, and then discuss two oracle experiments that attempt to quantify how much value could we obtain from a perfect borrowing model (since not all mistakes made by MT systems involve borrowed words). Armed with this understanding, we then explore how much improvement can be obtained using our system.

Datasets and software

The Swahili–English parallel corpus was crawled from the Global Voices project website[20]. For the Maltese–English language pair, we sample a parallel corpus of the same size from the EUbookshop corpus from the OPUS collection (Tiedemann, 2012). Similarly, to simulate resource-poor scenario for the Romanian–English language pair, we sample a corpus from the transcribed TED talks (Cettolo *et al.*, 2012). To evaluate translation improvement on corpora of different sizes we conduct experiments with sub-sampled 4,000, 8,000, and 14,000 parallel sentences from the training corpora (the smaller the training corpus, the more OOVs it has). Corpora sizes along with statistics of source-side OOV tokens and types are given in Table 2.8. Statistics of the held-out dev and test sets used in all translation experiments are given in table 2.9.

The Arabic–English pivot translation system was trained on a parallel corpus of about 5.4 million sentences available from the Linguistic Data Consortium (LDC), and optimized on the standard NIST MTEval dataset for the year 2005 (MT05). The Italian–English system was

[19] We chose examples from the Arabic–Swahili system because this is a more challenging case due to linguistic discrepancies.

[20] `sw.globalvoicesonline.org`

trained on 11 million sentences from the OPUS corpus. The French–English pivot system was trained on about 400,000 sentences from the transcribed TED talks, and optimized on the dev talks from the Romanian–English system; test talks from the Romanian–English system were removed from the French–English training corpus.

In all the MT experiments, we use the `cdec`[21] translation toolkit (Dyer *et al.*, 2010), and optimize parameters with MERT (Och, 2003). English 4-gram language models with Kneser-Ney smoothing (Kneser and Ney, 1995) were trained using KenLM (Heafield, 2011) on the target side of the parallel training corpora and on the Gigaword corpus (Parker *et al.*, 2009). Results are reported using case-insensitive BLEU with a single reference (Papineni *et al.*, 2002). To verify that our improvements are consistent and are not just an effect of optimizer instability, we train three systems for each MT setup; reported BLEU scores are averaged over systems.

		4K	8K	14K
SW–EN	Tokens	84,764	170,493	300,648
	Types	14,554	23,134	33,288
	OOV tokens	4,465 (12.7%)	3,509 (10.0%)	2,965 (8.4%)
	OOV types	3,610 (50.3%)	2,950 (41.1%)	2,523 (35.1%)
MT–EN	Tokens	104,181	206,781	358,373
	Types	14,605	22,407	31,176
	OOV tokens	4,735(8.7%)	3,497 (6.4%)	2,840 (5.2%)
	OOV types	4,171 (44.0%)	3,236 (34.2%)	2,673 (28.2%)
RO–EN	Tokens	35,978	71,584	121,718
	Types	7,210	11,144	15,112
	OOV tokens	3,268 (16.6%)	2,585 (13.1%)	2,177 (11.1%)
	OOV types	2,382 (55.0%)	1,922 (44.4%)	1,649 (38.1%)

Table 2.8: Statistics of the Swahili–English, Maltese–English, and Romanian–English corpora and source-side OOV rates for 4K, 8K, 14K parallel training sentences.

	SW–EN		MT–EN		RO–EN	
	dev	test	dev	test	dev	test
Sentences	1,552	1,732	2,000	2,000	2,687	2,265
Tokens	33,446	35,057	54,628	54,272	24,754	19,659
Types	7,008	7,180	9,508	9,471	5,141	4,328

Table 2.9: Dev and test corpora sizes.

[21] `www.cdec-decoder.org`

Upper bounds

The goal of our experiments is not only to evaluate the contribution of the OOV dictionaries that we extract when pivoting via borrowing, but also to understand the potential contribution of exploiting borrowing. What is the overall improvement that would be achieved if we could correctly translate all OOVs that were borrowed from another language? What is the overall improvement that can be achieved if we correctly translate all OOVs? We answer this question by defining "upper bound" experiments. In the upper bound experiments we word-align all available parallel corpora, including dev and test sets, and extract from the alignments oracle translations of OOV words. Then, we append the extracted OOV dictionaries to the training corpora and re-train SMT setups without OOVs. Translation scores of the resulting system provide an upper bound of an improvement from correctly translating all OOVs. When we append oracle translations of the subset of OOV dictionaries, in particular translations of all OOVs for which the output of the borrowing system is not empty, we obtain an upper bound that can be achieved using our method (if the borrowing system provided perfect outputs relative to the reference translations). Understanding the upper bounds is relevant not only for our experiments, but for any experiments that involve augmenting translation dictionaries; however, we are not aware of prior work providing similar analysis of upper bounds, and we recommend this as a calibrating procedure for future work on OOV mitigation strategies.

Borrowing-augmented setups

As described in §2.3, we integrate translations of OOV loanwords in the translation model using the synthetic phrase paradigm. Due to data sparsity, we conjecture that non-OOVs that occur only few times in the training corpus can also lack appropriate translation candidates, these are target-language OOVs. We therefore plug into the borrowing system OOVs and non-OOV words that occur less than 3 times in the training corpus. We list in table 2.10 the size of resulting borrowed lexicons that we integrate in translation tables.[22]

	4K	8K	14K
Loan OOVs in SW–EN	5,050	4,219	3,577
Loan OOVs in MT–EN	10,138	6,456	4,883
Loan OOVs in RO–EN	347	271	216

Table 2.10: The size of dictionaries extracted using pivoting via borrowing and integrated in translation models.

[22] Differences in statistics stem from differences in types of corpora, such as genre, domain, and morphological richness of the source language.

Transliteration-augmented setups

In addition to the standard baselines, we evaluate transliteration baselines, where we replace the borrowing model by the baselines described in §2.4.1. As in the borrowing system, transliteration outputs are filtered to contain only target language lexicons. We list in table 2.11 the size of obtained translated lexicons.

	4K	8K	14K
Transliteration OOVs in SW–EN	49	32	22
Transliteration OOVs in MT–EN	26,734	19,049	15,008
Transliteration OOVs in RO–EN	906	714	578

Table 2.11: The size of translated lexicons extracted using pivoting via transliteration and integrated in translation models.

Results

Translation results are shown in table 2.12. We evaluate separately the contribution of the integrated OOV translations, and the same translations annotated with phonetic and semantic features. We also provide upper bound scores for integrated loanword dictionaries as well as for recovering all OOVs.

Swahili–English MT performance is improved by up to +1.6 BLEU when we augment it with translated OOV loanwords leveraged from the Arabic–Swahili borrowing and then Arabic–English MT. The contribution of the borrowing dictionaries is +0.6–1.1 BLEU, and phonetic and semantic features contribute additional half BLEU. More importantly, upper bound results show that the system can be improved more substantially with better dictionaries of OOV loanwords. This result confirms that OOV borrowed words is an important type of OOVs, and with proper modeling it has the potential to improve translation by a large margin. Maltese–English system is also improved substantially, by up to +0.8 BLEU, but the contribution of additional features is less pronounced. Romanian–English systems obtain only small but significant improvement for 4K and 8K, $p < .01$ (Clark *et al.*, 2011). However, this is expected as the rate of borrowing from French into Romanian is smaller, and, as the result, the integrated loanword dictionaries are small. Transliteration baseline, conversely, is more effective in Romanian–French language pair, as the languages are related typologically, and have common cognates in addition to loanwords. Still, even with these dictionaries the translations with pivoting via borrowing/transliteration improve, and even almost approach the upper bounds results.

		4K	8K	14K
SW–EN	Baseline	13.2	15.1	17.1
	+ Transliteration OOVs	13.4	15.3	17.2
	+ Loan OOVs	14.3	15.7	18.2
	+ Features	**14.8**	**16.4**	**18.4**
	Upper bound loan	18.9	19.1	20.7
	Upper bound all OOVs	19.2	20.4	21.1
MT–EN	Baseline	26.4	31.4	35.2
	+ Transliteration OOVs	26.5	30.8	34.9
	+ Loan OOVs	**27.2**	31.7	**35.3**
	+ Features	26.9	**31.9**	34.5
	Upper bound loan	28.5	32.2	35.7
	Upper bound all OOVs	31.6	35.6	38.0
RO–EN	Baseline	15.8	18.5	20.7
	+ Transliteration OOVs	15.8	**18.7**	**20.8**
	+ Loan OOVs	**16.0**	**18.7**	20.7
	+ Features	**16.0**	18.6	20.6
	Upper bound loan	16.6	19.4	20.9
	Upper bound all OOVs	28.0	28.8	30.4

Table 2.12: BLEU scores in the Swahili–English, Maltese–English, and Romanian–English MT experiments.

Error analysis

Our augmented MT systems combine three main components: the translation system itself, a borrowing system, and a pivot translation system. At each step of the application errors may occur that lead to erroneous translations. To identify main sources of errors in the Swahili–English end-to-end system, we conducted a manual analysis of errors in translations of OOV types produced by the Swahili–English 4K translation systems. As a gold standard corpus we use the Helsinki Corpus of Swahili[23] (Hurskainen, 2004a, HCS). HCS is a morphologically, syntactically, and semantically annotated corpus of about 580K sentences (12.7M tokens). In the corpus 52,351 surface forms (1.5M tokens) are marked as Arabic loanwords. Out of the 3,610 OOV types in the Swahili–English 4K translation systems, 481 word types are annotated in the HCS. We manually annotated these 481 words and identified 353 errors; the remaining 128 words were translated correctly in the end-to-end system. Our analysis reveals the error sources detailed below. In table 2.13 we summarize the statistics of the error sources.

1. **Reachability of the borrowing system.**
   Only 368 out of 481 input words produced loanword candidates. The main reason for the unreachable paths is complex morphology of Swahili OOVs, not taken into account

[23] www.aakkl.helsinki.fi/cameel/corpus/intro.htm

Error source	#	%
Reachability of the borrowing system	113	32.0
Loanword production errors	191	54.1
Arabic–English translation errors	20	5.7
Swahili–English translation errors	29	8.2

Table 2.13: Sources of errors.

by our borrowing system. For example, *atakayehusika* 'who will be involved', the lemma is *husika* 'involve'.

2. **Loanword production errors.**
   About half of errors are due to incorrect outputs of the borrowing system. This is in line with the Arabic–Swahili borrowing system accuracy reported in table 2.6. For example, all morphological variants of the lemma *wahi* 'never' (*hayajawahi*, *halijawahi*, *hazijawahi*), incorrectly produced an Arabic donor word جاوه (*jAwh*) 'java'. Additional examples include all variants of the lemma *saidia* 'help' (*isaidie*, *kimewasaidia*) produced Arabic donor candidates that are variants of the proper name *Saidia*.

3. **Arabic–English translation errors.**
   As the most frequent source of errors in the Arabic–English MT system, we have identified OOV Arabic words. For example, although for the Swahili loanword *awashukuru* 'thank you' the borrowing system correctly produced a plausible donor word وشكور (*w$kwr*) 'and thank you' (rarely used), the only translation variant produced by the Arabic–English MT was *kochkor*.

4. **Swahili–English translation errors.**
   In some cases, although the borrowing system produced a correct donor candidate, and the Arabic–English translation was also correct, translation variants were different from the reference translations in the Swahili–English MT system. For example, the word *alihuzunika* 'he grieved' correctly produced an Arabic donor الحزن (*AlHzn*) 'grief'. Translation variants produced by the Arabic–English MT are *sadness, grief, saddened, sorrow, sad, mourning, grieved, saddening, mourn, distressed*, whereas the expected translation in the Swahili–English reference translations is *disappointed*. Another source of errors that occurred despite correct outputs of borrowing and translation systems is historical meaning change of words. An interesting example of such semantic shift is the word *sakafu* 'floor', that was borrowed from the Arabic word سقف (*sqf*) 'ceiling'.

Complex morphology of both Swahili and Arabic is the most frequent source of errors at all steps of the application. Concatenation of several prefixes in Swahili affects the reachability of the borrowing system. Some Swahili prefixes flip the meaning of words, for example *kutoadhibiwa* ‘impunity’, produces the lemma *adhibiwa* ‘punishment’, and consequently translations *torture, torturing, tortured*. Finally, derivational processes in both languages are not handled by our system, for example, a verb *aliyorithi* ‘he inherited’, produces an Arabic noun الوارثة (*AlwArvp*) ‘the heiress’, and its English translations *heiress*. Jointly reasoning about morphological processes in the donor and recipient languages suggests a possible avenue for remedying these issues.

2.5 Additional Related Work

With the exception of a study conducted by Blair and Ingram (2003) on generation of borrowed phonemes in English–Japanese language pair (the method does not generalize from borrowed phonemes to borrowed words, and does not rely on linguistic insights), we are not aware of any prior work on computational modeling of lexical borrowing. Few papers only mention or tangentially address borrowing, we briefly list them here. Daumé III (2009) focuses on areal effects on linguistic typology, a broader phenomenon that includes borrowing and genetic relations across languages. This study is aimed at discovering language areas based on typological features of languages. Garley and Hockenmaier (2012) train a maximum entropy classifier with character *n*-gram and morphological features to identify anglicisms (which they compare to loanwords) in an online community of German hip hop fans. Finally, List and Moran (2013) have published a toolkit for computational tasks in historical linguistics but remark that “Automatic approaches for borrowing detection are still in their infancy in historical linguistics.”

2.6 Summary

Given a loanword, our model identifies plausible donor words in a contact language. We show that a discriminative model with Optimality Theoretic features effectively models systematic phonological changes in Arabic–Swahili loanwords. We also found that the model and methodology is generally applicable to other language pairs with minimal engineering effort. Our translation results substantially improve over the baseline and confirm that OOV loanwords are important and merit further investigation.

There are numerous research questions that we would like to explore further. Is it possible to monolingually identify borrowed words in a language? Can we automatically identify a donor language (or its phonological properties) for a borrowed word? Since languages may borrow from many sources, can jointly modeling this process lead to better performance? Can we reduce the amount of language-specific engineering required to deploy our model? Can we integrate knowledge of borrowing in additional downstream NLP applications? We intend to address these questions in future work.

Chapter 3

Semantic Knowledge Transfer

This chapter concludes the first part of the thesis on using linguistic knowledge for procuring data and building better generalizations for resource-poor languages. From low-level phonological generalizations studied in the previous chapter, we now proceed to cross-lingual semantics. Languages partition human experiences in different ways. In particular, non-literal expressions, such as idioms and metaphors, vary widely across languages and across cultures, in a way that often makes them non-intelligible to non-native speakers and do not allow their literal translation into other languages. This chapter shows how a careful choice of features informed by linguistic research helps overcome this cross-lingual discrepancy, and to use English models to identify metaphors in low-resource languages. The theory of conceptual metaphor, used to guide model choices, helps obtain robust cross-lingual models, sparing the need for manual annotation efforts in different languages. In the other direction, the chapter results contribute to linguistic research, providing confirmation of the hypotheses stated by the TCM. This research was done in collaboration with Chris Dyer, Leonid Boytsov, Anatole Gershman, and Eric Nyberg, and presented at *ACL 2014* (Tsvetkov *et al.*, 2014c).

3.1 Background

Lakoff and Johnson (1980) characterize metaphor as reasoning about one thing in terms of another, i.e., a metaphor is a type of *conceptual mapping*, where words or phrases are applied to objects and actions in ways that do not permit a literal interpretation. They argue that metaphors play a fundamental communicative role in verbal and written interactions, claiming that much of our everyday language is delivered in metaphorical terms. There is empirical evidence supporting the claim: recent corpus studies have estimated that the proportion of words used metaphorically ranges from 5% to 20% (Steen *et al.*, 2010), and Thibodeau and Boroditsky (2011) provide evidence that a choice of metaphors affects decision making.

Given the prevalence and importance of metaphoric language, effective automatic detection of metaphors would have a number of benefits, both practical and scientific. Language processing applications that need to understand language or preserve meaning (information extraction, machine translation, dialog systems, sentiment analysis, and text analytics, etc.) would have access to a potentially useful high-level bit of information about whether something is to be understood literally or not. Second, scientific hypotheses about metaphoric language could be tested more easily at a larger scale with automation.

However, metaphor detection is a hard problem. On one hand, there is a subjective component: humans may disagree whether a particular expression is used metaphorically or not, as there is no clear-cut semantic distinction between figurative and metaphorical language (Shutova, 2010). On the other, metaphors can be domain- and context-dependent.[1]

Previous work has focused on metaphor identification in English, using both extensive manually-created linguistic resources (Mason, 2004; Gedigian *et al.*, 2006; Krishnakumaran and Zhu, 2007; Turney *et al.*, 2011; Broadwell *et al.*, 2013) and corpus-based approaches (Birke and Sarkar, 2007; Shutova *et al.*, 2013; Neuman *et al.*, 2013; Shutova and Sun, 2013; Hovy *et al.*, 2013). We build on this foundation and also extend metaphor detection into other languages in which few resources may exist. This chapter makes the following contributions: (1) we develop a new state-of-the-art English metaphor detection system that uses *conceptual* semantic features, such as a degree of abstractness and semantic supersenses;[2] (2) we create new metaphor-annotated corpora for Russian and English;[3] (3) using a paradigm of model transfer (McDonald *et al.*, 2011a; Täckström *et al.*, 2013; Kozhenikov and Titov, 2013), we provide support for the hypothesis that metaphors are conceptual (rather than lexical) in nature by showing that our English-trained model can detect metaphors in Spanish, Farsi, and Russian.

3.2 Methodology

Our task in this work is to define features that distinguish between metaphoric and literal uses of two syntactic constructions: subject-verb-object (SVO) and adjective-noun (AN) tuples.[4] We give examples of a prototypical metaphoric usage of each type:

- **SVO metaphors.** A sentence containing a metaphoric SVO relation is *my car drinks gasoline*. According to Wilks (1978), this metaphor represents a violation of selectional preferences for the verb *drink*, which is normally associated with animate subjects (the car is inanimate and, hence, cannot drink in the literal sense of the verb).

[1] For example, *drowning students* could be used metaphorically to describe the situation where students are overwhelmed with work, but in the sentence *a lifeguard saved drowning students*, this phrase is used literally.

[2] `https://github.com/ytsvetko/metaphor`

[3] `http://www.cs.cmu.edu/~ytsvetko/metaphor/datasets.zip`

[4] Our decision to focus on SVO and AN metaphors is justified by corpus studies that estimate that verb- and adjective-based metaphors account for a substantial proportion of all metaphoric expressions, approximately 60% and 24%, respectively (Shutova and Teufel, 2010; Gandy *et al.*, 2013).

- **AN metaphors.** The phrase *broken promise* is an AN metaphor, where attributes from a concrete domain (associated with the concrete word *broken*) are transferred to a more abstract domain, which is represented by the relatively abstract word *promise*. That is, we map an abstract concept *promise* to a concrete domain of physical things, where things can be literally broken to pieces.

Motivated by Lakoff's (1980) argument that metaphors are systematic conceptual mappings, we will use coarse-grained *conceptual*, rather than fine-grained *lexical* features, in our classifier. Conceptual features pertain to concepts and ideas as opposed to individual words or phrases expressed in a particular language. In this sense, as long as two words in two different languages refer to the same concepts, their conceptual features should be the same. Furthermore, we hypothesize that our coarse semantic features give us a language-invariant representation suitable for metaphor detection. To test this hypothesis, we use a cross-lingual model transfer approach: we use bilingual dictionaries to project words from other syntactic constructions found in other languages into English and then apply the English model on the derived conceptual representations.

Each SVO (or AN) instance will be represented by a triple (duple) from which a feature vector will be extracted.[5] The vector will consist of the concatenation of the conceptual features (which we discuss below) for all participating words, and conjunction features for word pairs.[6] For example, to generate the feature vector for the SVO triple $(car, drink, gasoline)$, we compute all the features for the individual words *car*, *drink*, *gasoline* and combine them with the conjunction features for the pairs *car drink* and *drink gasoline*.

We define three main feature categories (1) abstractness and imageability, (2) supersenses, (3) unsupervised vector-space word representations; each category corresponds to a group of features with a common theme and representation.

- **Abstractness and imageability.** Abstractness and imageability were shown to be useful in detection of metaphors (it is easier to invoke mental pictures of concrete and imageable words) (Turney *et al.*, 2011; Broadwell *et al.*, 2013). We expect that abstractness, used in conjunction features (e.g., a feature denoting that the subject is abstract and the verb is concrete), is especially useful: semantically, an abstract agent performing a concrete action is a strong signal of metaphorical usage.

  Although often correlated with abstractness, imageability is not a redundant property. While most abstract things are hard to visualize, some call up images, e.g., *vengeance* calls up an emotional image, *torture* calls up emotions and even visual images. There

[5] Looking at components of the syntactic constructions independent of their context has its limitations, as discussed above with the *drowning students* example; however, it simplifies the representation challenges considerably.

[6] If word one is represented by features $\mathbf{u} \in \mathbb{R}^n$ and word two by features $\mathbf{v} \in \mathbb{R}^m$ then the conjunction feature vector is the vectorization of the outer product $\mathbf{u}\mathbf{v}^\top$.

are concrete things that are hard to visualize too, for example, *abbey* is harder to visualize than *banana* (B. MacWhinney, personal communication).

- **Supersenses.** Supersenses[7] are coarse semantic categories originating in WordNet. For nouns and verbs there are 45 classes: 26 for nouns and 15 for verbs, for example, NOUN.BODY, NOUN.ANIMAL, VERB.CONSUMPTION, or VERB.MOTION (Ciaramita and Altun, 2006; Schneider, 2014). English adjectives do not, as yet, have a similar high-level semantic partitioning in WordNet, thus we constructed a 13-class taxonomy of adjective supersenses (Tsvetkov *et al.*, 2014a) (discussed in §3.3.2, but not included as a contribution of this thesis).

  Supersenses are particularly attractive features for metaphor detection: coarse sense taxonomies can be viewed as semantic concepts, and since concept mapping is a process in which metaphors are born, we expect different supersense co-occurrences in metaphoric and literal combinations. In "drinks gasoline", for example, mapping to supersenses would yield a pair <VERB.CONSUMPTION, NOUN.SUBSTANCE>, contrasted with <VERB.CONSUMPTION, NOUN.FOOD> for "drinks juice". In addition, this coarse semantic categorization is preserved in translation (Schneider *et al.*, 2013), which makes supersense features suitable for cross-lingual approaches such as ours.

- **Vector space word representations.** Vector space word representations learned using unsupervised algorithms are often effective features in supervised learning methods (Turian *et al.*, 2010a). In particular, many such representations are designed to capture lexical semantic properties and are quite effective features in semantic processing, including named entity recognition (Turian *et al.*, 2009), word sense disambiguation (Huang *et al.*, 2012), and lexical entailment (Baroni *et al.*, 2012). In a recent study, Mikolov *et al.* (2013c) reveal an interesting cross-lingual property of distributed word representations: there is a strong similarity between the vector spaces across languages that can be easily captured by linear mapping. Thus, vector space models can also be seen as vectors of (latent) semantic concepts, that preserve their "meaning" across languages.

3.3 Model and Feature Extraction

In this section we describe a classification model, and provide details on mono- and cross-lingual implementation of features.

[7] Supersenses are called "lexicographer categories" in WordNet documentation (Fellbaum, 1998), `http://wordnet.princeton.edu/man/lexnames.5WN.html`

3.3.1 Classification using Random Forests

To make classification decisions, we use a random forest classifier (Breiman, 2001), an ensemble of decision tree classifiers learned from many independent subsamples of the training data. Given an input, each tree classifier assigns a probability to each label; those probabilities are averaged to compute the probability distribution across the ensemble. Random forest ensembles are particularly suitable for our resource-scarce scenario: rather than overfitting, they produce a limiting value of the generalization error as the number of trees increases,[8] and no hyperparameter tuning is required. In addition, decision-tree classifiers learn non-linear responses to inputs and often outperform logistic regression (Perlich *et al.*, 2003).[9] Our random forest classifier models the probability that the input syntactic relation is metaphorical. If this probability is above a threshold, the relation is classified as metaphoric, otherwise it is literal. We used the `scikit-learn` toolkit to train our classifiers (Pedregosa *et al.*, 2011).

3.3.2 Feature extraction

Abstractness and imageability. The MRC psycholinguistic database is a large dictionary listing linguistic and psycholinguistic attributes obtained experimentally (Wilson, 1988).[10] It includes, among other data, 4,295 words rated by the degrees of abstractness and 1,156 words rated by the imageability. Similarly to Tsvetkov *et al.* (2013a), we use a logistic regression classifier to propagate abstractness and imageability scores from MRC ratings to all words for which we have vector space representations. More specifically, we calculate the degree of abstractness and imageability of all English items that have a vector space representation, using vector elements as features. We train two separate classifiers for abstractness and imageability on a seed set of words from the MRC database. Degrees of abstractness and imageability are posterior probabilities of classifier predictions. We binarize these posteriors into abstract-concrete (or imageable-unimageable) boolean indicators using pre-defined thresholds.[11] Performance of these classifiers, tested on a sampled held-out data, is 0.94 and 0.85 for the abstractness and imageability classifiers, respectively.

Supersenses. In the case of SVO relations, we incorporate supersense features for nouns and verbs; noun and adjective supersenses are used in the case of AN relations.

Supersenses of nouns and verbs. A lexical item can belong to several synsets, which are associated with different supersenses. Degrees of membership in different supersenses are

[8] See Theorem 1.2 in (Breiman, 2001) for details.

[9] In our experiments, random forests model slightly outperformed logistic regression and SVM classifiers.

[10] `http://ota.oucs.ox.ac.uk/headers/1054.xml`

[11] Thresholds are equal to 0.8 for abstractness and to 0.9 for imageability. They were chosen empirically based on accuracy during cross-validation.

represented by feature vectors, where each element corresponds to one supersense. For example, the word *head* (when used as a noun) participates in 33 synsets, three of which are related to the supersense NOUN.BODY. The value of the feature corresponding to this supersense is $3/33 \approx 0.09$.

Supersenses of adjectives. English WordNet offers a fine-grained inventory of semantic senses for adjectives. Like nouns, verbs, and adverbs, these are organized into *synsets* (synonym sets). Unlike nouns and verbs, however, there is no hierarchical taxonomy for adjectives; instead, adjective synsets are organized in **clusters** consisting of a core synset and linked satellite synsets with closely related meanings (Gross and Miller, 1990).[12] Members of these clusters are sometimes linked to nouns or verbs, or to other clusters via "see also" or antonymy links, but there is no systematic organization connecting these clusters. For example, *exasperated* and *cheesed off* are listed as synonyms and *displeased* as closely related, but there is nothing to indicate that these as well as *ashamed* all describe emotional states.

In Tsvetkov *et al.* (2014a) we presented an approach to eliciting high-level groupings of adjective synsets into a small number of coarse classes. Inspired by WordNet's partitioning of nouns and verbs into supersenses, we borrow and adapt to English the top-level adjectival classification scheme from GermaNet (i.e., the German-language WordNet; Hamp and Feldweg, 1997). We use 13 top-level classes from the adapted taxonomy of GermaNet. For example, the top-level classes in GermaNet include: ADJ.FEELING (e.g., willing, pleasant, cheerful); ADJ.SUBSTANCE (e.g., dry, ripe, creamy); ADJ.SPATIAL (e.g., adjacent, gigantic).[13] For each adjective type in WordNet, we produce a vector with a classifier posterior probabilities corresponding to degrees of membership of this word in one of the 13 semantic classes,[14] similar to the feature vectors we build for nouns and verbs. For example, for a word *calm* the top-2 categories (with the first and second highest degrees of membership) are ADJ.BEHAVIOR and ADJ.FEELING.

Vector space word representations. We employ 64-dimensional vector-space word representations constructed by Faruqui and Dyer (2014b).[15] Vector construction algorithm is a variation on traditional latent semantic analysis (Deerwester *et al.*, 1990) that uses multilingual information to produce representations in which synonymous words have similar vectors. The vectors were trained on the news commentary corpus released by WMT-2011,[16] comprising 180,834 types.

[12] There are 18,156 adjective synsets in WordNet (7,463 main synsets and 10,693 "satellite" synsets representing variations on main synsets (mapped with a "similar to" link). The lemmas in these synsets capture 21,479 adjective types, 4,993 of which are polysemous.

[13] For the full taxonomy see `http://www.sfs.uni-tuebingen.de/lsd/adjectives.shtml`

[14] `http://www.cs.cmu.edu/~ytsvetko/adj-supersenses.tar.gz`

[15] `http://www.cs.cmu.edu/~mfaruqui/soft.html`

[16] `http://www.statmt.org/wmt11/`

3.3.3 Cross-lingual feature projection

For languages other than English, feature vectors are projected to English features using translation dictionaries. We used the Babylon dictionary,[17] which is a proprietary resource, but any bilingual dictionary can in principle be used. For a non-English word in a source language, we first obtain all translations into English. Then, we average all feature vectors related to these translations. Consider an example related to projection of WordNet supersenses. A Russian word голова is translated as *head* and *brain*. Hence, we select all the synsets of the nouns *head* and *brain*. There are 38 such synsets (33 for *head* and 5 for *brain*). Four of these synsets are associated with the supersense NOUN.BODY. Therefore, the value of the feature NOUN.BODY is $4/38 \approx 0.11$.

3.4 Datasets

In this section we describe a training and testing dataset as well a data collection procedure.

3.4.1 English training sets

To train an SVO metaphor classifier, we employ the TroFi (Trope Finder) dataset.[18] TroFi includes 3,737 manually annotated English sentences from the *Wall Street Journal* (Birke and Sarkar, 2007). Each sentence contains either literal or metaphorical use for one of 50 English verbs. First, we use a dependency parser (Martins *et al.*, 2010) to extract subject-verb-object (SVO) relations. Then, we filter extracted relations to eliminate parsing-related errors, and relations with verbs which are not in the TroFi verb list. After filtering, there are 953 metaphorical and 656 literal SVO relations which we use as a training set.

In the case of AN relations, we construct and make publicly available a training set containing 884 metaphorical AN pairs and 884 pairs with literal meaning. It was collected by two annotators using public resources (collections of metaphors on the web). At least one additional person carefully examined and culled the collected metaphors, by removing duplicates, weak metaphors, and metaphorical phrases (such as *drowning students*) whose interpretation depends on the context.

3.4.2 Multilingual test sets

We collect and annotate metaphoric and literal test sentences in four languages. Thus, we compile eight test datasets, four for SVO relations, and four for AN relations. Each dataset has an equal number of metaphors and non-metaphors, i.e., the datasets are balanced. English

[17] `http://www.babylon.com`

[18] `http://www.cs.sfu.ca/~anoop/students/jbirke/`

(EN) and Russian (RU) datasets have been compiled by our team and are publicly available. Spanish (ES) and Farsi (FA) datasets are published elsewhere (Levin *et al.*, 2014). Table 3.1 lists test set sizes.

	SVO	AN
EN	222	200
RU	240	200
ES	220	120
FA	44	320

Table 3.1: Sizes of the eight test sets. Each dataset is balanced, i.e., it has an equal number of metaphors and non-metaphors. For example, English SVO dataset has 222 relations: 111 metaphoric and 111 literal.

We used the following procedure to compile the EN and RU test sets. A moderator started with seed lists of 1000 most common verbs and adjectives.[19]

Then she used the SketchEngine, which provides searching capability for the TenTen Web corpus,[20] to extract sentences with words that frequently co-occurred with words from the seed lists. From these sentences, she removed sentences that contained more than one metaphor, and sentences with non-SVO and non-AN metaphors. Remaining sentences were annotated by several native speakers (five for English and six for Russian), who judged AN and SVO phrases in context. The annotation instructions were general: *"Please, mark in bold all words that, in your opinion, are used non-literally in the following sentences. In many sentences, all the words may be used literally."* The Fleiss' Kappas for 5 English and 6 Russian annotators are: EN-AN = .76, RU-AN = .85, EN-SVO = .75, RU-SVO = .78. For the final selection, we filtered out low-agreement (<.8) sentences.

The test candidate sentences were selected by a person who did not participate in the selection of the training samples. No English annotators of the test set, and only one Russian annotator out of 6 participated in the selection of the training samples. Thus, we trust that annotator judgments were not biased towards the cases that the system is trained to process.

3.5 Experiments

3.5.1 English experiments

Our task, as defined in Section 3.2, is to classify SVO and AN relations as either metaphoric or literal. We first conduct a 10-fold cross-validation experiment on the training set defined

[19] Selection of 1000 most common verbs and adjectives achieves much broader lexical and domain coverage than what can be realistically obtained from continuous text. Our test sentence domains are, therefore, diverse: economic, political, sports, etc.

[20] http://trac.sketchengine.co.uk/wiki/Corpora/enTenTen

in Section 3.4.1. We represent each candidate relation using the features described in Section 3.3.2, and evaluate performance of the three feature categories and their combinations. This is done by computing an accuracy in the 10-fold cross validation. Experimental results are given in Table 3.2, where we also provide the number of features in each feature set.

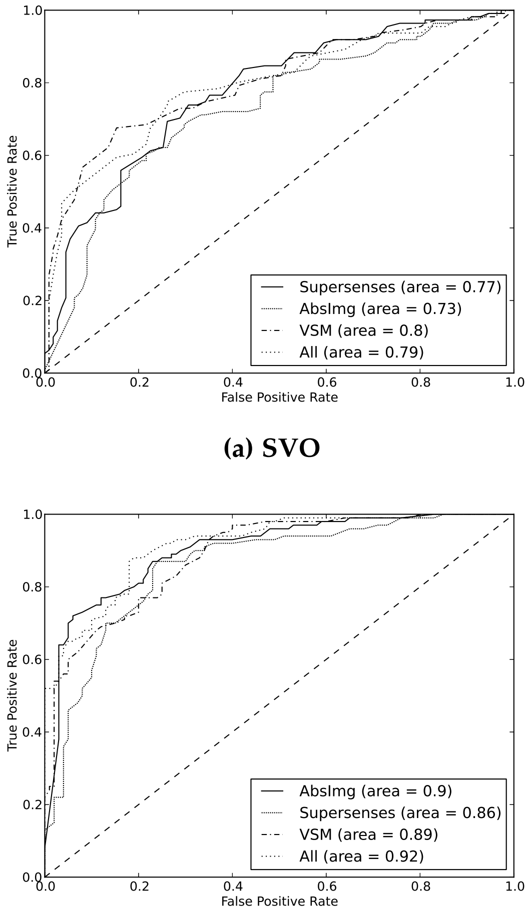
	SVO		AN	
	# FEAT	ACC	# FEAT	ACC
AbsImg	20	0.73*	16	0.76*
Supersense	67	0.77*	116	0.79*
AbsImg+Sup.	87	0.78*	132	0.80*
VSM	192	0.81	228	0.84*
All	279	**0.82**	360	**0.86**

Table 3.2: 10-fold cross validation results for three feature categories and their combination, for classifiers trained on English SVO and AN training sets. # FEAT column shows a number of features. ACC column reports an accuracy score in the 10-fold cross validation. Statistically significant differences ($p < 0.01$) from the all-feature combination are marked with a star.

These results show superior performance over previous state-of-the-art results, confirming our hypothesis that conceptual features are effective in metaphor classification. For the SVO task, the cross-validation accuracy is about 10% better than that of Tsvetkov *et al.* (2013a). For the AN task, the cross validation accuracy is better by 8% than the result of Turney *et al.* (2011) (two baseline methods are described in Section 3.5.2). We can see that all types of features have good performance on their own (VSM is the strongest feature type). Noun supersense features alone allow us to achieve an accuracy of 75%, i.e., adjective supersense features contribute 4% to adjective-noun supersense feature combination. Experiments with the pairs of features yield better results than individual features, implying that the feature categories are not redundant. Yet, combining all features leads to even higher accuracy during cross-validation. In the case of the AN task, a difference between the All feature combination and any other combination of features listed in Table 3.2 is statistically significant ($p < 0.01$ for both the sign and the permutation test).

Although the first experiment shows very high scores, the 10-fold cross-validation cannot fully reflect the generality of the model, because all folds are parts of the same corpus. They are collected by the same human judges and belong to the same domain. Therefore, experiments on out-of-domain data are crucial. We carry out such experiments using held-out SVO and AN EN test sets, described in Section 3.4.2 and Table 3.1. In this experiment, we measure the f-score. We classify SVO and AN relations using a classifier trained on the All feature combination and balanced thresholds. The values of the f-score are 0.76, both for SVO and AN tasks. This out-of-domain experiment suggests that our classifier is portable across domains and genres.

However, (1) different application may have different requirements for recall/precision, and (2) classification results may be skewed towards having high precision and low recall (or vice versa). It is possible to trade precision for recall by choosing a different threshold. Thus, in addition to giving a single f-score value for balanced thresholds, we present a Receiver Operator Characteristic (ROC) curve, where we plot a fraction of true positives against the fraction of false positives for 100 threshold values in the range from zero to one. The area under the ROC curve (AUC) can be interpreted as the probability that a classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example.[21] For a randomly guessing classifier, the ROC curve is a dashed diagonal line. A bad classifier has an ROC curve that goes close to the dashed diagonal or even below it.



(a) SVO

(b) AN

Figure 3.1: ROC curves for classifiers trained using different feature sets (English SVO and AN test sets).

According to ROC plots in Figure 3.1, all three feature sets are effective, both for SVO and for AN tasks. Abstractness and Imageability features work better for adjectives and nouns,

[21] Assuming that positive examples are labeled by ones, and negative examples are labeled by zeros.

which is in line with previous findings (Turney *et al.*, 2011; Broadwell *et al.*, 2013). It can be also seen that VSM features are very effective. This is in line with results of Hovy *et al.* (2013), who found that it is hard to improve over the classifier that uses only VSM features.

3.5.2 Comparison to baselines

In this section, we compare our method to state-of-the-art methods of Tsvetkov *et al.* (2013a) and of Turney *et al.* (2011), who focused on classifying SVO and AN relations, respectively.

In the case of SVO relations, we use software and datasets from Tsvetkov *et al.* (2013a). These datasets, denoted as an SVO-baseline, consist of 98 English and 149 Russian sentences. We train SVO metaphor detection tools on SVO relations extracted from TroFi sentences and evaluate them on the SVO-baseline dataset. We also use the same thresholds for classifier posterior probabilities as Tsvetkov *et al.* (2013a). Our approach is different from that of Tsvetkov *et al.* (2013a) in that it uses additional features (vector space word representations) and a different classification method (we use random forests while Tsvetkov *et al.* (2013a) use logistic regression). According to Table 3.3, we obtain higher performance scores for both Russian and English.

	EN	RU
SVO-baseline	0.78	0.76
This work	0.86	0.85

Table 3.3: Comparing f-scores of our SVO metaphor detection method to the baselines.

In the case of AN relations, we use the dataset (denoted as an AN-baseline) created by Turney *et al.* (2011) (see Section 4.1 in the referred paper for details). Turney *et al.* (2011) manually annotated 100 pairs where an adjective was one of the following: *dark*, *deep*, *hard*, *sweet*, and *worm*. The pairs were presented to five human judges who rated each pair on a scale from 1 (very literal/denotative) to 4 (very non-literal/connotative). Turney *et al.* (2011) train logistic-regression employing only abstractness ratings as features. Performance of the method was evaluated using the 10-fold cross-validation separately for each judge.

We replicate the above described evaluation procedure of Turney *et al.* (2011) using their model and features. In our classifier, we use the All feature combination and the balanced threshold as described in Section 3.5.1.

According to results in Table 3.4, almost all of the judge-specific f-scores are slightly higher for our system, as well as the overall average f-score.

In both baseline comparisons, we obtain performance at least as good as in previously published studies.

	AN-baseline	This work
Judge 1	0.73	0.75
Judge 2	0.81	0.84
Judge 3	0.84	0.88
Judge 4	0.79	0.81
Judge 5	0.78	0.77
average	0.79	0.81

Table 3.4: Comparing AN metaphor detection method to the baselines: accuracy of the 10-fold cross validation on annotations of five human judges.

3.5.3 Cross-lingual experiments

In the next experiment we corroborate the main hypothesis of this chapter: a model trained on English data can be successfully applied to other languages. Namely, we use a trained English model discussed in Section 3.5.1 to classify literal and metaphoric SVO and AN relations in English, Spanish, Farsi and Russian test sets, listed in Section 3.4.2. This time we used all available features.

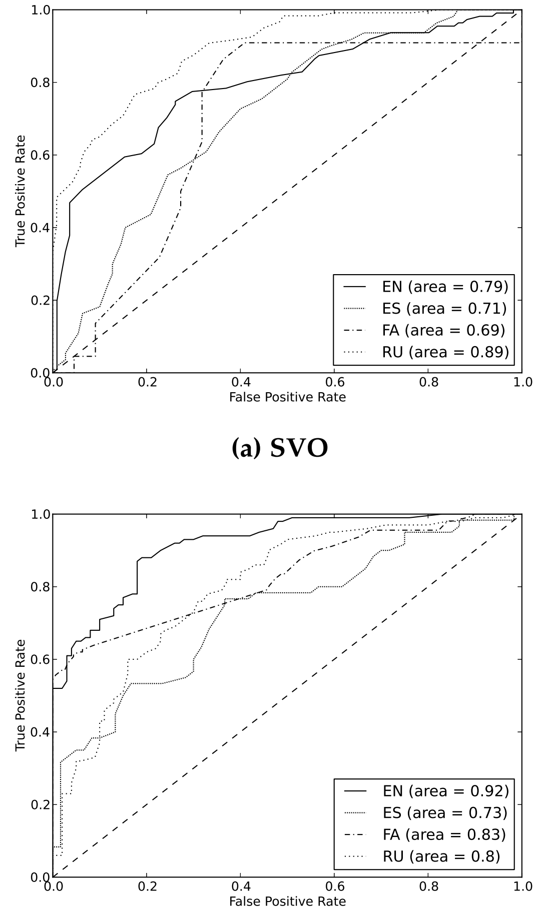
Experimental results for all four languages, are given in Figure 3.2. The ROC curves for SVO and AN tasks are plotted in Figure 3.2a and Figure 3.2b, respectively. Each curve corresponds to a test set described in Table 3.1. In addition, we perform an oracle experiment, to obtain actual f-score values for best thresholds. Detailed results are shown in Table 3.5.

Consistent results with high f-scores are obtained across all four languages. Note that higher scores are obtained for the Russian test set. We hypothesize that this happens due to a higher-quality translation dictionary (which allows a more accurate model transfer). Relatively lower (yet reasonable) results for Farsi can be explained by a smaller size of the bilingual dictionary (thus, fewer feature projections can be obtained). Also note that, in our experience, most of Farsi metaphors are adjective-noun constructions. This is why the AN FA dataset in Table 3.1 is significantly larger than SVO FA. In that, for the AN Farsi task we observe high performance scores.

	SVO	AN
EN	0.79	0.85
RU	0.84	0.77
ES	0.76	0.72
FA	0.75	0.74

Table 3.5: Cross-lingual experiment: f-scores for classifiers trained on the English data using a combination of all features, and applied, with optimal thresholds, to SVO and AN metaphoric and literal relations in four test languages: English, Russian, Spanish, and Farsi.

Figure 3.2 and Table 3.5 confirm, that *we obtain similar, robust results on four very different*



(a) SVO

(b) AN

Figure 3.2: Cross-lingual experiment: ROC curves for classifiers trained on the English data using a combination of all features, and applied to SVO and AN metaphoric and literal relations in four test languages: English, Russian, Spanish, and Farsi.

languages, using the same English classifiers. We view this result as a strong evidence of language-independent nature of our metaphor detection method. In particular, this shows that proposed conceptual features can be used to detect selectional preferences violation across languages.

To summarize the experimental section, our metaphor detection approach obtains state-of-the-art performance in English, is effective when applied to out-of-domain English data, and works cross-lingually.

3.5.4 Examples

Manual data analysis on adjective-noun pairs supports an abstractness-concreteness hypothesis formulated by several independent research studies. For example, in English we classify as metaphoric *dirty word* and *cloudy future*. Word pairs *dirty diaper* and *cloudy weather* have same adjectives. Yet they are classified as literal. Indeed, *diaper* is a more concrete term than *word* and *weather* is more concrete than *future*. Same pattern is observed in non-English datasets. In Russian, больное общество "sick society" and пустой звук "empty sound" are classified as metaphoric, while больная бабушка "sick grandmother" and пустая чашка "empty cup" are classified as literal. Spanish example of an adjective-noun metaphor is a well-known músculo económico "economic muscle". We also observe that non-metaphoric adjective noun pairs tend to have more imageable adjectives, such as literal derecho humano "human right". In Spanish, *human* is more imageable than *economic*.

Verb-based examples that are correctly classified by our model are: *blunder escaped notice* (metaphoric) and *prisoner escaped jail* (literal). We hypothesize that supersense features are instrumental in the correct classification of these examples: <NOUN.PERSON,VERB.MOTION> is usually used literally, while <NOUN.ACT,VERB.MOTION> is used metaphorically.

3.6 Related Work

For a historic overview and a survey of common approaches to metaphor detection, we refer the reader to recent reviews by Shutova et al. (Shutova, 2010; Shutova *et al.*, 2013). Here we focus only on recent approaches.

Shutova et al. (2010) proposed a bottom-up method: one starts from a set of seed metaphors and seeks phrases where verbs and/or nouns belong to the same cluster as verbs or nouns in seed examples.

Turney et al. (2011) show how abstractness scores could be used to detect metaphorical AN phrases. Neuman et al. (2013) describe a Concrete Category Overlap algorithm, where co-occurrence statistics and Turney's abstractness scores are used to determine WordNet supersenses that correspond to literal usage of a given adjective or verb. For example, given an adjective, we can learn that it modifies concrete nouns that usually have the supersense NOUN.BODY. If this adjective modifies a noun with the supersense NOUN.FEELING, we conclude that a metaphor is found.

Broadwell *et al.* (2013) argue that metaphors are highly imageable words that do not belong to a discussion topic. To implement this idea, they extend MRC imageability scores to all dictionary words using links among WordNet supersenses (mostly hypernym and hyponym relations). Strzalkowski *et al.* (2013) carry out experiments in a specific (government-related) domain for four languages: English, Spanish, Farsi, and Russian. Strzalkowski *et al.* (2013) explain the algorithm only for English and say that is the same for Spanish, Farsi, and Russian. Because they heavily rely on WordNet and availability of imageability scores, their approach may not be applicable to low-resource languages.

Hovy *et al.* (2013) applied tree kernels to metaphor detection. Their method also employs WordNet supersenses, but it is not clear from the description whether WordNet is essential or can be replaced with some other lexical resource. We cannot compare directly our model with this work because our classifier is restricted to detection of only SVO and AN metaphors.

Tsvetkov *et al.* (2013a) propose a cross-lingual detection method that uses only English lexical resources and a dependency parser. Their study focuses only on the verb-based metaphors. Tsvetkov *et al.* (2013a) employ only English and Russian data. Current work builds on this study, and incorporates new syntactic relations as metaphor candidates, adds several new feature sets and different, more reliable datasets for evaluating results. We demonstrate results on two new languages, Spanish and Farsi, to emphasize the generality of the method.

A recent book on metaphor processing (Veale *et al.*, 2016) summarizes comprehensively additional related work, including research done after our study.

Word sense disambiguation (WSD) is a related problem, where one identifies meanings of polysemous words. The difference is that in the WSD task, we need to select an already existing sense, while for the metaphor detection, the goal is to identify cases of sense borrowing. Studies showed that cross-lingual evidence allows one to achieve a state-of-the-art performance in the WSD task, yet, most cross-lingual WSD methods employ parallel corpora (Navigli, 2009).

3.7 Summary

The key contribution of this chapter is that we show how to identify metaphors across languages by building a model in English and applying it—without adaptation—to other languages: Spanish, Farsi, and Russian. This model uses language-independent (rather than lexical or language specific) conceptual features, motivated by the Conceptual Metaphor Theory. Not only do we establish benchmarks for Spanish, Farsi, and Russian, but we also achieve state-of-the-art performance in English. In addition, we present a comparison of relative contributions of several types of features. We concentrate on metaphors in the context of two kinds of syntactic relations: subject-verb-object (SVO) relations and adjective-noun (AN) relations, which account for a majority of all metaphorical phrases.

Future work will expand the scope of metaphor identification by including nominal metaphoric relations as well as explore techniques for incorporating contextual features, which can play a key role in identifying certain kinds of metaphors. Second, cross-lingual model transfer can be improved with more careful cross-lingual feature projection.

Part II

Linguistic Knowledge in Resource-Rich NLP: Understanding and Integrating Linguistic Knowledge in Distributed Representations

Chapter 4

Linguistic Knowledge in Training Data

Previous two chapters focused on low-data problems emerging either because of cost limits or because of inherent limits on the kind of data that would apply. I now introduce a sequence of three chapters targeting problems in which data is not a limiting factor; the chapters are on interpreting and integrating linguistic knowledge in neural NLP for learning improved, linguistically-informed generalizations. This chapter focuses on optimizing the content, structure, and order of training data for constructing task-specific distributed representations of words. Alongside its practical contribution—improvement of downstream tasks with task-specific representations—this work facilitates analysis and better understanding of desired characteristics of training data for a task. Work described in this chapter is an extended version of the *ACL 2016* publication (Tsvetkov *et al.*, 2016b), conducted in collaboration with Chris Dyer, Manaal Faruqui, Wang Ling, and Brian MacWhinney.

4.1 Background

It is well established that in language acquisition, there are robust patterns in the order by which phenomena are acquired. For example, prototypical concepts are acquired earlier; concrete words tend to be learned before abstract ones (Rosch, 1978). The acquisition of lexical knowledge in artificial systems proceeds differently. In general, models will improve during the course of parameter learning, but the time course of acquisition is not generally studied beyond generalization error as a function of training time or data size. We revisit this issue of choosing the order of learning—**curriculum learning**—framing it as an optimization problem so that a rich array of factors—including nuanced measures of difficulty, as well as prototypicality and diversity—can be exploited.

Prior research focusing on curriculum strategies in NLP is scarce, and has conventionally been following a paradigm of "starting small" (Elman, 1993), i.e., initializing the learner with "simple" examples first, and then gradually increasing data complexity (Bengio *et al.*, 2009;

Spitkovsky *et al.*, 2010). In language modeling, this preference for increasing complexity has been realized by curricula that increase the entropy of training data by growing the size of the training vocabulary from frequent to less frequent words (Bengio *et al.*, 2009). In unsupervised grammar induction, an effective curriculum comes from increasing length of training sentences as training progresses (Spitkovsky *et al.*, 2010). These case studies have demonstrated that carefully designed curricula can lead to better results. However, they have relied on heuristics in selecting curricula or have followed the intuitions of human and animal learning (Kail, 1990; Skinner, 1938). Had different heuristics been chosen, the results would have been different. In this chapter, we use curriculum learning to create improved word representations. However, rather than testing a small number of curricula, we search for an optimal curriculum using Bayesian optimization. A curriculum is defined to be the ordering of the training instances, in our case it is the ordering of paragraphs in which the representation learning model reads the corpus. We use a linear ranking function to conduct a systematic exploration of interacting factors that affect curricula of representation learning models. We then analyze our findings, and compare them to human intuitions and learning principles.

We treat curriculum learning as an outer loop in the process of learning and evaluation of vector-space representations of words; the iterative procedure is (1) predict a curriculum; (2) train word embeddings; (3) evaluate the embeddings on tasks that use word embeddings as the sole features. Through this model we analyze the impact of curriculum on word representation models and on extrinsic tasks. To quantify curriculum properties, we define three groups of features aimed at analyzing statistical and linguistic content and structure of training data: (1) diversity, (2) simplicity, and (3) prototypicality. A function of these features is computed to score each paragraph in the training data, and the curriculum is determined by sorting corpus paragraphs by the paragraph scores. We detail the model in §4.2. Word vectors are learned from the sorted corpus, and then evaluated on part-of-speech tagging, parsing, named entity recognition, and sentiment analysis (§4.3). Our experiments confirm that training data curriculum affects model performance, and that models with optimized curriculum consistently outperform baselines trained on shuffled corpora (§4.4). We analyze our findings in §4.5.

The contributions of this chapter are twofold. First, this is the first framework that formulates curriculum learning as an optimization problem, rather than shuffling data or relying on human intuitions. We experiment with optimizing the curriculum of word embeddings, but in principle the curriculum of other models can be optimized in a similar way. Second, to the best of our knowledge, this study is the first to analyze the impact of distributional and linguistic properties of training texts on the quality of task-specific word embeddings.

4.2 Curriculum Learning Model

We are considering the problem of maximizing a performance of an NLP task through sequentially optimizing the curriculum of training data of word vector representations that are used as features in the task.

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ be the training corpus with n lines (sentences or paragraphs). The curriculum of word representations is quantified by scoring each of the paragraphs according to the linear function $\mathbf{w}^\intercal \boldsymbol{\phi}(\mathcal{X})$, where $\boldsymbol{\phi}(\mathcal{X}) \in \mathbb{R}^{\ell \times 1}$ is a real-valued vector containing ℓ linguistic features extracted for each paragraph, and $\mathbf{w} \in \mathbb{R}^{\ell \times 1}$ denote the weights learned for these features. The feature values $\boldsymbol{\phi}(\mathcal{X})$ are z-normalized across all paragraphs. These scores are used to specify the order of the paragraphs in the corpus—the curriculum: we sort the paragraphs by their scores.

After the paragraphs are curriculum-ordered, the reordered training corpus is used to generate word representations. These word representations are then used as features in a subsequent NLP task. We define the objective function $eval : \mathcal{X} \to \mathbb{R}$, which is the quality estimation metric for this NLP task performed on a held-out dataset (e.g., correlation, accuracy, F_1 score, BLEU). Our goal is to define the features $\boldsymbol{\phi}(\mathcal{X})$ and to find the optimal weights $\mathbf{w}$ that maximize $eval$.

We optimize the feature weights using Bayesian optimization; we detail the model in §4.2.1. Distributional and linguistic features inspired by prior research in language acquisition and second language learning are described in §4.2.2. Figure 4.1 shows the computation flow diagram.

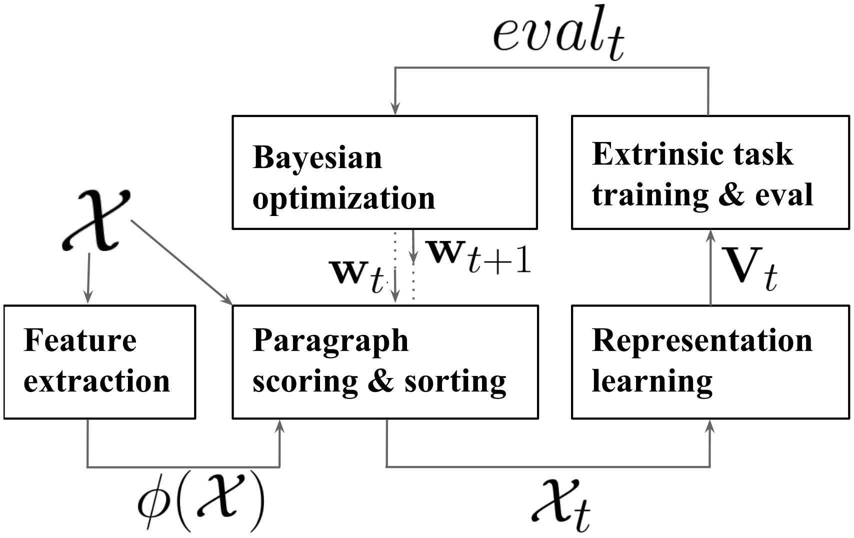


Figure 4.1: Curriculum optimization framework.

4.2.1 Bayesian optimization for curriculum learning

As no assumptions are made regarding the form of $eval(\mathbf{w})$, gradient-based methods cannot be applied, and performing a grid search over parameterizations of $\mathbf{w}$ would require a exponentially growing number of parameterizations to be traversed. Thus, we propose to use Bayesian Optimization (BayesOpt) as the means to maximize $eval(\mathbf{w})$. BayesOpt is a methodology to globally optimize *expensive*, multimodal black-box functions (Shahriari *et al.*, 2016; Bergstra *et al.*, 2011; Snoek *et al.*, 2012). It can be viewed as a sequential approach to performing a regression from high-level model parameters (e.g., learning rate, number of layers in a neural network, and in our model–curriculum weights $\mathbf{w}$) to the loss function or the performance measure (*eval*).

An arbitrary objective function, *eval*, is treated as a black-box, and BayesOpt uses Bayesian inference to characterize a posterior distribution over functions that approximate *eval*. This model of *eval* is called the **surrogate model**. Then, the BayesOpt exploits this model to make decisions about *eval*, e.g., where is the expected maximum of the function, and what is the expected improvement that can be obtained over the best iteration so far. The strategy function, estimating the next set of parameters to explore given the current beliefs about *eval* is called the **acquisition function**. The surrogate model and the acquisition function are the two key components in the BayesOpt framework; their interaction is shown in Algorithm 1.

The surrogate model allows us to cheaply approximate the quality of a set of parameters $\mathbf{w}$ without running $eval(\mathbf{w})$, and the acquisition function uses this surrogate to choose a new value of $\mathbf{w}$. However, a trade-off must be made: should the acquisition function move $\mathbf{w}$ into a region where the surrogate believes an optimal value will be found, or should it explore regions of the space that reveal more about how *eval* behaves, perhaps discovering even better values? That is, acquisition functions balance a tradeoff between exploration—by selecting $\mathbf{w}$ in the regions where the uncertainty of the surrogate model is high, and exploitation—by querying the regions where the model prediction is high.

Popular choices for the surrogate model are Gaussian Processes (Rasmussen, 2006; Snoek *et al.*, 2012, GP), providing convenient and powerful prior distribution on functions, and tree-structured Parzen estimators (Bergstra *et al.*, 2011, TPE), tailored to handle conditional spaces. Choices of the acquisition functions include probability of improvement (Kushner, 1964), expected improvement (EI) (Močkus *et al.*, 1978; Jones, 2001), GP upper confidence bound (Srinivas *et al.*, 2010), Thompson sampling (Thompson, 1933), entropy search (Hennig and Schuler, 2012), and dynamic combinations of the above functions (Hoffman *et al.*, 2011); see Shahriari *et al.* (2016) for an extensive comparison. Yogatama *et al.* (2015) found that the combination of EI as the acquisition function and TPE as the surrogate model performed favorably in Bayesian optimization of text representations; we follow this choice in our model.

Algorithm 1 Bayesian optimization

```
$\mathcal{H} \leftarrow \emptyset$                                   ▷ Initialize observation history
$\mathcal{A} \leftarrow EI$                                  ▷ Initialize acquisition function
$\mathcal{S}_0 \leftarrow TPE$                                    ▷ Initialize surrogate model
for t ← 1 to $T$ do
    $\mathbf{w}_t \leftarrow \text{argmax}_{\mathbf{w}} \mathcal{A}(\mathbf{w}; \mathcal{S}_{t-1}, \mathcal{H})$    ▷ Predict $\mathbf{w}_t$ by optimizing acquisition function
    $eval(\mathbf{w}_t)$                              ▷ Evaluate $\mathbf{w}_t$ on extrinsic task
    $\mathcal{H} \leftarrow \mathcal{H} \cup (\mathbf{w}_t, eval(\mathbf{w}_t))$             ▷ Update observation history
    Estimate $\mathcal{S}_t$ given $\mathcal{H}$
end for
return $\mathcal{H}$
```

4.2.2 Distributional and linguistic features

To characterize and quantify a curriculum, we define three categories of features, focusing on various distributional, syntactic, and semantic aspects of training data. We now detail the feature categories along with motivations for feature selection.

Diversity. Diversity measures capture the distributions of types in data. Entropy is the best-known measure of diversity in statistical research, but there are many others (Tang *et al.*, 2006; Gimpel *et al.*, 2013). Common measures of diversity are used in many contrasting fields, from ecology and biology (Rosenzweig, 1995; Magurran, 2013), to economics and social studies (Stirling, 2007). Diversity has been shown effective in related research on curriculum learning in language modeling, vision, and multimedia analysis (Bengio *et al.*, 2009; Jiang *et al.*, 2014).

Let p_i and p_j correspond to empirical frequencies of word types t_i and t_j in the training data. Let d_{ij} correspond to their semantic similarity, calculated as the cosine similarity between embeddings of t_i and t_j learned from the training data. We annotate each paragraph with the following diversity features:

- Number of word types: $\#types$
- Type-token ratio: $\frac{\#types}{\#tokens}$
- Entropy: $-\sum_i p_i \ln(p_i)$
- Simpson's index (Simpson, 1949): $\sum_i {p_i}^2$
- Quadratic entropy (Rao, 1982):[1] $\sum_{i,j} d_{ij} p_i p_j$

Simplicity. Spitkovsky *et al.* (2010) have validated the utility of syntactic simplicity in curriculum learning for unsupervised grammar induction by showing that training on sentences

[1] Intuitively, this feature promotes paragraphs that contain semantically similar high-probability words.

in order of increasing lengths outperformed other orderings. We explore the simplicity hypothesis, albeit without prior assumptions on specific ordering of data, and extend it to additional simplicity/complexity measures of training data. Our features are inspired by prior research in second language acquisition, text simplification, and readability assessment (Schwarm and Ostendorf, 2005; Heilman *et al.*, 2007; Pitler and Nenkova, 2008; Vajjala and Meurers, 2012). We use an off-the-shelf syntactic parser[2] (Zhang and Clark, 2011) to parse our training corpus. Then, the following features are used to measure phonological, lexical, and syntactic complexity of training paragraphs:

- Language model score
- Character language model score
- Average sentence length
- Verb-token ratio
- Noun-token ratio
- Parse tree depth
- Number of noun phrases: $\#NPs$
- Number of verb phrases: $\#VBs$
- Number of prepositional phrases: $\#PPs$

Prototypicality. This is a group of semantic features that use insights from cognitive linguistics and child language acquisition. The goal is to characterize the curriculum of representation learning in terms of the curriculum of human language learning. We resort to the Prototype theory (Rosch, 1978), which posits that semantic categories include more central (or prototypical) as well as less prototypical words. For example, in the ANIMAL category, *dog* is more prototypical than *sloth* (because *dog* is more frequent); *dog* is more prototypical than *canine* (because *dog* is more concrete); and *dog* is more prototypical than *bull terrier* (because *dog* is less specific). According to the theory, more prototypical words are acquired earlier. We use lexical semantic databases to operationalize insights from the prototype theory in the following semantic features; the features are computed on token level and averaged over paragraphs:

- Age of acquisition (AoA) of words was extracted from the crowd-sourced database, containing over 50 thousand English words (Kuperman *et al.*, 2012). For example, the AoA of *run* is 4.47 (years), of *flee* is 8.33, and of *abscond* is 13.36. If a word was not found in the database it was assigned the maximal age of 25.

[2] `http://http://people.sutd.edu.sg/~yue_zhang/doc`

- Concreteness ratings on the scale of 1–5 (1 is most abstract) for 40 thousand English lemmas (Brysbaert *et al.*, 2014). For example, *cookie* is rated as 5, and *spirituality* as 1.07.
- Imageability ratings are taken from the MRC psycholinguistic database (Wilson, 1988). Following §3 (Tsvetkov *et al.*, 2014c), we used the MRC annotations as seed, and propagated the ratings to all vocabulary words using the word embeddings as features in an ℓ_2-regularized logistic regression classifier.
- Conventionalization features count the number of "conventional" words and phrases in a paragraph. Assuming that a Wikipedia title is a proxy to a conventionalized concept, we counted the number of existing titles (from a database of over 4.5 million titles) in the paragraph.
- Number of syllables scores are also extracted from the AoA database; out-of-database words were annotated as 5-syllable words.
- Relative frequency in a supersense was computed by normalizing the word frequencies in the training corpus by their coarse semantic categories defined in the WordNet.[3] There are 41 supersense types: 26 for nouns and 15 for verbs, e.g., NOUN.ANIMAL and VERB.MOTION. For example, in NOUN.ANIMAL the relative frequency of *human* is 0.06, of *dog* is 0.01, of *bird* is 0.01, of *cattle* is 0.009, and of *bumblebee* is 0.0002.
- Relative frequency in a synset was calculated similarly to the previous feature category, but word frequencies were normalized by their WordNet synsets (more fine-grained synonym sets). For example, in the synset {*vet, warhorse, veteran, oldtimer, seasoned stager*}, *veteran* is the most prototypical word, scoring 0.87.

4.3 Evaluation Benchmarks

We evaluate the utility of the pretrained word embeddings as features in downstream NLP tasks. We choose the following off-the-shelf models that utilize pretrained word embeddings as features:

Sentiment Analysis (Senti). Socher *et al.* (2013a) created a treebank of sentences annotated with fine-grained sentiment labels on phrases and sentences from movie review excerpts. The coarse-grained treebank of positive and negative classes has been split into training, development, and test datasets containing 6,920, 872, and 1,821 sentences, respectively. We use the average of the word vectors of a given sentence as a feature vector for classification (Faruqui *et al.*, 2015; Sedoc *et al.*, 2016). The ℓ_2-regularized logistic regression classifier is tuned on the development set and accuracy is reported on the test set.

[3] Supersenses are introduced in more detail in chapter 3.

Named Entity Recognition (NER). Named entity recognition is the task of identifying proper names in a sentence, such as names of persons, locations etc. We use the recently proposed LSTM-CRF NER model (Lample *et al.*, 2016) which trains a forward-backward LSTM on a given sequence of words (represented as word vectors), the hidden units of which are then used as (the only) features in a CRF model (Lafferty *et al.*, 2001) to predict the output label sequence. We use the CoNLL 2003 English NER dataset (Tjong Kim Sang and De Meulder, 2003) to train our models and present results on the test set.

Part of Speech Tagging (POS). For POS tagging, we again use the LSTM-CRF model (Lample *et al.*, 2016), but instead of predicting the named entity tag for every word in a sentence, we train the tagger to predict the POS tag of the word. The tagger is trained and evaluated with the standard Penn TreeBank (PTB) (Marcus *et al.*, 1993) training, development and test set splits as described by Collins (2002).

Dependency Parsing (Parse). Dependency parsing is the task of identifying syntactic relations between the words of a sentence. For dependency parsing, we train the stack-LSTM parser of Dyer *et al.* (2015) for English on the Universal Dependency v1.1 Treebank (Agić *et al.*, 2015) with the standard development and test splits, reporting unlabeled attachment scores (UAS) on the test data. We remove all part-of-speech and morphology features from the data, and prevent the model from optimizing the word embeddings used to represent each word in the corpus, thereby forcing the parser to rely completely on the pretrained embeddings.

4.4 Experiments

Data. All models were trained on Wikipedia articles, split to paragraph-per-line. Texts were cleaned, tokenized, numbers were normalized by replacing each digit with "DG", all types that occur less than 10 times were replaces by the "UNK" token, the data was not lowercased. We list data sizes in table 4.1.

# paragraphs	# tokens	# types
2,532,361	100,872,713	156,663

Table 4.1: Training data sizes.

Setup. 100-dimensional word embeddings were trained using the cbow model implemented in the word2vec toolkit (Mikolov *et al.*, 2013b).[4] All training data was used, either shuffled

[4] To evaluate the impact of curriculum learning, we enforced sequential processing of data organized in a pre-defined order of training examples. To control for sequential processing, word embedding were learned by running the cbow using a single thread for one iteration.

or ordered by a curriculum. As described in §4.3, we modified the extrinsic tasks to learn solely from word embeddings, without additional features. All models were learned under same conditions, across curricula: in Parse, NER, and POS we limited the number of training iterations to 3, 3, and 1, respectively. This setup allowed us to evaluate the effect of curriculum without additional interacting factors.

Experiments. In all the experiments we first train word embedding models, then the word embeddings are used as features in four extrinsic tasks (§4.3). We tune the tasks on development data, and report results on the test data. The only component that varies across the experiments is order of paragraphs in the training corpus—the curriculum. We compare the following experimental setups:

- **Shuffled** baselines: the curriculum is defined by random shuffling the training data. We shuffled the data 10 times, and trained 10 word embeddings models, each model was then evaluated on downstream tasks. Following Bengio *et al.* (2009), we report test results for the system that is closest to the median in dev scores. To evaluate variability and a range of scores that can be obtained from shuffling the data, we also report test results for systems that obtained the highest dev scores.
- **Sorted** baselines: the curriculum is defined by sorting the training data by sentence length in increasing/decreasing order, similarly to (Spitkovsky *et al.*, 2010).
- **Coherent** baselines: the curriculum is defined by just concatenating Wikipedia articles. The goal of this experiment is to evaluate the importance of semantic coherence in training data. Our intuition is that a coherent curriculum can improve models, since words with similar meanings and similar contexts are grouped when presented to the learner.
- **Optimized curriculum** models: the curriculum is optimized using the BayesOpt. We evaluate and compare models optimized using features from one of the three feature groups (§4.2.2). As in the shuffled baselines, we fix the number of trials (here, BayesOpt iterations) to 10, and we report test results of systems that obtained best dev scores.

Results. Experimental results are listed in table 4.2. Most systems trained with curriculum substantially outperform the strongest of all baselines. These results are encouraging, given that all word embedding models were trained on the same set of examples, only in different order, and display the indirect influence of the data curriculum on downstream tasks. These results support our assumption that curriculum matters. Albeit not as pronounced as with optimized curriculum, sorting paragraphs by length can also lead to substantial improvements over random baselines, but there is no clear recipe on whether the models prefer curricula sorted in an increasing or decreasing order. These results also support the advantage of a task-specific optimization framework over a general, intuition-guided recipe. An interesting result, also, that shuffling is not essential: systems trained on coherent data are on par (or better) than the shuffled systems.[5] In the next section, we analyze these results qualitatively.

		Senti	NER	POS	Parse
Shuffled	median	66.01	85.88	96.35	75.08
	best	66.61	85.50	96.38	76.40
Sorted	long→short	66.78	85.22	96.47	75.85
	short→long	66.12	85.49	96.20	75.31
Coherent	original order	66.23	85.99	96.47	76.08
Optimized curriculum	diversity	66.06	86.09	96.59	**76.63**
	prototypicality	**67.44**	85.96	96.53	75.81
	simplicity	67.11	**86.42**	**96.62**	76.54

Table 4.2: Evaluation of the impact of the curriculum of word embeddings on the downstream tasks.

4.5 Analysis

What are task-specific curriculum preferences? We manually inspect learned features and curriculum-sorted corpora, and find that best systems are obtained when their embeddings are learned from curricula appropriate to the downstream tasks. We discuss below several examples.

POS and Parse systems converge to the same set of weights, when trained on features that provide various measures of syntactic simplicity. The features with highest coefficients (and thus the most important features in sorting) are #*NPs*, Parse tree depth, #*VPs*, and #*PPs* (in this order). The sign in the #*NPs* feature weight, however, is the opposite from the other three feature weights (i.e., sorted in different order). #*NPs* is sorted in the increasing order of the number of noun phrases in a paragraph, and the other features are sorted in the decreasing order. Since Wikipedia corpus contains a lot of partial phrases (titles and headings), such curriculum promotes more complex, full sentences, and demotes partial sentences.

Best Senti system is sorted by prototypicality features. Most important features (with the highest coefficients) are Concreteness, Relative frequency in a supersense, and the Number of syllables. First two are sorted in decreasing order (i.e. paragraphs are sorted from more to less concrete, and from more to less prototypical words), and the Number of syllables is sorted in increasing order (this also promotes simpler, shorter words which are more prototypical). We

[5] Note that in the shuffled NER baselines, best dev results yield lower performance on the test data. This implies that in the standard development/test splits the development and test sets are not fully compatible or not large enough. We also observe this problem in the curriculum-optimized Parse-prototypicality and Senti-diversity systems. The dev scores for the Parse systems are 76.99, 76.47, 76.47 for diversity, prototypicality, and simplicity, respectively, but the prototypicality-sorted parser performs poorly on test data. Similarly in the sentiment analysis task, the dev scores are 69.15, 69.04, 69.49 for diversity, prototypicality, and simplicity feature groups. Senti-diversity scores, however, are lower on the test data, although the dev results are better than in Senti-simplicity. This limitation of the standard dev/test splits is beyond the scope of this work.

hypothesize that this soring reflects the type of data that Sentiment analysis task is trained on: it is trained on movie reviews, that are usually written in a simple, colloquial language.

Unlike POS, Parse, and Senti systems, all NER systems prefer curricula in which texts are sorted from short to long paragraphs. The most important features in the best (simplicity-sorted) system are *#PPs* and Verb-token ratio, both sorted from less to more occurrences of prepositional and verb phrases. Interestingly, most of the top lines in the NER system curricula contain named entities, although none of our features mark named entities explicitly. We show top lines in the simplicity-optimized system in figure 4.2.

Trimingham “ Golf ” ball .
Adélie penguin
“ **Atriplex** ” leaf UNK UNK
Hồng Lĩnh mountain
Anneli Jäätteenmäki UNK cabinet
Gävle goat
Early telescope observations .
Scioptric ball
Matryoshka doll
Luxembourgian passport
Australian Cashmere goat
Plumbeous water redstart
Dagebüll lighthouse
Vecom FollowUs . tv
Syracuse Junction railroad .
San Clemente Island goat
Tychonoff plank

Figure 4.2: Most of the top lines in best-scoring NER system contain named entities, although our features do not annotate named entities explicitly.

Finally, in all systems sorted by prototypicality, the last line is indeed not a prototypical word *Donaudampfschiffahrtselektrizitätenhauptbetriebswerkbauunterbeamtengesellschaft*, which is an actual word in German, frequently mentioned as an example of compounding in synthetic languages, but rarely used by German speakers.

Weighting examples according to curriculum. Another way to integrate curriculum in word embedding training is to weight training examples according to curriculum during word representation training. We modify the cbow objective $\sum_{t=1}^{T} \log p(w_t|w_{t-c}..w_{t+c})$ as follows:[6]

$$\sum_{t=1}^{T}(\frac{1}{1+e^{-weight(w_t)}}+\lambda)\log p(w_t|w_{t-c}..w_{t+c})$$

Here, $weight(w_t)$ denotes the score attributed to the token w_t, which is the z-normalized score of the paragraph; λ=0.5 is determined empirically. $\log p(w_t)|w_{t-c}..w_{t+c})$ computes the probability of predicting word w_t, using the context of c words to the left and right of w_t. Notice that this quantity is no longer a proper probability, as we are not normalizing over the weights $weight(w_t)$ over all tokens. However, the optimization in word2vec is performed using stochastic gradient descent, optimizing for a single token at each iteration. This yields a normalizer of 1 for each iteration, yielding the same gradient as the original cbow model.

We retrain our best curriculum-sorted systems with the modified objective, also controlling for curriculum. The results are shown in table 4.3. We find that the benefit of integrating curriculum in training objective of word representations is not evident across tasks: Senti and NER systems trained on vectors with the modified objective substantially outperform best results in table 4.2; POS and Parse perform better than the baselines but worse than the systems with the original objective.

	Senti	NER	POS	Parse
curriculum	67.44	86.42	**96.62**	**76.63**
cbow+curric	**68.26**	**86.49**	96.48	76.54

Table 4.3: Evaluation of the impact of curriculum integrated in the cbow objective.

Are we learning task-specific curricula? One way to assess whether we learn meaningful task-specific curriculum preferences is to compare curricula learned by one downstream task across different feature groups. If learned curricula are similar in, say, NER system, despite being optimized once using diversity features and once using prototypicality features—two disjoint feature sets—we can infer that the NER task prefers word embeddings learned from examples presented in a certain order, regardless of specific optimization features. For each downstream task, we thus measure Spearman's rank correlation between the curricula optimized using diversity (D), or prototypicality (P), or simplicity (S) feature sets. Prior to measuring correlations, we remove duplicate lines from the training corpora. Correlation results across tasks and across feature sets are shown in table 4.4.

The general pattern of results is that if two systems score higher than baselines, training sentences of their feature embeddings have similar curricula (i.e., the Spearman's ρ is positive), and if two systems disagree (one is above and one is below the baseline), then their curricula

[6] The modified word2vec tool is located at https://github.com/wlin12/wang2vec .

also disagree (i.e., the Spearman's ρ is negative or close to zero). NER systems all outperform the baselines and their curricula have high correlations. Moreover, NER sorted by diversity and simplicity have better scores than NER sorted by prototypicality, and in line with these results $\rho(\text{S,D})_{NER} > \rho(\text{P,S})_{NER}$ and $\rho(\text{S,D})_{NER} > \rho(\text{D,P})_{NER}$. Similar pattern of results is in POS correlations. In Parse systems, also, diversity and simplicity features yielded best parsing results, and $\rho(\text{S,D})_{Parse}$ has high positive correlation. The prototypicality-optimized parser performed poorly, and its correlations with better systems are negative. The best parser was trained using the diversity-optimized curriculum, and thus $\rho(\text{D,P})_{Parse}$ is the lowest. Senti results follow similar pattern of curricula correlations.

	Senti	NER	POS	Parse
ρ(D, P)	-0.68	0.76	0.66	-0.76
ρ(P, S)	0.33	0.75	0.75	-0.45
ρ(S, D)	-0.16	0.81	0.51	0.67

Table 4.4: Curricula correlations across feature groups.

Curriculum learning vs. data selection. We compare the task of curriculum learning to the task of data selection (reducing the set of training instances to more important or cleaner examples). We reduce the training data to the subset of 10% of tokens, and train downstream tasks on the reduced training sets. We compare system performance trained using the top 10% of tokens in the best curriculum-sorted systems (Senti-prototypicality, NER-implicity, POS-simplicity, Parse-diversity) to the systems trained using the top 10% of tokens in a corpus with randomly shuffled paragraphs.[7] The results are listed in table 4.5.

	Senti	NER	POS	Parse
random	63.97	**82.35**	96.22	69.11
curriculum	**64.47**	76.96	**96.55**	**72.93**

Table 4.5: Data selection results.

The curriculum-based systems are better in POS and in Parse systems, mainly because these tasks prefer vectors trained on curricula that promote well-formed sentences (as discussed above). Conversely, NER prefers vectors trained on corpora that begin with named entities, so most of the tokens in the reduced training data are constituents in short noun phrases. These results suggest that the tasks of data selection and curriculum learning are different. Curriculum is about strong initialization of the models and time-course learning, which is not necessarily sufficient for data reduction.

[7] Top $n\%$ tokens are used rather than top $n\%$ paragraphs because in all tasks except NER curriculum-sorted corpora begin with longer paragraphs. Thus, with top $n\%$ paragraphs our systems would have an advantage over random systems due to larger vocabulary sizes and not necessarily due to a better subset of data.

A risk of overfitting to development data with Bayesian optimization. The number of BayesOpt iterations as well as the number of shuffled baselines was set to 10 to keep the overall runtime of the experiments lower. Even with small number of BayesOpt iterations, our models show promising improvements. However, when we retrain four best curriculum-sorted systems—Senti-prototypicality, NER-implicity, POS-simplicity, Parse-diversity—with 100 iterations, as shown in table 4.6, performance on development data increases across all tasks but test results drop for three out of four systems. These results suggest that with 100 iterations the BayesOpt model is overfitting to development data either due to a limitation of the optimization algorithm or, as noted in footnote 5, due to limitations of standard dev/test splits. We leave to future work further investigation of this issue, discussing it in §7.2.3.

	Senti	NER	POS	Parse
dev-10	69.04	90.44	96.47	77.90
test-10	67.44	86.42	96.62	76.63
dev-100	70.18	90.58	96.54	78.33
test-100	66.23	86.10	96.62	75.54

Table 4.6: Overfitting to development sets with Bayesian optimization.

4.6 Related Work

Two prior studies on curriculum learning in NLP are discussed in the chapter (Bengio *et al.*, 2009; Spitkovsky *et al.*, 2010). Curriculum learning and related research on self-paced learning has been explored more deeply in computer vision (Bengio *et al.*, 2009; Kumar *et al.*, 2010; Lee and Grauman, 2011) and in multimedia analysis (Jiang *et al.*, 2015). Bayesian optimization has also received little attention in NLP. GPs were used in the task of machine translation quality estimation (Cohn and Specia, 2013) and in temporal analysis of social media texts (Preotiuc-Pietro and Cohn, 2013); TPEs were used by Yogatama *et al.* (2015) for optimizing choices of feature representations—*n*-gram size, regularization choice, etc.—in supervised classifiers.

4.7 Summary

We used Bayesian optimization to optimize curricula for training dense distributed word representations, which, in turn, were used as the sole features in NLP tasks. Our experiments confirmed that better curricula yield stronger models. We also conducted an extensive analysis, which sheds better light on understanding of text properties that are beneficial for model initialization. The proposed novel technique for finding an optimal curriculum is general, and can be used with other datasets and models.

Chapter 5

Linguistic Knowledge in Deep Learning Models

This chapter continues the line of work on integrating explicit linguistic knowledge in neural NLP with the goal to learn better generalizations that are guided by linguistic theory. The chapter tackles the problem of poor generalization of neural models when a model is trained on data from different languages (or in general, data domains), and proposes a solution relying on linguistic typology as a mediator between inputs in different languages. The method focuses on integrating knowledge directly into the neural architecture at training time. The novel "polyglot" model—a neural architecture integrated with hand-engineered linguistic attributes, trained on and applied to any number of languages—obtains better performance than a corresponding standard architecture trained monolingually, and is useful in downstream applications. Work described in this chapter is the *NAACL 2016* publication (Tsvetkov *et al.*, 2016c), conducted in collaboration with Chris Dyer, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, and Lori Levin.

5.1 Background

Statistical language models (LMs) are a core component of machine translation, speech recognition, information retrieval, and other language processing tasks, that estimates semantic and morphosyntactic fluency of a sequence in a language. Traditional LMs are computed as n-gram counts with smoothing (Kneser and Ney, 1995; Chen and Goodman, 1996). More recently, neural probabilistic language models have been shown to outperform the count-based models, due to their ability to incorporate longer contexts. Most prominent neural LM architectures include feed-forward networks (Bengio *et al.*, 2003; Schwenk, 2007), log-bilinear models (Mnih and Hinton, 2007), and recurrent neural networks (Mikolov *et al.*, 2010, 2011; Sundermeyer *et al.*, 2012).

Nearly all existing LM architectures are designed to model one language at a time. This is unsurprising considering the historical importance of count-based models in which every surface form of a word is a separately modeled entity (English *cat* and Spanish *gato* would not likely benefit from sharing counts). However, recent models that use distributed representations—in particular models that share representations across languages (Hermann and Blunsom, 2014; Faruqui and Dyer, 2014b; Huang *et al.*, 2015; Lu *et al.*, 2015; Ammar *et al.*, 2016b, *inter alia*)—suggest universal models applicable to multiple languages are a possibility. This chapter takes a step in this direction.

We introduce **polyglot language models**: neural network language models that are trained on and applied to any number of languages. Our goals with these models are the following. First, to facilitate data and parameter sharing, providing more training resources to languages, which is especially valuable in low-resource settings. Second, models trained on diverse languages with diverse linguistic properties will better be able to learn naturalistic representations that are less likely to "overfit" to a single linguistic outlier. Finally, polyglot models offer convenience in a multilingual world: a single model replaces dozens of different models.

Exploration of polyglot language models at the sentence level—the traditional domain of language modeling—requires dealing with a massive event space (i.e., the union of words across many languages). To work in a more tractable domain, we evaluate our model on phone-based language modeling, the modeling sequences of *sounds*, rather than words. We choose this domain since a common assumption of many theories of phonology is that all spoken languages construct words from a finite inventory of phonetic symbols (represented conveniently as the elements of the the International Phonetic Alphabet; IPA) which are distinguished by language-universal features (e.g., place and manner of articulation, voicing status, etc.). Although our focus is on sound sequences, our solution can be ported to the semantic/syntactic problem as resulting from adaptation to constraints on semantic/syntactic structure.

This chapter makes two primary contributions: in modeling and in applications. In §5.2, we introduce a novel polyglot neural language model (NLM) architecture. Despite being trained on multiple languages, the multilingual model is more effective (9.5% lower perplexity) than individual models, and substantially more effective than naive baselines (over 25% lower perplexity). Our most effective polyglot architecture conditions not only on the identity of the language being predicted in each sequence, but also on a vector representation of its phono-typological properties. In addition to learning representations of phones as part of the polyglot language modeling objective, the model incorporates features about linguistic typology to improve generalization performance (§5.3). Our second primary contribution is to show that downstream applications are improved by using polyglot-learned phone representations. We focus on two tasks: predicting adapted word forms in models of cross-lingual lexical borrowing and speech synthesis (§5.4). Our experimental results (§5.5) show that in borrowing, we improve over the current state-of-the-art, and in speech synthesis, our features are more effective than manually-designed phonetic features. Finally, we analyze the phonological content of learned representations, finding that our polyglot models discover standard phonological categories such as length and nasalization, and that these are grouped correctly across languages with different phonetic inventories and contrastive features.

5.2 Model

In this section, we first describe in §5.2.1 the underlying framework of our model—RNNLM—a standard recurrent neural network based language model (Mikolov *et al.*, 2010; Sundermeyer *et al.*, 2012). Then, in §5.2.2, we define a Polyglot LM—a modification of RNNLM to incorporate language information, both learned and hand-crafted.

Problem definition. In the phonological LM, *phones* (sounds) are the basic units. Mapping from words to phones is defined in pronunciation dictionaries. For example, "cats" [kæts] is a sequence of four phones. Given a prefix of phones $\phi_1, \phi_2, \ldots, \phi_{t-1}$, the task of the LM is to estimate the conditional probability of the next phone $p(\phi_t \mid \phi_1, \phi_2, \ldots, \phi_{t-1})$.

5.2.1 RNNLM

In NLMs, a vocabulary V (here, a set of phones composing all word types in the language) is represented as a matrix of parameters $\mathbf{X} \in \mathbb{R}^{d \times |V|}$, with $|V|$ phone types represented as d-dimensional vectors. $\mathbf{X}$ is often denoted as lookup table. Phones in the input sequence are first converted to phone vectors, where ϕ_i is represented by $\mathbf{x}_i$ by multiplying the phone indicator (one-hot vector of length $|V|$) and the lookup table.

At each time step t, most recent phone prefix vector[1] $\mathbf{x}_t$ and hidden state $\mathbf{h}_{t-1}$ are transformed to compute a new hidden representation:

$$\mathbf{h}_t = f(\mathbf{x}_t, \mathbf{h}_{t-1}),$$

where f is a non-linear transformation. In the original RNNLMs (Mikolov *et al.*, 2010), the transformation is such that:

$$\mathbf{h}_t = \tanh(\mathbf{W}_{h_x}\mathbf{x}_t + \mathbf{W}_{h_h}\mathbf{h}_{t-1} + \mathbf{b}_h).$$

[1] We are reading at each time step the most recent n-gram context rather than—as is more common in RNNLMs—a single phone context. Empirically, this works better for phone sequences, and we hypothesize that this lets the learner rely on direct connections for local phenomena (which are abundant in phonology) and minimally use the recurrent state to model longer-range effects.

To overcome the notorious problem in recurrent neural networks of vanishing gradients (Bengio *et al.*, 1994), following Sundermeyer *et al.* (2012), in recurrent layer we use long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997):[2]

$$\mathbf{h}_t = \text{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}).$$

Given the hidden sequence $\mathbf{h}_t$, the output sequence is then computed as follows:

$$p(\phi_t = i \mid \phi_1, \ldots, \phi_{t-1}) = \text{softmax}(\mathbf{W}_{out}\mathbf{h}_t + \mathbf{b}_{out})_i,$$

where $\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$ ensures a valid probability distribution over output phones.

5.2.2 Polyglot LM

We now describe our modifications to RNNLM to account for multilinguality. The architecture is depicted in figure 5.1. Our task is to estimate the conditional probability of the next phone given the preceding phones and the language (ℓ): $p(\phi_t \mid \phi_1, \ldots, \phi_{t-1}, \ell)$.

In a multilingual NLM, we define a vocabulary V^* to be the union of vocabularies of all training languages, assuming that all language vocabularies are mapped to a shared representation (here, IPA). In addition, we maintain V_ℓ with a special symbol for each language (e.g., $\phi_{english}$, ϕ_{arabic}). Language symbol vectors are parameters in the new lookup table $\mathbf{X}_\ell \in \mathbb{R}^{d \times |\#langs|}$ (e.g., $\mathbf{x}_{english}$, $\mathbf{x}_{arabic}$). The inputs to the Polyglot LM are the phone vectors $\mathbf{x}_t$, the language character vector $\mathbf{x}_\ell$, and the typological feature vector constructed externally $\mathbf{t}_\ell$. The typological feature vector will be discussed in the following section.

The input layer is passed to the hidden local-context layer:

$$\mathbf{c}_t = \mathbf{W}_{c_x}\mathbf{x}_t + \mathbf{W}_{c_{lang}}\mathbf{x}_{lang} + \mathbf{b}_c.$$

The local-context vector is then passed to the hidden LSTM global-context layer, similarly to the previously described RNNLM:

$$\mathbf{g}_t = \text{LSTM}(\mathbf{c}_t, \mathbf{g}_{t-1}).$$

In the next step, the global-context vector $\mathbf{g}_t$ is "factored" by the typology of the training language, to integrate manually-defined language features. To obtain this, we first project the (potentially high-dimensional) $\mathbf{t}_\ell$ into a low-dimensional vector, and apply non-linearity. Then, we multiply the $\mathbf{g}_t$ and the projected language layer, to obtain a global-context-language matrix:

[2] For brevity, we omit the equations describing the LSTM cells; they can be found in (Graves, 2013, eq. 7–11).

$$\mathbf{f}_\ell = \tanh(\mathbf{W}_\ell \mathbf{t}_\ell + \mathbf{b}_\ell),$$
$$\mathbf{G}_t^\ell = \mathbf{g}_t \otimes \mathbf{f}_\ell^\top.$$

Finally, we vectorize the resulting matrix into a column vector and compute the output sequence as follows:

$$p(\phi_t = i \mid \phi_1, \ldots, \phi_{t-1}, \ell) = \text{softmax}(\mathbf{W}_{out}\text{vec}(\mathbf{G}_t^\ell) + \mathbf{b}_{out})_i.$$

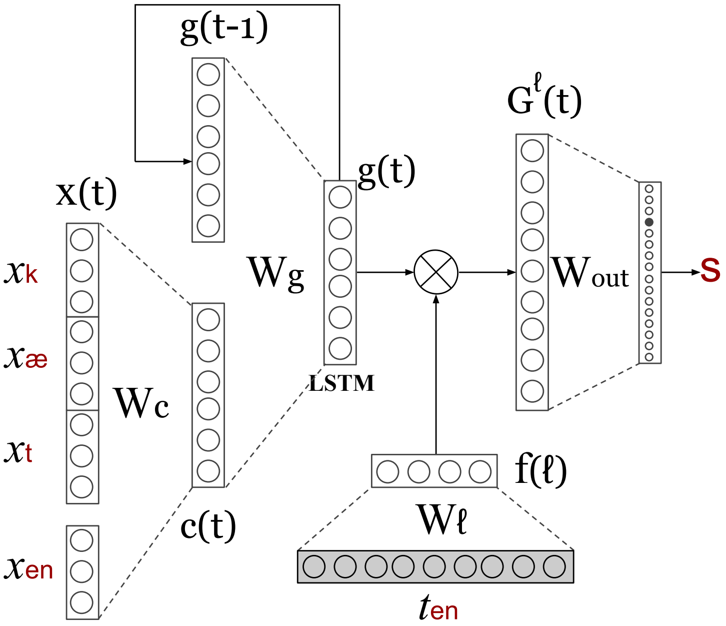


Figure 5.1: Architecture of the Polyglot LM.

Model training. Parameters of the models are the lookup tables $\mathbf{X}$ and $\mathbf{X}_\ell$, weight matrices $\mathbf{W}_i$, and bias vectors $\mathbf{b}_i$. Parameter optimization is performed using stochastic updates to minimize the categorical cross-entropy loss (which is equivalent to minimizing perplexity and maximizing likelihood): $H(\phi, \hat{\phi}) = -\Sigma_i \hat{\phi}_i \log \phi_i$, where ϕ is predicted and $\hat{\phi}$ is the gold label.

5.3 Typological Features

Typological information is fed to the model via vectors of 190 binary typological features, all of which are phonological (related to sound structure) in their nature. These feature vectors are derived from data from the WALS (Dryer and Haspelmath, 2013a), PHOIBLE (Moran *et al.*, 2014), and Ethnologue (Lewis *et al.*, 2015) typological databases via extensive post-processing and analysis.[3] The features primarily concern properties of sound inventories (i.e., the set of phones or phonemes occurring in a language) and are mostly of one of four types:

1. **Single segment represented in an inventory**; e.g., does language ℓ's sound inventory include /g/, a voiced velar stop?
2. **Class of segments represented in an inventory**; e.g., does language ℓ's sound inventory include voiced fricatives like /z/ and /v/?
3. **Minimal contrast represented in an inventory**; e.g., does language ℓ's sound inventory include two sounds that differ only in voicing, such as /t/ and /d/?
4. **Number of sounds representative of a class that are present in an inventory**; e.g., does language ℓ's sound inventory include exactly five vowels?

The motivation and criteria for coding each individual feature required extensive linguistic knowledge and analysis. Consider the case of tense vowels like /i/ and /u/ in "beet" and "boot" in contrast with lax vowels like /ɪ/ and /ʊ/ in "bit" and "book." Only through linguistic analysis does it become evident that (1) all languages have tense vowels—a feature based on the presence of tense vowels is uninformative and that (2) a significant minority of languages make a distinction between tense and lax vowels—a feature based on whether languages display a minimal difference of this kind would be more useful.

5.4 Applications of Phonetic Vectors

Learned continuous word representations—word vectors—are an important by-product of neural LMs, and these are used as features in numerous NLP applications, including chunking (Turian *et al.*, 2010b), part-of-speech tagging (Ling *et al.*, 2015a), dependency parsing (Lazaridou *et al.*, 2013; Bansal *et al.*, 2014; Dyer *et al.*, 2015; Watanabe and Sumita, 2015), named entity recognition (Guo *et al.*, 2014), and sentiment analysis (Socher *et al.*, 2013b; Wang *et al.*, 2015). We evaluate phone vectors learned by Polyglot LMs in two downstream applications that rely on phonology: modeling lexical borrowing (§5.4.1) and speech synthesis (§5.4.2).

5.4.1 Lexical borrowing

In the process of their nativization in a foreign language, loanwords undergo primarily **phonological adaptation**, namely insertion/deletion/substitution of phones to adapt to the phonotactic constraints of the recipient language. If a foreign phone is not present in the recipient

[3] This data resource, which provides standardized phono-typological information for 2,273 languages, is available at `https://github.com/dmort27/uriel-phonology/tarball/0.1`. It is a subset of the URIEL database, a comprehensive database of typological features encoding syntactic and morphological (as well as phonological) properties of languages. It is available at `http://cs.cmu.edu/~dmortens/uriel.html`.

language, it is usually replaced with its closest native equivalent—we thus hypothesize that cross-lingual phonological features learned by the Polyglot LM can be useful in models of borrowing to quantify cross-lingual similarities of sounds.

To test this hypothesis, we augment the hand-engineered models proposed in chapter 2 with features from phone vectors learned by our model. Inputs to the borrowing framework are loanwords (in Swahili, Romanian, Maltese), and outputs are their corresponding "donor" words in the donor language (Arabic, French, Italian, resp.). The framework is implemented as a cascade of finite-state transducers with insertion/deletion/substitution operations on sounds, weighted by high-level conceptual linguistic constraints that are learned in a supervised manner. Given a loanword, the system produces a candidate donor word with lower ranked violations than other candidates, using the shortest path algorithm. In the original borrowing model, insertion/deletion/substitution operations are unweighted. In this work, we integrate transition weights in the phone substitution transducers, which are cosine distances between phone vectors learned by our model. Our intuition is that similar sounds appear in similar contexts, even if they are not present in the same language (e.g., /sˤ/ in Arabic is adapted to /s/ in Swahili). Thus, if our model effectively captures cross-lingual signals, similar sounds should have smaller distances in the vector space, which can improve the shortest path results. Figure 5.2 illustrates our modifications to the original framework.

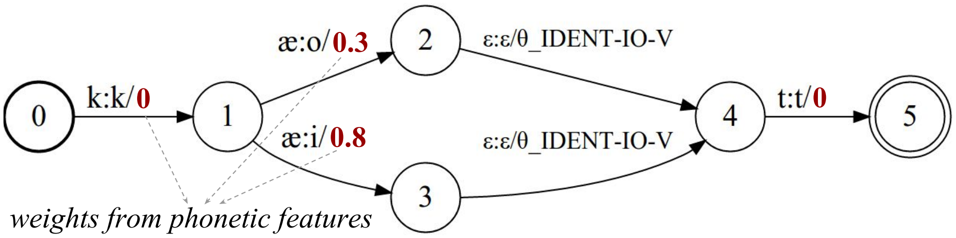


Figure 5.2: Distances between phone vectors learned by the Polyglot LM are integrated as substitution weights in the lexical borrowing transducers. An English word *cat* [kæt] is adapted to its Russian counterpart кот [kot]. The transducer has also an erroneous path to кит [kit] 'whale'. In the original system, both paths are weighted with the same feature IDENT-IO-V, firing on vowel substitution. Our modification allows the borrowing model to identify more plausible paths by weighting substitution operations.

5.4.2 Speech synthesis

Speech synthesis is the process of converting text into speech. It has various applications, such as screen readers for the visually impaired and hands-free voice based systems. Text-to-speech (TTS) systems are also used as part of speech-to-speech translation systems and spoken dialog systems, such as personal digital assistants. Natural and intelligible TTS systems exist for a number of languages in the world today. However, building TTS systems remains prohibitive for many languages due to the lack of linguistic resources and data. Many of these low resource languages have a high user population, or a large number of non-literate speakers for which developing speech-based systems may be beneficial.

The language-specific resources that are traditionally used for building TTS systems in a new language are: (1) audio recordings with transcripts; (2) pronunciation lexicon or letter to sound rules; and (3) a phone set definition. Standard TTS systems today use phone sets designed by experts. Typically, these phone sets also contain phonetic features for each phoneme, which are used as features in models of the spectrum and prosody. The phonetic features available in standard TTS systems are multidimensional vectors indicating various properties of each phoneme, such as whether it is a vowel or consonant, vowel length and height, place of articulation of a consonant, etc. Constructing these features by hand can be labor intensive, and coming up with such features automatically may be useful in low-resource scenarios.

In this work, we replace manually engineered phonetic features with phone vectors, which are then used by classification and regression trees for modeling the spectrum. Each phoneme in our phone set is assigned an automatically constructed phone vector, and each member of the phone vector is treated as a phoneme-level feature which is used in place of the manually engineered phonetic features. While prior work has explored TTS augmented with acoustic features (Watts *et al.*, 2015), to the best of our knowledge, we are the first to replace manually engineered phonetic features in TTS systems with automatically constructed phone vectors.

5.5 Experiments

Our experimental evaluation of our proposed polyglot models consists of two parts: (i) an intrinsic evaluation where phone sequences are modeled with independent models and (ii) an extrinsic evaluation of the learned phonetic representations. Before discussing these results, we provide details of the data resources we used.

5.5.1 Resources and experimental setup

Resources. We experiment with the following languages: Arabic (AR), French (FR), Hindi (HI), Italian (IT), Maltese (MT), Romanian (RO), Swahili (SW), Tamil (TA), and Telugu (TE). In our language modeling experiments, two main sources of data are pronunciation dictionaries and typological features described in §5.3. The dictionaries for AR, FR, HI, TA, and TE are taken from in-house speech recognition/synthesis systems. For remaining languages, the dictionaries are automatically constructed using the Omniglot grapheme-to-IPA conversion rules.[4]

[4] http://omniglot.com/writing/

We use two types of pronunciation dictionaries: (1) AR, FR, HI, IT, MT, RO, and SW dictionaries used in experiments with lexical borrowing; and (2) EN, HI, TA, and TE dictionaries used in experiments with speech synthesis. The former are mapped to IPA, with the resulting phone vocabulary size—the number of distinct phones across IPA dictionaries—of 127 phones. The latter are encoded using the UniTran universal transliteration resource (Qian *et al.*, 2010), with a vocabulary of 79 phone types.

	AR	FR	HI	IT	MT	RO	SW
train	1,868/18,485	238/1,851	193/1,536	988/901	114/1,152	387/4,661	659/7,239
dev	366/3,627	47/363	38/302	19/176	22/226	76/916	130/1,422
test	208/2,057	27/207	22/173	11/100	13/128	43/524	73/806

Table 5.1: Train/dev/test counts for IPA pronunciation dictionaries for words (phone sequences) and phone tokens, in thousands: #thousands of sequences/# thousands of tokens.

From the (word-type) pronunciation dictionaries, we remove 15% of the words for development, and a further 10% for testing; the rest of the data is used to train the models. In tables 5.1 and 5.2 we list—for both types of pronunciation dictionaries—train/dev/test data statistics for words (phone sequences) and phone tokens. We concatenate each phone sequence with beginning and end symbols.

Hyperparameters. We used the following network architecture: 100-dimensional phone vectors, with hidden local-context and LSTM layers of size 100, and hidden language layer of size 20. All language models were trained using the left context of 3 phones (4-gram LMs). Across all language modeling experiments, parameter optimization was performed on the dev set using the Adam algorithm (Kingma and Ba, 2014) with mini-batches of size 100 to train the models for 5 epochs.

5.5.2 Intrinsic perplexity evaluation

Perplexity is the standard evaluation measure for language models, which has been shown to correlate strongly with error rates in downstream applications (Klakow and Peters, 2002). We evaluated perplexities across several architectures, and several monolingual and multilingual setups. We kept the same hyper-parameters across all setups, as detailed in §5.5. Perplexities

	EN	HI	TA	TE
train	101/867	191/1,523	74/780	71/690
dev	20/169	37/300	14/152	14/135
test	11/97	21/171	8/87	8/77

Table 5.2: Train/dev/test statistics for UniTran pronunciation dictionaries for words (phone sequences) and phone tokens, in thousands: #thousands of sequences/# thousands of tokens.

of LMs trained on the two types of pronunciation dictionaries were evaluated separately; table 5.3 summarizes perplexities of the models trained on IPA dictionaries, and table 5.4 summarizes perplexities of the UniTran LMs.

In columns, we compare three model architectures: *baseline* denotes the standard RNNLM architecture described in §5.2.1; *+lang* denotes the Polyglot LM architecture described in §5.2.2 with input language vector but without typological features and language layer; finally, *+typology* denotes the full Polyglot LM architecture. This setup lets us separately evaluate the contribution of modified architecture and the contribution of auxiliary set of features introduced via the language layer.

Test languages are IT in table 5.3, and HI in table 5.4. The rows correspond to different sets of training languages for the models: *monolingual* is for training and testing on the same language; *+similar* denotes training on three typologically similar languages: IT, FR, RO in table 5.3, and HI, TA, TE in table 5.4; *+dissimilar* denotes training on four languages, three similar and one typologically dissimilar language, to evaluate robustness of multilingual systems to diverse types of data. The final sets of training languages are IT, FR, RO, HI in table 5.3, and HI, TA, TE, EN in table 5.4.

	Perplexity (↓)			
training set	baseline	+lang	+typology	
monolingual	4.36	–	–	
+similar	5.73	4.93	**4.24**	(↓ 26.0%)
+dissimilar	5.88	4.98	**4.41**	(↓ 25.0%)

Table 5.3: Perplexity experiments with IT as test language. Training languages: monolingual: IT; +similar: IT, FR, RO; +dissimilar: IT, FR, RO, HI.

	Perplexity (↓)			
training set	baseline	+lang	+typology	
monolingual	3.70	–	–	
+similar	4.14	3.78	**3.35**	(↓ 19.1%)
+dissimilar	4.29	3.82	**3.42**	(↓ 20.3%)

Table 5.4: Perplexity experiments with HI as test language. Training languages: monolingual: HI; +similar: HI, TA, TE; +dissimilar: HI, TA, TE, EN.

We see several patterns of results. First, polyglot models require, unsurprisingly, information about what language they are predicting to obtain good modeling performance. Second, typological information is more valuable than letting the model learn representations of the language along with the characters. Finally, typology-augmented polyglot models outperform their monolingual baseline, providing evidence in support of the hypothesis that cross-lingual evidence is useful not only for learning cross-lingual representations and models, but monolingual ones as well.

5.5.3 Lexical borrowing experiments

We used lexical borrowing models described in chapter 2 for three language pairs: AR–SW, FR–RO, and IT–MT. Train and test corpora are donor–loanword pairs in the language pairs. We use the original systems as the baselines, and compare these to the corresponding systems augmented with phone vectors, as described in §5.4.1.

Integrated vectors were obtained from a single polyglot model with typology, trained on all languages with IPA dictionaries; only vectors trained by the model were used and not the full polyglot language model. For comparison with the results in table 5.3, perplexity of the model on the IT dataset (used for evaluation is §5.5.2) is 4.16, even lower than in the model trained on four languages. To retrain the high-level conceptual linguistic features learned by the borrowing models, we initialized the augmented systems with feature weights learned by the baselines, and retrained. Final weights were established using cross-validation. Then, we evaluated the accuracy of the augmented borrowing systems on the held-out test data.

Accuracies are shown in table 5.5. We observe improvements of up to 5% in accuracies of FR–RO and IT–MT pairs. Effectiveness of the same polyglot model trained on multiple languages and integrated in different downstream systems supports our assumption that the model remains stable and effective with addition of languages. Our model is less effective for the AR–SW language pair. We speculate that the results are worse, because this is a pair of (typologically) more distant languages; consequently, the phonological adaptation processes that happen in loanword assimilation are more complex than mere substitutions of similar phones that we are targeting via the integration of phone vectors.

	Accuracy (↑)		
	AR–SW	FR–RO	IT–MT
baseline	**48.4**	75.6	83.3
+multilingual	46.9	**80.6**	**87.1**

Table 5.5: Accuracies of the baseline models of lexical borrowing and the models augmented with phone vectors. In all the experiments, we use vectors from a single Polyglot LM model trained on AR, SW, FR, RO, IT, MT.

5.5.4 Speech synthesis experiments

TTS systems are evaluated using a variety of objective and subjective metrics. Subjective metrics, which require humans to rate or compare systems by listening to them can be expensive and time consuming. A popular objective metric for measuring the quality of synthetic speech is the Mel Cepstral Distortion (MCD) (Hu and Loizou, 2008). The MCD metric calculates an L2 norm of the Mel Frequency Cepstral Coefficients (MFCCs) of natural speech from a held out test set, and synthetic speech generated from the same test set. Since this is a distance metric, a lower value of MCD suggests better synthesis. The MCD is a database-specific metric, but experiments by Kominek et al. (Kominek *et al.*, 2008) have shown that a decrease in MCD of 0.08 is perceptually significant, and a decrease of 0.12 is equivalent to doubling the size of the TTS database. In our experiments, we use MCD to measure the relative improvement obtained by our techniques.

We conducted experiments on the IIIT-H Hindi voice database (Prahallad *et al.*, 2012), a 2 hour single speaker database recorded by a professional male speaker. We used the same front end (UniTran) to build all the Hindi TTS systems, with the only difference between the systems being the presence or absence of phonetic features and our vectors. For all our voice-based experiments, we built CLUSTERGEN Statistical Parametric Synthesis voices (Black, 2006) using the Festvox voice building tools (Black and Lenzo, 2003) and the Festival speech synthesis engine (Black and Taylor, 1997).

The baseline TTS system was built using no phonetic features. We also built a TTS system with standard hand-crafted phonetic features. Table 5.6 shows the MCD for the HI baseline, the standard TTS with hand-crafted features, and augmented TTS systems built using monolingual and multilingual phone vectors constructed with Polyglot LMs.

	MCD (↓)
baseline	4.58
+hand-crafted	4.41
+monolingual	4.40
+multilingual	**4.39**

Table 5.6: MCD for the HI TTS systems. Polyglot LM training languages: monolingual: HI; +multilingual: HI, TA, TE, EN.

Our multilingual vectors outperform the baseline, with a significant decrease of 0.19 in MCD. Crucially, TTS systems augmented with the Polyglot LM phone vectors outperform also the standard TTS with hand-crafted features. We found that using both feature sets added no value, suggesting that learned phone vectors are capturing information that is equivalent to the hand-engineered vectors.

5.5.5 Qualitative analysis of vectors

Phone vectors learned by Polyglot LMs are mere sequences of real numbers. An interesting question is whether these vectors capture linguistic (phonological) qualities of phones they are encoding. To analyze to what extent our vectors capture linguistic properties of phones, we use the QVEC—a method and a tool to quantify and interpret linguistic content of vector space models that we introduce in chapter 6 (Tsvetkov *et al.*, 2015b). The tool aligns dimensions in a matrix of learned distributed representations with dimensions of a hand-crafted linguistic matrix. Alignments are induced via correlating columns in the distributed and the linguistic matrices. To analyze the content of the distributed matrix, annotations from the linguistic matrix are projected via the maximally-correlated alignments.

We constructed a phonological matrix in which 5,059 rows are IPA phones and 21 columns are boolean indicators of universal phonological properties, e.g. *consonant*, *voiced*, *labial*.[5] We the projected annotations from the linguistic matrix and manually examined aligned dimensions in the phone vectors from §5.5.3 (trained on six languages). In the maximally-correlated columns—corresponding to linguistic features *long*, *consonant*, *nasalized*—we examined phones with highest coefficients. These were: [ɐː, ʊː, iː, ɔː, ɛː] for *long*; [v, ɲ, d͡ʒ, d, f, j, t͡s, ŋ] for *consonant*; and [ɔ̃, ɛ̃, ɑ̃, œ̃] for *nasalized*. Clearly, the learned representation discover standard phonological features. Moreover, these top-ranked sounds are not grouped by a single language, e.g., /d͡ʒ/ is present in Arabic but not in French, and /ɲ, ŋ/ are present in French but not in Arabic. From this analysis, we conclude that (1) the model discovers linguistically meaningful phonetic features; (2) the model induces meaningful related groupings across languages.

5.6 Related Work

Multilingual language models. Interpolation of monolingual LMs is an alternative to obtain a multilingual model (Harbeck *et al.*, 1997; Weng *et al.*, 1997). However, interpolated models still require a trained model per language, and do not allow parameter sharing at training time. Bilingual language models trained on concatenated corpora were explored mainly in speech recognition (Ward *et al.*, 1998; Wang *et al.*, 2002; Fügen *et al.*, 2003). Adaptations have been proposed to apply language models in bilingual settings in machine translation (Niehues *et al.*, 2011) and code switching (Adel *et al.*, 2013). These approaches, however, require adaptation to every pair of languages, and an adapted model cannot be applied to more than two languages.

Multimodal neural language models. Multimodal language modeling is integrating image/video modalities in text LMs. Our work is inspired by the neural multimodal LMs (Kiros and Salakhutdinov, 2013; Kiros *et al.*, 2015), which defined language models conditional on visual contexts, although we use a different language model architecture (recurrent vs. log-bilinear) and a different approach to gating modality.

Multilingual neural architectures. Concurrently with our work, Ammar *et al.* (2016a) used a different polyglot architecture for multilingual dependency parsing. This work has also confirmed the utility of polyglot architectures in leveraging multilinguality, but used a different mechanism for integrating linguistic features that did not yield gains in performance. Firat *et al.* (2016) introduced a multilingual architecture for machine translation, in which to

[5] This matrix is described in Littell *et al.* (2016) and is available at `https://github.com/dmort27/panphon/`.

mediate between languages they used an attention mechanism, shared across all language pairs.

5.7 Summary

We presented a novel *multilingual* language model architecture that uses an array of typological features to mediate between input languages. The model obtains substantial gains in perplexity, and improves downstream text and speech applications. Although we focus on phonology, our approach is general, and can be applied in problems that integrate divergent modalities, e.g., topic modeling, and multilingual tagging and parsing.

Chapter 6

Linguistic Knowledge in Evaluation of Distributed Representations

This chapter concludes the sequence of chapters on the synergy of linguistics and neural NLP. It focuses on the evaluation of distributed representations of words. The chapter introduces a general strategy to evaluate word vectors, that is based on correlating the model with rich linguistic resources. Work described in this chapter is based on the *EMNLP 2015* publication (Tsvetkov *et al.*, 2015b), and its extension in the *RepEval* workshop (Tsvetkov *et al.*, 2016a). Research was conducted in collaboration with Chris Dyer, Manaal Faruqui, Wang Ling, and Guillaume Lample.

6.1 Background

A major attraction of vector space word representations is that they can be derived from large unannotated corpora, and they are useful as a source of features for downstream NLP tasks that are learned from small amounts of supervision. Despite their ubiquity, there is no standard scheme for intrinsically evaluating the quality of word vectors. This lack of standardized evaluation is due, in part, to word vectors' major criticism: word vectors are linguistically opaque in a sense that it is still not clear how to interpret individual vector dimensions, and, consequently, it is not clear how to score a non-interpretable representation. Nevertheless, to facilitate development of better word vector models and for better error analysis of word vectors, it is desirable to compare word vector models easily, without recourse to multiple extrinsic applications whose implementation and runtime can be costly, and to understand how features in word vectors contribute to downstream tasks. Yet, unless it is coupled with an extrinsic task, intrinsic evaluation of word vectors has little value in itself. The main purpose of an intrinsic evaluation is to serve as a *proxy* for the downstream task the embeddings are tailored for. This chapter advocates a novel approach to constructing such a proxy.

What are the desired properties of an intrinsic evaluation measure of word embeddings? First, retraining models that use word embeddings as features is often expensive. A *computationally efficient* intrinsic evaluation that *correlates with extrinsic scores* is useful for faster prototyping. Second, an intrinsic evaluation that enables *interpretation* and analysis of properties encoded by vector dimensions is an auxiliary mechanism for analyzing how these properties affect the target downstream task. It thus facilitates refinement of word vector models and, consequently, improvement of the target task. Finally, an intrinsic evaluation that approximates a range of related downstream tasks (e.g., semantic text-classification tasks) allows to assess *generality* (or specificity) of a word vector model, without actually implementing all the tasks.

We propose QVEC and QVEC-CCA—two variants of a simple intrinsic evaluation measure for word vectors (§6.3). Our measures are based on component-wise correlations with annotated by domain experts "linguistic" word vectors whose components have well-defined linguistic properties (§6.2). QVEC helps shed new light on how vector spaces encode meaning, thus facilitating the interpretation and a qualitative evaluation of word vectors; QVEC-CCA is better suited for quantitative evaluation: it is faster than QVEC and obtains better correlations with downstream tasks. Since vectors are typically used to provide features to downstream learning problems, our measures favor *recall* (rather than precision), which captures our intuition that meaningless dimensions in induced vector representations are less harmful than important dimensions that are missing. To show that our proposed score is meaningful, we compare our intrinsic evaluation model to the standard extrinsic evaluation benchmarks (§6.4). For nine off-the-shelf word vector representation models, both measures are shown to correlate well with downstream semantic and syntactic tasks, and can easily be adjusted to a specific task (e.g., part-of-speech tagging) by selecting task-specific linguistic resources (e.g., part-of-speech annotations) (§6.5).

Both QVEC and QVEC-CCA, as well as their further extensions to multilingual evaluation, and to many-to-many alignment of distributional and linguistic matrices using Integer Linear Programming, are publicly released at `https://github.com/ytsvetko/qvec`, and are used by other researchers.

6.2 Linguistic Dimension Word Vectors

Both QVEC and QVEC-CCA rely on a matrix of linguistic properties constructed from a manually crafted linguistic resource. Linguistic resources are invaluable as they capture generalizations made by domain experts. However, resource construction is expensive, therefore it is not always possible to find an existing resource that captures exactly the set of optimal lexical properties for a downstream task. Resources that capture more coarse-grained, general properties can be used instead, for example, WordNet for semantic evaluation, or Penn Treebank

(Marcus *et al.*, 1993, PTB) for syntactic evaluation. Since these properties are not an exact match to the task, the intrinsic evaluation tests for a necessary (but possibly not sufficient) set of generalizations.

Semantic vectors. To evaluate the semantic content of word vectors, we exploit supersense annotations in a WordNet-annotated corpus—SemCor (Miller *et al.*, 1993). SemCor is a WordNet-annotated corpus that captures, among others, supersense annotations of WordNet's 13,174 noun lemmas and 5,686 verb lemmas at least once. We construct term frequency vectors normalized to probabilities for all nouns and verbs that occur in SemCor at least 5 times. The resulting supersense-dimension matrix has 4,199 rows (supersense-annotated nouns and verbs that occur in SemCor at least 5 times[1]), and 41 columns: 26 for nouns and 15 for verbs.

Syntactic vectors. Similar to semantic vectors, we construct syntactic vectors for all words with 5 or more occurrences in the training part of the PTB. Vector dimensions are probabilities of the part-of-speech (POS) annotations in the corpus. This results in 10,865 word vectors with 45 interpretable columns, each column corresponds to a POS tag from the PTB. Example vectors are shown in table 6.1.

WORD	NN.ANIMAL	NN.FOOD	···	VB.MOTION
fish	0.68	0.16	···	0.00
duck	0.31	0.00	···	0.69
chicken	0.33	0.67	···	0.00
cock	0.20	0.00	···	0.40
fly	0.07	0.00	···	0.75

WORD	PTB.NN	PTB.VB	···	PTB.JJ
summer	0.96	0.00	···	0.00
spring	0.94	0.02	···	0.00
fall	0.49	0.43	···	0.00
light	0.52	0.02	···	0.41
clear	0.00	0.10	···	0.87

Table 6.1: Oracle linguistic word vectors, constructed from linguistic resources containing semantic/syntactic annotations.

6.3 Word Vector Evaluation Models

QVEC. We align dimensions of distributional word vectors to dimensions (linguistic properties) in the linguistic vectors described in §6.2 to maximize the cumulative correlation of the aligned dimensions. We now formally describe the model.

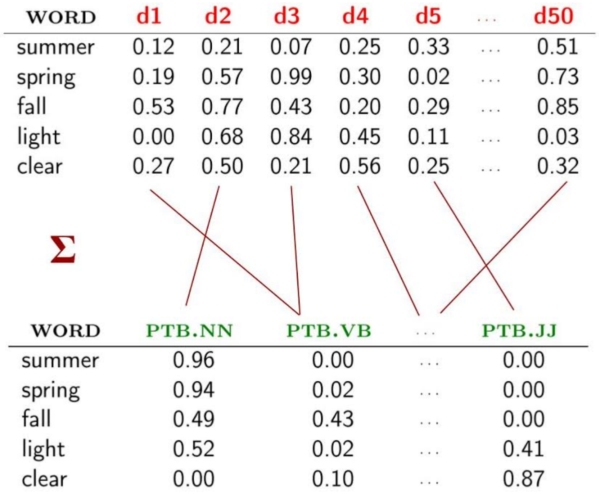
Let the number of common words in the vocabulary of the distributional and linguistic word vectors be N. We define, the distributional vector matrix $\mathbf{X} \in \mathbb{R}^{N \times D_X}$ with every column as a dimension vector $\mathbf{x} \in \mathbb{R}^{N \times 1}$. D_X denotes word vector dimensionality. Similarly, $\mathbf{S} \in \mathbb{R}^{N \times D_S}$ is the linguistic property matrix with every row as a semantic/syntactic property vector $\mathbf{s} \in \mathbb{R}^{N \times 1}$. D_S denotes linguistic properties obtained from a manually-annotated linguistic resource.

[1] We exclude sparser word types to avoid skewed probability estimates of senses of polysemous words.

We obtain an alignment between the word vector dimensions and the linguistic dimensions which maximizes the correlation between the aligned dimensions of the two matrices. This is 1:n alignment: one distributional dimension is aligned to at most one linguistic property, whereas one linguistic property can be aligned to n distributional dimensions. Let $\mathbf{A} \in \{0,1\}^{D_x \times D_s}$ be a matrix of alignments such that $a_{ij} = 1$ iff $\mathbf{x}_i$ is aligned to $\mathbf{s}_j$, otherwise $a_{ij} = 0$. If $r(\mathbf{x}_i, \mathbf{s}_j)$ is the Pearson's correlation between vectors $\mathbf{x}_i$ and $\mathbf{s}_j$, then our objective is defined as:

$$\textsc{qvec} = \max_{\mathbf{A}:\sum_j a_{ij} \leq 1} \sum_{i=1}^{D_X} \sum_{j=1}^{D_S} r(\mathbf{x}_i, \mathbf{s}_j) \times a_{ij}$$

The constraint $\sum_j a_{ij} \leq 1$, warrants that one distributional dimension is aligned to at most one linguistic dimension. The total correlation between two matrices QVEC is our intrinsic evaluation measure of a set of word vectors relative to a set of linguistic properties. An illustration is given in figure 6.1.



WORD	d1	d2	d3	d4	d5	...	d50
summer	0.12	0.21	0.07	0.25	0.33	...	0.51
spring	0.19	0.57	0.99	0.30	0.02	...	0.73
fall	0.53	0.77	0.43	0.20	0.29	...	0.85
light	0.00	0.68	0.84	0.45	0.11	...	0.03
clear	0.27	0.50	0.21	0.56	0.25	...	0.32

WORD	PTB.NN	PTB.VB	...	PTB.JJ
summer	0.96	0.00	...	0.00
spring	0.94	0.02	...	0.00
fall	0.49	0.43	...	0.00
light	0.52	0.02	...	0.41
clear	0.00	0.10	...	0.87

Figure 6.1: The horizontal vectors represent the word vectors: **X** is the upper matrix, it is the distributional matrix, and on the bottom is the semantic/syntactic property matrix **S**. The vertical vectors represent the "distributional dimension vector" in **X** and "linguistic dimension vector" in **S**. QVEC is the sum of alignments between distributional and linguistic vector dimensions that maximize correlation between the dimensions.

The QVEC's underlying hypothesis is that dimensions in distributional vectors correspond to linguistic properties of words. It is motivated, among others, by the effectiveness of word vectors in linear models implying that linear combinations of features (vector dimensions) produce relevant, salient content. Via the alignments a_{ij} we obtain labels on dimensions in the distributional word vectors. The magnitude of the correlation $r(\mathbf{x}_i, \mathbf{s}_j)$ corresponds to the annotation confidence: the higher the correlation, the more salient the linguistic content of the dimension. Clearly, dimensions in the linguistic matrix S do not capture every possible linguistic property, and low correlations often correspond to the missing information in the linguistic matrix. Thus, QVEC is a recall-oriented measure: highly-correlated alignments provide evaluation and annotation of vector dimensions, and missing information or noisy dimensions do not significantly affect the score since the correlations are low.

QVEC-CCA. Although QVEC meets the requirements of a desired evaluation measure for word embedding—it is fast, correlates with downstream tasks, and facilitates interpretability—it suffers from two weaknesses. First, it is not invariant to linear transformations of the embeddings' basis, whereas the bases in word embeddings are generally arbitrary (Szegedy *et al.*, 2014). Second, it produces an unnormalized score: the more dimensions in the embedding matrix the higher the score. This precludes comparison of models of different dimensionality. We now introduce QVEC-CCA, which simultaneously addresses both problems.

To measure correlation between the embedding matrix $\mathbf{X}$ and the linguistic matrix $\mathbf{S}$, instead of cumulative dimension-wise correlation we employ canonical correlation analysis (Hardoon *et al.*, 2004, CCA). CCA finds two sets of basis vectors, one for $\mathbf{X}^\top$ and the other for $\mathbf{S}^\top$, such that the correlations between the projections of the matrices onto these basis vectors are maximized. Formally, CCA finds a pair of basis vectors $\mathbf{v}$ and $\mathbf{w}$ such that

$$\text{QVEC-CCA} = \mathrm{CCA}(\mathbf{X}^\top, \mathbf{S}^\top) = \max_{\mathbf{v},\mathbf{w}} r(\mathbf{X}^\top\mathbf{v}, \mathbf{S}^\top\mathbf{w})$$

Thus, QVEC-CCA ensures invariance to the matrices' bases' rotation, and since it is a single correlation, it produces a score in $[-1, 1]$.

6.4 Experimental Setup

6.4.1 Word Vector Models

To test the QVEC, we select a diverse suite of popular/state-of-the-art word vector models. All vectors are trained on 1 billion tokens (213,093 types) of English Wikipedia corpus with vector dimensionality 50, 100, 200, 300, 500, 1000.

CBOW and Skip-Gram (SG). The WORD2VEC tool (Mikolov *et al.*, 2013b) is fast and widely-used. In the SG model, each word's Huffman code is used as an input to a log-linear classifier with a continuous projection layer and words within a given context window are predicted. In the CBOW model a word is predicted given the context words.[2]

[2] `https://code.google.com/p/word2vec`

CWindow and Structured Skip-Gram (SSG). Ling *et al.* (2015c) propose a syntactic modification to the WORD2VEC models that accounts for word order information, obtaining state-of-the-art performance in syntactic downstream tasks.[3]

CBOW with Attention (Attention). Ling *et al.* (2015b) further improve the WORD2VEC CBOW model by employing an attention model which finds, within the contextual words, the words that are relevant for each prediction. These vectors have been shown to benefit both semantically and syntactically oriented tasks.

GloVe. Global vectors for word representations (Pennington *et al.*, 2014) are trained on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations show interesting linear substructures of the vector space.[4]

Latent Semantic Analysis (LSA). We construct word-word co-occurrence matrix $\mathbf{X}$; every element in the matrix is the pointwise mutual information between the two words (Church and Hanks, 1990). Then, truncated singular value decomposition is applied to factorize $\mathbf{X}$, where we keep the k largest singular values. Low dimensional word vectors of dimension k are obtained from $\mathbf{U_k}$ where $\mathbf{X} \approx \mathbf{U_k}\Sigma\mathbf{V_k}^\top$ (Landauer and Dumais, 1997).

GloVe+WN, GloVe+PPDB, LSA+WN, LSA+PPDB. We use retrofitting (Faruqui *et al.*, 2015) as a post-processing step to enrich GloVe and LSA vectors with semantic information from WordNet and Paraphrase database (PPDB) (Ganitkevitch *et al.*, 2013).[5]

6.4.2 Evaluation benchmarks

We compare the QVEC to eight standard semantic and syntactic tasks for evaluating word vectors; we now briefly describe the tasks.

Word Similarity. We use three different benchmarks to measure word similarity. The first one is the **WS-353** dataset (Finkelstein *et al.*, 2001), which contains 353 pairs of English words that have been assigned similarity ratings by humans. The second is the **MEN** dataset (Bruni *et al.*, 2012) of 3,000 words pairs sampled from words that occur at least 700 times in a large web corpus. The third dataset is **SimLex-999** (Hill *et al.*, 2014) which has been constructed to overcome the shortcomings of WS-353 and contains 999 pairs of adjectives, nouns and verbs. Word similarity is computed using cosine similarity between two words and the performance

[3] https://github.com/wlin12/wang2vec

[4] http://www-nlp.stanford.edu/projects/glove/

[5] https://github.com/mfaruqui/retrofitting

of word vectors is computed by Spearman's rank correlation between the rankings produced by vector model against the human rankings.[6]

Text Classification. We consider four binary categorization tasks from the 20 Newsgroups (**20NG**) dataset.[7] Each task involves categorizing a document according to two related categories with training/dev/test split in accordance with Yogatama and Smith (2014). For example, a classification task is between two categories of Sports: baseball vs hockey. We report the average classification accuracy across the four tasks. Our next downstream semantic task is the sentiment analysis task (**Senti**) (Socher *et al.*, 2013b) which is a binary classification task between positive and negative movie reviews using the standard training/dev/test split and report accuracy on the test set. In both cases, we use the average of the word vectors of words in a document (and sentence, respectively) and use them as features in an ℓ_2-regularized logistic regression classifier. Finally, we evaluate vectors on the metaphor detection method presented in chapter 3 (**Metaphor**) (Tsvetkov *et al.*, 2014c).[8] The system uses word vectors as features in a random forest classifier to label adjective-noun pairs as literal/metaphoric. We report the system accuracy in 5-fold cross validation.

Syntactic Benchmarks. The first out of two syntactic tasks is **POS** tagging. We use the LSTM-CRF model (Lample *et al.*, 2016), which trains a forward-backward LSTM on a given sequence of words (represented as word vectors). The hidden are then used as the sole features in a CRF model (Lafferty *et al.*, 2001) to predict the output label sequence. The tagger is trained to predict the POS tags and evaluated with the standard Penn TreeBank (PTB) (Marcus *et al.*, 1993) training, development and test set splits as described by Collins (2002). The second syntactic benchmark is dependency parsing (**Parse**). We use the stack-LSTM model of Dyer *et al.* (2015) for English on the universal dependencies treebank (Agić *et al.*, 2015) with the standard development and test splits, reporting unlabeled attachment scores (UAS) on the test data. We remove all part-of-speech and morphology features from the data, and prevent the model from optimizing the word embeddings used to represent each word in the corpus, thereby forcing the parser to rely completely on the pretrained embeddings.

[6] We employ an implementation of a suite of word similarity tasks at `wordvectors.org` (Faruqui and Dyer, 2014a).

[7] `http://qwone.com/~jason/20Newsgroups`

[8] `https://github.com/ytsvetko/metaphor`

6.5 Results

6.5.1 Linear correlation

To test the efficiency of QVEC and QVEC-CCA in capturing the semantic content of word vectors, we evaluate how well their scores correspond to the scores of word vector models on benchmarks described in §6.4.2. We compute the Pearson's correlation coefficient r to quantify the linear relationship between the scorings. We begin with comparison of QVEC with one extrinsic task—Senti—evaluating 300-dimensional vectors. To account for variance in WORD2VEC representations (due to their random initialization and negative sampling strategies, the representations are different for each run of the model), and to compare the QVECs to a larger set of vectors, we now train three versions of vector sets per model. This results in 21 word vector sets: three vector sets per five WORD2VEC models plus GloVe, LSA, and retrofitting vectors. In the rest of this section we report results on the extended suite of word vectors. The Pearson's correlation computed on the extended set of comparison points is shown in table 6.2. Both QVEC and QVEC-CCA obtain high positive correlation with the Senti task.

	Model	Senti	QVEC	QVEC-CCA
1.	CBOW-1	81.0	40.3	.498
2.	CBOW-2	80.6	39.6	.498
3.	CBOW-3	80.0	40.9	.498
4.	SG-1	80.5	35.9	.499
5.	SG-2	81.2	36.3	.499
6.	SG-3	80.5	36.9	.499
7.	CWindow-1	76.2	28.1	.453
8.	CWindow-2	77.4	27.0	.453
9.	CWindow-3	77.5	28.3	.457
10.	SSG-1	81.5	40.5	.512
11.	SSG-2	80.6	41.0	.511
12.	SSG-3	81.3	41.3	.510
13.	Attention-1	80.1	40.8	.502
14.	Attention-2	80.3	41.0	.502
15.	Attention-3	80.4	40.6	.502
16.	GloVe	79.4	34.4	.461
17.	GloVe+WN	79.6	42.1	.487
18.	GloVe+PPDB	79.7	39.2	.478
19.	LSA	76.9	19.8	.419
20.	LSA+WN	77.5	29.4	.447
21.	LSA+PPDB	77.3	28.4	.442
	Correlation (r)		0.88	0.93

Table 6.2: Intrinsic (QVEC/QVEC-CCA) and extrinsic scores of the 300-dimensional vectors trained using different word vector models and evaluated on the Senti task. Both QVEC and QVEC-CCA obtain high positive Pearson's correlation between the intrinsic and extrinsic scores.

We now extend the table 6.2 results, and show correlations between the intrinsic and extrinsic scores across semantic benchmarks for the same 300-dimensional vectors. In table 6.3, we show that correlations between our proposed intrinsic scores (QVEC/QVEC-CCA) and the extrinsic scores are on par with or higher than the correlation between the word similarity and text classification tasks. QVEC-CCA obtains high positive correlation with all the semantic tasks, and outperforms QVEC on two tasks.

	20NG	Metaphor	Senti
WS-353	0.55	0.25	0.46
MEN	0.76	0.49	0.55
SimLex	0.56	0.44	0.51
QVEC	0.74	0.75	0.88
QVEC-CCA	0.77	0.73	0.93

Table 6.3: Pearson's correlations between word similarity/QVEC/QVEC-CCA scores and the downstream text classification tasks.

In table 6.4, we evaluate QVEC and QVEC-CCA on syntactic benchmarks. We first use linguistic vectors with dimensions corresponding to part-of-speech tags (denoted as PTB). Then, we use linguistic vectors which are a concatenation of the semantic and syntactic matrices described in §6.2 for words that occur in both matrices; this setup is denoted as PTB+SST.

		POS	Parse
	WS-353	-0.38	0.68
	MEN	-0.32	0.51
	SimLex	0.20	-0.21
PTB	QVEC	0.23	0.39
	QVEC-CCA	0.23	0.50
PTB+SST	QVEC	0.28	0.37
	QVEC-CCA	0.23	0.63

Table 6.4: Pearson's correlations between word similarity/QVEC/QVEC-CCA scores and the downstream syntactic tasks.

Although some word similarity tasks obtain high correlations with syntactic applications, these results are inconsistent, and vary from a high negative to a high positive correlation. Conversely, QVEC and QVEC-CCA consistently obtain moderate-to-high positive correlations with the downstream tasks.

Comparing performance of QVEC-CCA in PTB and PTB+SST setups sheds light on the importance of linguistic signals captured by the linguistic matrices. Appending supersense-annotated columns to the linguistic matrix which already contains POS-annotated columns does not affect correlations of QVEC-CCA with the POS tagging task, since the additional linguistic information is not relevant for approximating how well dimensions of word embeddings encode POS-related properties. In the case of dependency parsing—the task which encodes not only syntactic, but also semantic information (e.g., captured by subject-verb-object relations)—supersenses introduce relevant linguistic signals that are not present in POS-annotated columns. Thus, appending supersense-annotated columns to the linguistic matrix improves correlation of QVEC-CCA with the dependency parsing task.

6.5.2 Rank correlation

Since QVEC favors recall over precision, larger numbers of dimensions will *ceteris paribus* result in higher scores—but not necessarily higher correlations with downstream tasks. We therefore impose the restriction that QVEC only be used to compare vectors of the same size, but we now show that its correlation with downstream tasks is stable, conditional on the size of the vectors being compared. We aggregate rankings by individual downstream tasks into a global ranking using the Kemeny–Young rank aggregation algorithm, for each dimension separately (Kemeny, 1959). The algorithm finds a ranking which minimizes pairwise disagreement of individual rankers. Table 6.5 shows Spearman's rank correlation between the rankings produced by the QVEC and the Senti task/the aggregated ranking. For example, ranking of 300-dimensional models produced by Senti is *{SSG, CBOW, SG, Attention, GloVe+PPDB, GloVe+WN, GloVe, LSA+WN, LSA+PPDB, LSA, CWindow}*, and the QVEC's ranking is *{GloVe+WN, Attention, SSG, CBOW, GloVe+PPDB, SG, GloVe, LSA+WN, LSA+PPDB, CWindow, LSA}*. The Spearman's ρ between the two rankings is 0.78. We note, however, that there is a considerable variation between rankings across all models and across all dimensions, for example the SimLex ranking produced for the same 300-dimensional vectors is *{GloVe+PPDB, GloVe+WN, SG, LSA+PPDB, SSG, CBOW, Attention, CWindow, LSA+WN, GloVe, LSA}*, and ρ(Senti, SimLex) = 0.46. In a recent related study, Schnabel *et al.* (2015) also observe that existing word similarity and text categorization evaluations yield different orderings of word vector models. This task-specifity of rankings emphasizes the deficiency of evaluating word vector models solely on downstream tasks, and the need of a standardized intrinsic evaluation approach that quantifies linguistic content of word vectors.

	50	100	200	300	500	1000
ρ(QVEC, Senti)	0.32	0.57	0.73	0.78	0.72	0.60
ρ(QVEC, All)	0.66	0.59	0.63	0.65	0.62	0.59

Table 6.5: Spearman's rank-order correlation between the QVEC ranking of the word vector models and the ranking produced by (1) the Senti task, or (2) the aggregated ranking of all tasks (All). We rank separately models of vectors of different dimensionality (table columns).

6.6 Summary

We introduced an approach to intrinsic evaluation of word embeddings quantifying word vector content with respect to desired linguistic properties. We showed that both QVEC and QVEC-CCA show strong correlations with downstream semantic and syntactic tasks. Semantic and syntactic linguistic features that we use to construct linguistic dimension matrices are rather coarse, thus the proposed evaluation can approximate a range of downstream tasks, but may not be sufficient to evaluate finer-grained task-specific features. In the future work we propose to investigate the problem of interpreting word vector in depth, and exploit existing monolingual and multilingual resources to construct better linguistic matrices, suited for task-specific evaluations across a range of NLP tasks.

Chapter 7

Conclusion

In this thesis, I argue that cultivating a symbiosis between machine learning and linguistic theory is not only an opportunity, it is crucial for NLP. Scientifically, awareness to linguistic research is essential for understanding which research questions to pose and how to define tasks processing language, which resources to construct, how to define methodologies to effectively exploit resources and knowledge, what knowledge is expected to be captured by machine learning models, and how to evaluate and analyze the models. Practically, such a symbiosis can alleviate the inherent limits of relevant data and facilitate building better NLP models by exploiting deeper linguistic knowledge acquired throughout decades of linguistic research.

Each chapter in this thesis is a case study that exemplifies the benefit of interdisciplinary, linguistically-grounded language processing. Each chapter poses a novel research question that is motivated by (i) understanding the weaknesses of contemporary computational approaches, and (ii) awareness to linguistic studies addressing the same question. Each chapter introduces a robust computational framework—operationalizing linguistic insights—that learns linguistically-informed generalizations and improves practical NLP problems. In the following section, I summarize the contributions of this thesis.

7.1 Summary of Contributions

- I introduce the task of computationally modeling lexical borrowing, and provide the first computational model of lexical borrowing used in a downstream NLP task. I show that lexical correspondences induced using this model can project resources—namely, translations—leading to improved performance in a downstream translation system.
- While there have been software implementations of OT (Hayes *et al.*, 2013), they have been used chiefly to facilitate linguistic analysis; I show how to use OT to formulate a model that can be learned with less supervision than linguistically naïve models.

- I develop a state-of-the-art English metaphor detection system that uses *conceptual* semantic features, such as a degree of abstractness and semantic supersenses.

- I introduce the task of cross-lingual metaphor detection. Using a paradigm of model transfer, I provide support for the hypothesis that metaphors are conceptual (rather than lexical) in nature by showing that our English-trained model can detect metaphors in Spanish, Farsi, and Russian.

- I introduce the task of optimizing the curriculum, and provide the first computational framework that formulates curriculum learning as an optimization problem, rather than shuffling data or relying on human intuitions. I show that optimizing the curriculum improves the performance of a range of NLP tasks that employ word embeddings as features. The framework is general: although I experiment with optimizing the curriculum of word embeddings, the curriculum of other models can be optimized in a similar way.

- I conduct the first study that analyzes the impact of distributional and linguistic properties of training texts on the quality of task-specific word embeddings.

- I introduce a novel polyglot neural phonological language model architecture, and show how to integrate linguistic typology to mediate between languages and improve generalization performance. Conditioning not only on the identity of the language being predicted in each sequence, but also on a vector representation of its phono-typological properties, the multilingual model is more effective than individual monolingual models, and substantially more effective than naive baselines.

- I show that downstream applications are improved by using polyglot-learned phone representations. In borrowing, integrating polyglot features improves over the state-of-the-art models presented in chapter 2, and in speech synthesis, polyglot features are more effective than manually-designed phonetic features.

- I formulate the desiderata for intrinsic evaluation of distributed representations of words, introduce the task of evaluating word embeddings by their subspace alignment to linguistic resources, and propose two novel evaluation methods that obtain higher correlation with downstream semantic and syntactic applications than standard approaches to evaluation which rely on word similarity and relatedness. In this work, I also make first steps towards interpreting knowledge encoded in distributed representations.

7.2 Future Work

In the future work, many more research questions can be addressed using the holistic, interdisciplinary view in the spirit of this thesis. I discuss below some of the immediate future directions. At a higher level, developing language-aware models that use linguistic knowledge to compensate for dearth of resources and to learn better generalizations can improve speech and text translation systems, educational and healthcare applications, and dialog systems for thousands of languages.

This thesis is focused on leveraging linguistic knowledge to benefit NLP. In the other direction, linguistics can also benefit greatly from exploiting computational models. Sophisticated probabilistic models can guide language inquiry, and provide new insight into languages, their nuts and bolts, their life cycle, their role. Among some inspiring examples that automate linguistic inquiry are models for discovering linguistic typology (Daumé III and Campbell, 2007), for reconstruction of ancient languages (Bouchard-Côté *et al.*, 2013), for language acquisition (Kwiatkowski *et al.*, 2012), for understanding properties of translationese (Rabinovich and Wintner, 2015), for discovery and tracking word meaning change (Kulkarni *et al.*, 2015), to name a few.

7.2.1 Future work on models of lexical borrowing

Models of lexical borrowing for typologically diverse languages. To establish the robustness and generality of the proposed approach to modeling lexical borrowing, and to practically contribute to more languages, in future work I propose to build borrowing systems for a diverse set of languages, listed in table 7.1. The list of focus languages was selected based on the following considerations:

- Recipients are susceptible to borrowing, typologically diverse, and also diverse in terms of availability of resources, from resource-rich English to extremely resource-poor Yoruba. We also paid attention to selecting recipient languages that have more than one prominent donor language.
- Donors are resource-rich and typologically diverse.
- Donor–recipient language pairs have been extensively studied in contact linguistics, in particular using Optimality Theoretic view, and linguistic studies contain lexicons of donor–loanword examples that we can use as training sets.
- Availability of pronunciation dictionaries (or phonographic writing system); availability of small parallel corpora, dependency treebanks and POS tagged corpora to enable extrinsic evaluation; and, when possible, language presence it Twitter data.

Recipient	Family	# speakers (millions)	Resource-rich donors	Studies on lexical borrowing
English	Germanic	1500	French, German, Spanish, Arabic	(Grant, 2009; Durkin, 2014)
Russian	Slavic	260	English	(Benson, 1959; Styblo Jr, 2007)
Japanese	Altaic	125	English, Chinese	(Kay, 1995; Daulton, 2008; Schmidt, 2009)
Vietnamese	Austro-Asiatic	80	Chinese, French	(Barker, 1969; Alves, 2009)
Korean	Koreanic	75	English, Chinese	(Kang, 2003; Kenstowicz, 2005)
Swahili	Bantu	40	Arabic	(Mwita, 2009; Schadeberg, 2009)
Yoruba	Edekiri	28	English	(Ojo, 1977; Kenstowicz, 2006)
Romanian	Romance	24	French	(Friesner, 2009; Schulte, 2009)
Quechua	Quechuan	9	Spanish	(Rendón, 2008; Rendón and Adelaar, 2009)
Hebrew	Semitic	7.5	English, Arabic	(Schwarzwald, 1998; Cohen, 2009, 2013)
Finnish	Uralic	5	German, English	(Orešnik, 1982; Kallio, 2006; Johnson, 2014)

Table 7.1: Recipient and donor languages to be used for borrowing systems constructed in this project.

Models of lexical borrowing in cross-lingual core-NLP. Approaches to low-resource NLP can roughly be categorized into unsupervised (Christodoulopoulos *et al.*, 2010; Berg-Kirkpatrick and Klein, 2010; Cohen *et al.*, 2011; Naseem *et al.*, 2012), and semi-supervised methods relying on cross-lingual projection from resource-rich languages via parallel data. Cross-lingual semi-supervised learning strategies of syntactic structure—both shallow and deep—have been a promising vein of research over the last years; we will continue this research direction, but will augment or replace word alignments with multilingual borrowing lexicons.

The majority of semi-supervised approaches project annotations from English, both in POS tagging (Yarowsky *et al.*, 2001; Das and Petrov, 2011; Li *et al.*, 2012; Täckström *et al.*, 2013), and in induction of dependency relations (Wu, 1997; Kuhn, 2004; Smith and Smith, 2004; Hwa *et al.*, 2005; Xi and Hwa, 2005; Burkett and Klein, 2008; Snyder *et al.*, 2009; Ganchev *et al.*, 2009; Tiedemann, 2014). I hypothesize that in the case of syntactic annotation, projection via borrowing is a more plausible method than projection via word alignment, especially when large parallel corpora are unavailable. First, projecting via borrowing will enable annotation projection from typologically more similar languages to the target language. For example, some Austronesian languages (e.g., Malagasy) have borrowed extensively from Arabic, and these are typologically closer to Arabic than to English (e.g., like Arabic, these are verb-subject-object languages). Second, linguistic studies report statistics on "borrowability" of syntactic categories pertaining to specific languages; Haspelmath and Tadmor (2009), for example, provide relative percentages of loan nouns, verbs, adjectives, and adverbs for 41 languages. This prior knowledge can be used to bias the learned taggers/parsers. Finally, integrating borrowing will allow to use signals from multiple source languages, which have proven to work better than solely exploiting English (McDonald *et al.*, 2011b; Berg-Kirkpatrick and Klein, 2010; Cohen *et al.*, 2011; Naseem *et al.*, 2012).

I propose experiment with existing cross-lingual approaches to POS tagging (cf. Täckström *et al.*, 2013), adapted to projecting via donor–loan lexicons. These approaches can employ borrowing both in projecting token annotations (in addition to or instead of parallel corpora), and in projecting tag dictionaries. Similarly, techniques for cross-lingual dependency grammar induction (cf. McDonald *et al.*, 2011b) can be adapted. Prior work on non-parallel cross-lingual structure prediction has already shown that phylogenetic relations are a valuable signal (Berg-Kirkpatrick and Klein, 2010), and this vein of research is worth pursuing focusing on non-parallel projection via borrowing.

The universal POS tagset (Petrov *et al.*, 2012) and the universal dependency treebank (McDonald *et al.*, 2013) can be used to train resource-rich taggers/parsers, the sources of projection. Leipzig Corpora Collection (Biemann *et al.*, 2007) can be used as a source of monolingual data, and publicly available parallel or multi-parallel corpora such as WIT^3 (translated TED talks) (Cettolo *et al.*, 2012) and Common Crawl parallel data (Smith *et al.*, 2013) for translation and cross-lingual projection. In addition, Wiktionary[1] can be used as a source of multilingual tag dictionaries (cf. Li *et al.*, 2012; Garrette and Baldridge, 2013), and the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013b) as a source of linguistic typological features (cf. Naseem *et al.*, 2012). An accuracy our taggers and parsers will be evaluated on existing tagged and parsed corpora (e.g., Erjavec, 2004; Hajič *et al.*, 2009). In addition, the effectiveness of this methodology can be evaluated by simulating resource-rich to resource-poor scenario in a resource-rich language pair (e.g., English–Spanish, English–French). In addition to manual qualitative evaluation and error analysis, it would be interesting to investigate the impact of typological similarity on the quality of cross-lingual projection.

Modeling the sociolinguistic aspects of lexical borrowing. As pointed out by Appel and Muysken (2005, p. 174), the literature on lexical borrowing combines an immense number of case studies with the absence of general works. Linguistic case studies focus on morpho-phonological properties of lexical, morphological, phonological, and grammatical borrowing; sociolinguistic case studies explore social and cultural determinants of borrowing. While in the first phases of this research we will develop general data-driven approaches to identify and model lexical borrowing as a linguistic phenomenon, the goal of the current phase is to develop general approaches to statistically model borrowing as a social and cultural phenomenon. We will focus on two research questions: (1) what are demographic and geographic conditions to the emergence and diffusion of borrowing?; and (2) is there a correlation between socio-political status quo in contact communities and meaning/semantic polarity that loanwords acquire in borrowing? Answering these questions using robust statistical methods will help researchers better understand and exploit lexical borrowing in NLP applications. In addition, these computational studies will corroborate or refute the generality of claims made in individual sociolinguistic case studies on borrowing, thereby contributing to sociolinguistic

[1] `www.wiktionary.org`

research.

Emergence and diffusion of borrowing. Linguistic borrowing emerges in bilingualism over time; wider-spread and longer language contact triggers more common and more far-reaching borrowing (McMahon, 1994). Twitter and other social media corpora provide particularly fruitful conditions to investigate the emergence and diffusion of lexical borrowing: these texts are an example of rapid language development, of emergence and diffusion of neologisms (Androutsopoulos, 2000; Anis, 2007; Herring, 2012), of dense interaction between diverse multilingual populations all over the world. Importantly, written neologisms—"neography" (Anis, 2007)—are often phonetic forms of spoken words, and, as recent research shows, are influenced by the phonological system of the source language (Eisenstein, 2013a, 2015). Presence of morpho-phonological influence indicates some level of linguistic adaptation, i.e., borrowing. While previous studies investigated the emergence and diffusion of lexical change in English Twitter data, relying only on word frequencies (Eisenstein *et al.*, 2014), the goal of this proposed study is to extend this research direction to multiple languages/dialects, and to explore examples containing morpho-phonological signals.

First, it would be interesting to learn to distinguish borrowed items among all neologisms, using the classifiers and learned morpho-phonological features from the models of borrowing or the phonological polyglot models. Then, it would be interesting to investigate demographic and geographic factors influencing borrowing and its linguistic pathways (cf. Eisenstein *et al.*, 2014): web-user age, social status, influence, geographic location, and metropolitan area influence. Our research hypotheses, based on linguistic accounts (Guy, 1990; McMahon, 1994; Sankoff, 2002; Appel and Muysken, 2005, among many others), are that borrowing emerges "from above", i.e., being initiated by higher-influence groups (of users, locations, languages) and adopted "from below", by lower-influence groups (as in the Labov's seminal work on dialect importation (1966)); that while phonetic transcription and context-switching are used by younger users, borrowing requires mature linguistic experience of adults to correctly identify and adopt the borrowed features.

Relation of prestige to sentiment polarity of borrowing. Susceptibility of a language to borrowing is determined by both linguistic and socio-political factors. Linguistic considerations involve attempts to fill a lexical gap for a concept that has no word in the recipient language (new information or technology, contact with foreign culture, or foreign flora and fauna, etc.). Socio-political factors are determined by political and social prestige. Presumably, speakers borrow from a prestige-group, although not always—when loanwords enter from less prestigious to more prestigious language, these often connote derogatory meanings (McMahon, 1994). For example, Romanian loanwords from Slavic tend to have positive semantic orientation, while loanwords from Turkish have more negative connotations, e.g., *prieten* 'friend'

(Romanian, borrowed from Slavic *prijatel*) vs. *dușman* ‘enemy’ (from Turkish *düşman*) (Schulte, 2009). Turkish, in turn, frequently borrowed words from Arabic and Persian, but with the founding of the modern Turkish state and the related rise of nationalism and national identity, this borrowing ceased more or less instantly (Curnow, 2001). I propose to conduct a qualitative study to measure correlation between sentiment polarity of borrowed words and prestige relations of lending/borrowing users.

Studies on emergence and diffusion of borrowing, and on dynamics in semantics of borrowing can shed new light to the longstanding goal of social studies—linguistic and computational—understanding the relationship between social forces and linguistic outcomes. Practically, this study can result in new analytics tools that can improve social media applications analyzing and predicting human behavior, e.g., predicting public opinion from social media feeds (O’Connor *et al.*, 2010), predicting personal attributes like geographic location, age, gender, and race/ethnicity of social media users (Nguyen *et al.*, 2011; Mislove *et al.*, 2011), measuring power differences between individuals based the patterns of words they speak or write (Danescu-Niculescu-Mizil *et al.*, 2012).

In the proposed experiments geo-tagged Twitter data can be used from the Twitter Gardenhose stream, which is a sample of approximately 10% of all publicly available Twitter messages (tweets). In addition to the messages themselves, the corpus contains information about the identity of the message sender, whether or not it was a “retweet” of another message, partial information from the “follower graph,” and, for some subset of the tweets, geographic location of the message sender (which can be, e.g., correlated with demographic information). All donor and recipient focus languages—except Swahili, Yoruba, and Quechua—are well-represented in the CMU Twitter archive (table 7.2).

Language	# of tweets
English	12,194,934
Japanese	6,257,572
Spanish	4,444,441
Arabic	1,963,931
French	792,591
Korean	456,838
Russian	428,989
German	184,889
Finnish	57,214
Vietnamese	17,594
Chinese	14,435
Romanian	7,555
Hebrew	4,835

Table 7.2: Statistics of the multilingual tweets.

7.2.2 Future work on metaphor detection

Direct extensions of chapter 3 are to apply the proposed technique to additional syntactic constructions, e.g., noun-noun (NN) metaphors, to improve cross-lingual model transfer with more careful cross-lingual feature projection, and to conduct an in-depth qualitative analysis of types of English SVO, AN, and NN metaphors that can be identified in other languages using our method. More interesting extensions include: (i) Monolingual and cross-lingual identification of *contextual* metaphors, i.e., expressions that function either as metaphorical or as literal, depending on a broader context of the sentence (cf. Jang *et al.*, 2016); (ii) With cross-lingual model transfer, adapting to multilinguality a recently proposed multimodal approach to identification of English metaphors using visual cues (Shutova *et al.*, 2016). In this setup, it would be especially interesting to learn whether cross-lingual lexical discrepancies in metaphorical usages are smoothed in visual signals; (iii) In socio-linguistic research, no prior computational work has done an extensive analysis of whether using metaphors in public speeches affects engagement or agreement from listeners/commenters (cf. Niculae and Danescu-Niculescu-Mizil, 2014), whether metaphors actually shape listeners' opinions as articulated by Lakoff and Johnson (1980); and finally, (iv) Exploiting the broad research on metaphor detection and interpretation by shifting the task into more practical settings, by integrating metaphor detection systems in other semantic tasks, e.g., semantic parsing, question answering, modeling argument structure, and others.

7.2.3 Future work on learning the curriculum

In chapter 4, I introduced a general framework for characterizing training data using rich linguistic knowledge, and then optimizing the curriculum based on data properties. The framework is the first step in linguistically and statistically analyzing training data for optimizing downstream performance in time-course learning. Multiple research questions can be posed to further rigorously study desirable properties of training data and curricula, across applications, and across approaches to optimization.

Learning the curriculum across NLP tasks, across linguistic domains, and across data characteristics. Learning the curriculum through optimizing word vector representations allowed us to evaluate the proposed framework across a diverse set of tasks, and thereby obtain a strong evidence of feasibility of the framework. We have corroborated our hypothesis that curriculum optimization is a promising new research direction. Further in-depth analysis of the framework and its variations is essential, however, to clearly understand characteristics of training data and curricula that are optimal for different learners. Had we optimized the task-specific neural architectures directly, using standard training sets and not out-of-domain training data from Wikipedia, the effect of curriculum could have been more pronounced. Had we optimized the models on data from other domains, genres, and languages, our findings would have been different. Finally, we learned weights for each group of features separately; a more detailed understanding of the proposed and other statistical and linguistic features and feature interactions will provide new insights into properties and into needs of existing data sets.

Using the architecture established in §4.2, in the future work I propose to investigate the curriculum learning problem along three (non-orthogonal) directions of research: (i) the varieties of training data and their impact on learning distributed word representations; (ii) the effect of curriculum learning on a variety of neural NLP models; and (iii) the role of linguistic and distributional, functional and structural properties of data in characterizing the curricula.

The biggest gains from the deep learning paradigm shift were achieved in NLP through the use of distributed word and concept representations (Manning, 2016). However, the majority of current research related to distributed representations of words—either focusing on improving learned representations for downstream applications (e.g., Ling *et al.*, 2015c; Levy and Goldberg, 2014; Faruqui *et al.*, 2015), or exploring various data-driven models that utilize embeddings as features (Lazaridou *et al.*, 2013; Bansal *et al.*, 2014; Guo *et al.*, 2014; Socher *et al.*, 2013b, among dozens)—resort to Wikipedia dumps as the default source of texts to train word vectors. No studies have been conducted to understand the impact of a specific variety of text—quality and type of text, domain, genre, register—as well as the impact of ordering of training data on the word vectors that are used as features in various tasks. This research direction is worth exploring, given the well-known contrast in efficacy of in- vs. out-of-domain training data in supervised and unsupervised problems, given a massive body of research on topic/domain adaptation (Foster *et al.*, 2010; Blitzer *et al.*, 2007; Daumé *et al.*, 2010; Aue and Gamon, 2005; Axelrod *et al.*, 2011), data selection (Eetemadi *et al.*, 2015; Lü *et al.*, 2007; Moore and Lewis, 2010), normalization of noisy data (Eisenstein, 2013b; Gimpel *et al.*, 2011; Han and Baldwin, 2011), and even studying the effect of a source language on translated training texts (Lembersky *et al.*, 2012, 2013). Using the proposed curriculum-optimization architecture, it would be interesting to explore the varieties of training data—novels, newswire, noisy social media texts, spoken language—mixed with Wikipedia texts or used separately, and their contribution to the effectiveness of word embeddings used in different NLP tasks.

It would be also interesting to study the effect of curriculum directly on the tasks, rather than via pre-trained word embeddings. In neural models, I propose to explore two cornerstone approaches to sequential decision making that exploit local and longer-distance history of observations and make a strategized structured predictions. One is the supervised sequence-to-sequence learning (Cho *et al.*, 2014; Sutskever *et al.*, 2014) with Long Short-Term Memory (LSTM) units (Hochreiter and Schmidhuber, 1997; Graves, 2013); and the other is deep reinforcement learning (Sutton and Barto, 1998; Mnih *et al.*, 2013). The former has proven itself as so far the most effective neural architecture for sequential language processing, used in a variety of NLP tasks (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2015; Ling *et al.*, 2015a; Dyer *et al.*, 2015; Lample *et al.*, 2016; Faruqui *et al.*, 2016, *inter alia*). The latter is gaining popularity in language processing (Narasimhan *et al.*, 2015; He *et al.*, 2016; Ranzato *et al.*, 2016; Li *et al.*, 2016) after its tremendous success in related research domains of artificial intelligence (Mnih *et al.*, 2015). Concrete applications for learning with sequence-to-sequence LSTMs will include *parsing*—a core-NLP problem that obtains state-of-the-art results when implemented with neural models both for dependency Dyer *et al.* (2015) and phrase structure grammars (Luong *et al.*, 2016)—and *discourse modeling* with reinforcement learning, that builds upon sequence-to-sequence models, but also optimizes long-term rewards (cf. Li *et al.*, 2016). In all these use-cases, it would be interesting to investigate the impact of functional and structural properties of curriculum on the task. In addition, it is important to study the relationship between a curriculum and specific design decisions of the learning models: is there a relationship between a size of the network, as defined by its hyperparameters or policy, and a curriculum, quality, and quantity of training data?

On the axis of data attributes, I propose to explore how training data characteristics—i.e., features corresponding to content, distributional properties, and structural patterns of data—affect learning across text varieties, neural models, and language processing tasks. The goal in this line of experiments is to identify subsets of features and feature combinations that are most informative to determine an optimal curriculum in curriculum-learning experiments proposed above. The research questions include (i) what features are universally effective across tasks and can be used, for example for multitask learning, and what features and feature combinations are effective in task-specific settings; (ii) how knowledge-rich features can be substituted by unsupervised features to facilitate curriculum learning in lower-resource domains; (iii) how to define and integrate in the proposed framework more complex, structural features that cross word- and sentence-boundaries. In addition to the proposed features in and their combinations, new hand-annotated features can be introduced, and features extracted from unlabeled data. Manually and semi-manually annotated resources can include knowledge bases, such as Freebase (Bollacker *et al.*, 2008) or NELL (Mitchell *et al.*, 2015) for extracting semantic relations, and the Universal Dependency Treebank (Agić *et al.*, 2015) for morpho-syntactic annotations. It is also important to explore a set of word-level semantic and syntactic properties that can be induced from unlabeled corpora in an unsupervised manner, e.g., Brown clusters (Brown *et al.*, 1992). If effective in English, these can then be ported to languages lacking manually-labeled resources.

Finally, it is worth exploring semantic and syntactic structural data characteristics. Inspired by prior research in summarization (Erkan and Radev, 2004), a group of features quantifying text coherence can be defined. In this line of work, text is represented as a network: an undirected weighted graph, with sentences (or paragraphs) as vertices, and thresholded sentence (syntactic or semantic) relatedness between sentences as edges. Subgraphs in this graph correspond to groupings of semantically/syntactically related sentences, and to quantify the sentence/subgraph importance, standard features from network analysis can be employed, such as subgraph density, degree centrality, eigenvector centrality, and others (Freeman, 1978). In the proposed setup, a curriculum is defined not as ordering of single unrelated sentences, but on chunks of text of longer context—groupings of sentences that are represented as subgraphs in the semantic network. Coherence-based features will facilitate discovery of a preferred organization of concepts in data at training time, and will answer the following questions: For a selected task, is it better to group related concepts or to mix them with unrelated concepts and increase network diversity? What is the size of of history to keep? How these preferences change in time-course training? These features, as well as the whole framework for learning the curriculum, can potentially be useful not only for computational curriculum learning, but also in psycholinguistic experiments, for example to corroborate a hypothesis proposed by the Competition Model (MacWhinney, 1987; Bates and MacWhinney, 1989), that lexical items compete with each other during comprehension and production.

Improving the framework for learning the curriculum. Overfitting is a significant challenge for Bayesian optimization (Snoek *et al.*, 2012), and this is one reason why in our experiments we used limited subsets of features, to keep the overall small number of hyperparameters to optimize. This restriction hinders constructing an expressive framework with rich features and their structural relations, and exploiting fully the time-course learning. There are several directions to explore to alleviate this problem. One option is to fully exploit the TPE algorithm that allows to incorporate structural feature relations by manually defining feature interdependencies (cf. Yogatama *et al.*, 2015); this would allow to better inspect relevant feature spaces in the current framework. To integrate more features, the TPE algorithm can be replaced by the Random EMbedding Bayesian Optimization (REMBO) algorithm (Wang *et al.*, 2013, 2016) which is better suited for optimizing high-dimensional spaces. Prior to optimization, REMBO reduces the dimensionality of the input feature vector using multiplication by a random matrix.

As an alternative to Bayesian optimization, reinforcement learning (RL) can be used in the same framework. As well as in Bayesian optimization, the classic balance between exploration and exploitation is a key part of reinforcement learning: the RL agent should find more information about the environment (exploration) in order to discover a good policy—the part of the learning space that maximizes the reward—and should exploit known information effectively (exploitation). RL is also well-suited to sequential, time-course learning advocated in chapter 4: the *eval* function from §4.2 can be directly translated to the reward R_t that indicates how well the learning progresses at step t. Given an observation O_t—training examples following a curriculum comprising the full corpus or a batch from $\mathcal{X}_t$—and the reward, the

RL agent will emit an action A_t—in §4.2 this is $\mathbf{w}_{t+1}$, an assignment of feature weights to explore—which maximized the reward according to an optimal policy. In the proposed RL framework, a policy defines a mapping from the current state S_t^e to the selection of the next sorted sequence of training examples to process. In §4.2, S_t^e is defined as a simple linear function $\mathbf{w}^\intercal \boldsymbol{\phi}(\mathcal{X})$, but in the RL framework it can capture richer knowledge including both specification of what was learned by the model by now (e.g., extracted using QVEC method but with a linguistic matrix that represents the task), and specification of the remaining training data represented using local and global feature weights, structural relations, etc. Hutter (2009) developed a method that given a complex specification of an environment automatically selects features that are necessary and sufficient for reducing the problem to a computationally tractable Markov Decision Processes. Thus, rich functional and structural features as well as additional knowledge of what has already been learned by the model can be exploited in a single reinforcement learning framework.

Proposed studies on learning the curriculum have a potential to result in better understanding of what linguistic and distributional characteristics of training texts are required for effective training of data-driven models. They can help transform and better leverage existing training datasets, to shed light to missing knowledge in the datasets, as well as to outliers and irrelevant training examples that can be filtered at training time. They are potentially useful also in lower-resource and non-English settings (i) through the use of unsupervised features, and (ii) to guide annotation for creating new datasets by identifying more urgent examples for annotation, in settings with large amounts of unlabeled data and a limited availability of annotators (e.g., Garrette and Baldridge, 2013). With awareness to content and structure of training data, this enterprise has the potential to improve NLP broadly, across a wide range of tasks, and give a better starting point to future annotations. In addition, the proposed framework is setting a platform for investigating a range of related problems: data selection, domain adaptation, and active learning. Beyond NLP, this framework can be used in psycholinguistic studies, in studies of language acquisition, and second language learning.

7.2.4 Future work on polyglot models

Linguistic generalizations learned by monolingual data-driven models typically require from thousands to millions of examples to learn, and are not directly transferable to out-of-domain data (i.e., to another text domain, style, genre, register, to code-switching, and to other languages). The key idea of the polyglot models is to overcome this data- and generalization-bottleneck by building a single model for resource-rich and resource-poor languages that leverages commonalities across languages. Polyglot approach thus allows to learn more robust, language-universal generalizations, and facilitates data transfer. In Chapter 5, I presented a multilingual architecture to effectively transfer phonological representations; Ammar *et al.* (2016a) also showed that languages with a small treebank benefit greatly from multilingual parsing.

The goal of the future research on polyglot models is to show how to integrate prior knowledge in diverse multilingual and multi-domain neural models to effectively transfer resources via linguistically-informed generalizations, and to enable lower-resource neural NLP. I propose to apply the §5.2 most immediately, to multilingual representation learning, adapting the proposed model from phonological to word representations. Then, armed with language-universal representations of lexical items, to proceed to additional multilingual NLP tasks that will leverage these representations, starting with the model of Ammar *et al.* (2016a) and integrating in it typological morpho-syntactic features using the proposed multiplicative architecture, and then addressing additional multilingual tasks, e.g., morphological generation (cf. Faruqui *et al.*, 2016; Kann and Schütze, 2016) and machine translation (cf. Firat *et al.*, 2016). In each of the tasks, it is essential to identify (i) architectural modifications, and (ii) the set of necessary and sufficient linguistic features that are required to mediate between hybrid inputs across tasks, to enable multi-dialect, multi-domain, code-switched, and multimodal neural models. Across the tasks, it would be interesting to conduct an in-depth exploration of data reduction, e.g., in one of the languages, and adaptation of the trained polyglot model to unseen languages. It would also be interesting to go beyond high-level typological features, and to explore how to integrate auxiliary linguistic knowledge, e.g., to integrate phrase-based translation tables into neural MT.

Extensive research on polyglot models has the potential to improve resource-constrained neural NLP applications, to bridge between symbolic and neural approaches, and to shed better light (i) to our understanding of what linguistic knowledge need to be integrated directly and what can be learned automatically, and (ii) to our understanding of generalization capacity of current techniques.

7.2.5 Future work on evaluation and interpretation of distributed representations

Aligning dimensions of linguistic and distributional vectors enables projection of linguistic annotations via the alignments, and thereby facilitates qualitative analysis of individual dimensions in distributional vectors. We find correspondence between the projected labels of distributional columns and the column content. For example, in the 50-dimensional SG model top-10 ranked words in a dimension aligned to NOUN.BODY with r=0.26 are *amputated, sprained, palsy, semenya, lacerations, genital, cervical, concussion, congenital, abdominal*. This interesting by-product of our method will be addressed in future work.

While we experiment with linguistic vectors capturing semantic/syntactic concepts, our methodology is generally applicable to other linguistic resources and more fine-grained linguistic features. Thus, QVEC and QVEC-CCA can be used as a task-specific evaluators, relying on semantic, syntactic, morphological, and typological resources (e.g., Universal Dependencies Treebank (Agić *et al.*, 2015) and WALS (Dryer and Haspelmath, 2013a)), and as evaluators for non-English word embeddings, when resources are available (e.g., Danish supersenses (Martínez Alonso *et al.*, 2015)).

A useful property of supersenses (features in our linguistic vectors) is that they are stable across languages (Schneider *et al.*, 2013; Tsvetkov *et al.*, 2014a). Cross-lingual vector evaluation and evaluation of multilingual word vectors with QVEC and QVEC-CCA is thus an additional promising research avenue. We have already started the investigation in (Ammar *et al.*, 2016b).

Finally, in our experiments we measured linear correlation of word vectors with a linguistic resource. Adapting the intrinsic evaluation to a non-linear matching to a linguistic resource, e.g. via transfer learning (cf. Qian *et al.*, 2016), or using deep CCA (Andrew *et al.*, 2013) could yield better evaluation results.

Bibliography

Adel, H., Vu, N. T., and Schultz, T. (2013). Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proc. ACL*, pages 206–211.

Adler, A. N. (2006). Faithfulness and perception in loanword adaptation: A case study from Hawaiian. *Lingua*, **116**(7), 1024–1045.

Agić, Ž., Aranzabe, M. J., Atutxa, A., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goenaga, I., Gojenola, K., Goldberg, Y., Hajič, J., Johannsen, A. T., Kanerva, J., Kuokkala, J., Laippala, V., Lenci, A., Lindén, K., Ljubešić, N., Lynn, T., Manning, C., Martínez, H. A., McDonald, R., Missilä, A., Montemagni, S., Nivre, J., Nurmi, H., Osenova, P., Petrov, S., Piitulainen, J., Plank, B., Prokopidis, P., Pyysalo, S., Seeker, W., Seraji, M., Silveira, N., Simi, M., Simov, K., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). Universal dependencies 1.1. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Ahn, S.-C. and Iverson, G. K. (2004). Dimensions in Korean laryngeal phonology. *Journal of East Asian Linguistics*, **13**(4), 345–379.

Al-Onaizan, Y. and Knight, K. (2002). Machine transliteration of names in Arabic text. In *Proc. the ACL workshop on Computational Approaches to Semitic Languages*, pages 1–13.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, pages 11–23. Springer.

Alves, M. (2009). Loanwords in Vietnamese. In M. Haspelmath and U. Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 617–637. Max Planck Institute for Evolutionary Anthropology.

Ammar, W., Dyer, C., and Smith, N. A. (2012). Transliteration by sequence labeling with lattice encodings and reranking. In *Proc. NEWS workshop at ACL*.

Ammar, W., Chahuneau, V., Denkowski, M., Hanneman, G., Ling, W., Matthews, A., Murray, K., Segall, N., Tsvetkov, Y., Lavie, A., and Dyer, C. (2013). The cmu machine translation systems at WMT 2013: Syntax, synthetic translation options, and pseudo-references. In *Proc. WMT*.

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016a). Many languages, one parser. *TACL*.

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016b). Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Andrew, G., Arora, R., Bilmes, J. A., and Livescu, K. (2013). Deep canonical correlation analysis. In *Proc. ICML*, pages 1247–1255.

Androutsopoulos, J. K. (2000). Non-standard spellings in media texts: The case of German fanzines. *Journal of Sociolinguistics*, **4**(4), 514–533.

Anis, J. (2007). Neography: Unconventional spelling in French SMS text messages. *The multilingual internet: Language, culture, and communication online*, pages 87–115.

Appel, R. and Muysken, P. (2005). *Language contact and bilingualism*. Amsterdam University Press.

Aue, A. and Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proc. RANLP*.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proc. EMNLP*, pages 355–362.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*.

Bansal, M., Gimpel, K., and Livescu, K. (2014). Tailoring continuous word representations for dependency parsing. In *Proc. ACL*.

Bar-Hillel, Y. (1960). A demonstration of the nonfeasibility of fully automatic high quality machine translation. *Advances in Computers*.

Barker, M. E. (1969). The phonological adaptation of French loanwords io Vietnamese. *Mon-Khmer Studies Journal*, **3**, 138–147.

Baroni, M., Bernardi, R., Do, N.-Q., and Shan, C.-c. (2012). Entailment above the word level in distributional semantics. In *Proc. EACL*, pages 23–32.

Bates, E. and MacWhinney, B. (1989). Functionalism and the competition model. *The crosslinguistic study of sentence processing*, **3**, 73–112.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, **5**(2), 157–166.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *JMLR*, **3**, 1137–1155.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proc. ICML*, pages 41–48.

Benson, M. (1959). English loanwords in Russian. *The Slavic and East European Journal*, **3**(3), 248–267.

Berg-Kirkpatrick, T. and Klein, D. (2010). Phylogenetic grammar induction. In *Proc. ACL*, pages 1288–1297.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Proc. NIPS*, pages 2546–2554.

Biemann, C., Heyer, G., Quasthoff, U., and Richter, M. (2007). The leipzig corpora collection-monolingual corpora of standard size. In *Proc. Corpus Linguistic*.

Birke, J. and Sarkar, A. (2007). Active learning for the identification of nonliteral language. In *Proc. the Workshop on Computational Approaches to Figurative Language*, FigLanguages '07, pages 21–28.

Black, A. W. (2006). CLUSTERGEN: a statistical parametric synthesizer using trajectory modeling. In *Proc. Interspeech*.

Black, A. W. and Lenzo, K. A. (2003). Building synthetic voices. `http://festvox.org/bsv/`.

Black, A. W. and Taylor, P. (1997). The Festival speech synthesis system: system documentation. Technical report, Human Communication Research Centre, University of Edinburgh.

Blair, A. D. and Ingram, J. (2003). Learning to predict the phonological structure of English loanwords in Japanese. *Applied Intelligence*, **19**(1-2), 101–108.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, Bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proc. ACL*, pages 187–205.

Boersma, P. and Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic inquiry*, **32**(1), 45–86.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proce. SIGMOD*, pages 1247–1250.

Bordes, A., Chopra, S., and Weston, J. (2014). Question answering with subgraph embeddings. In *Proc. EMNLP*.

Bouchard-Côté, A., Hall, D., Griffiths, T. L., and Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, **110**(11), 4224–4229.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32.

Broadwell, G. A., Boz, U., Cases, I., Strzalkowski, T., Feldman, L., Taylor, S., Shaikh, S., Liu, T., Cho, K., and Webb, N. (2013). Using imageability and topic chaining to locate metaphors in linguistic corpora. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 102–110. Springer.

Broselow, E. (2004). Language contact phonology: richness of the stimulus, poverty of the base. In *Proc. NELS*, volume 34, pages 1–22.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, **16**(2), 79–85.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, **18**(4), 467–479.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proc. ACL*.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, **46**(3), 904–911.

Burkett, D. and Klein, D. (2008). Two languages are better than one (for syntactic parsing). In *Proc. EMNLP*, pages 877–886.

Calabrese, A. and Wetzels, W. L. (2009). *Loan phonology*, volume 307. John Benjamins Publishing.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proc. ACL*.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web inventory of transcribed and translated talks. In *Proc. EAMT*, pages 261–268.

Chahuneau, V., Schlinger, E., Smith, N. A., and Dyer, C. (2013). Translating into morphologically rich languages with synthetic phrases. In *Proc. EMNLP*, pages 1677–1687.

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proc. ACL*, pages 310–318.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proc. EMNLP*.

Christodoulopoulos, C., Goldwater, S., and Steedman, M. (2010). Two decades of unsupervised POS induction: How far have we come? In *Proc. EMNLP*, pages 575–584.

Church, K. (2007). A pendulum swung too far. *LiLT*, **2**(4), 1–26.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1), 22–29.

Ciaramita, M. and Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. EMNLP*, pages 594–602.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. ACL*, pages 176–181.

Cohen, E.-G. (2009). *The role of similarity in phonology: Evidence from loanword adaptation in Hebrew*. Ph.D. thesis, Tel Aviv University.

Cohen, E.-G. (2013). The emergence of the unmarked: Vowel harmony in Hebrew loanword adaptation. *Lingua*, **131**, 66–79.

Cohen, S. B., Das, D., and Smith, N. A. (2011). Unsupervised structure prediction with non-parallel multilingual guidance. In *Proc. EMNLP*.

Cohn, T. and Specia, L. (2013). Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proc. ACL*, pages 32–42.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proc. EMNLP*, pages 1–8.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, **12**, 2493–2537.

Comrie, B. and Spagnol, M. (2015). Maltese loanword typology. Submitted.

Curnow, T. J. (2001). What language features can be 'borrowed'. *Areal diffusion and genetic inheritance: problems in comparative linguistics*, pages 412–436.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., and Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In *Proc. WWW*, pages 699–708.

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. ACL*, pages 600–609.

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loanwords*, volume 26. Multilingual Matters.

Daumé, III, H., Kumar, A., and Saha, A. (2010). Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59.

Daumé III, H. (2009). Non-parametric Bayesian areal linguistics. In *Proc. NAACL*, pages 593–601.

Daumé III, H. and Campbell, L. (2007). A bayesian model for discovering typological implications. In *Proc. ACL*, pages 65–72.

Davidson, L. and Noyer, R. (1997). Loan phonology in Huave: nativization and the ranking of faithfulness constraints. In *Proc. WCCFL*, volume 15, pages 65–79.

De Gispert, A. and Marino, J. B. (2006). Catalan-English statistical machine translation without parallel corpus: bridging through Spanish. In *Proc. LREC*, pages 65–68.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, **41**(6), 391–407.

Dholakia, R. and Sarkar, A. (2014). Pivot-based triangulation for low-resource languages. In *Proc. AMTA*.

Diab, M. and Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proc. ACL*.

Dryer, M. S. and Haspelmath, M., editors (2013a). *WALS Online*. Max Planck Institute for Evolutionary Anthropology. `http://wals.info/`.

Dryer, M. S. and Haspelmath, M. (2013b). Wals online. *Max Planck Institute for Evolutionary Anthropology, Leipzig*.

Durkin, P. (2014). *Borrowed Words: A History of Loanwords in English*. Oxford University Press.

Durrani, N., Sajjad, H., Fraser, A., and Schmid, H. (2010). Hindi-to-Urdu machine translation through transliteration. In *Pro. ACL*, pages 465–474.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., and Resnik, P. (2010). `cdec`: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proc. ACL*.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM Model 2. In *Proc. NAACL*.

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., and Smith, N. A. (2015). Transition-based dependency parsing with stack long short-term memory. In *Proc. ACL*.

Eetemadi, S., Lewis, W., Toutanova, K., and Radha, H. (2015). Survey of data-selection methods in statistical machine translation. *Machine Translation*, **29**(3-4), 189–223.

Eisenstein, J. (2013a). Phonological factors in social media writing. In *Proc. NAACL*, pages 11–19.

Eisenstein, J. (2013b). What to do about bad language on the internet. In *HLT-NAACL*, pages 359–369.

Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS ONE*, **9**(11), e113114.

Eisner, J. (1997). Efficient generation in primitive Optimality Theory. In *Proc. EACL*, pages 313–320.

Eisner, J. (2002). Comprehension and compilation in Optimality Theory. In *Proc. ACL*, pages 56–63.

Ellison, T. M. (1994). Phonological derivation in Optimality Theory. In *Proc. CICLing*, pages 1007–1013.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, **48**(1), 71–99.

Erjavec, T. (2004). Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc. LREC*.

Erkan, G. and Radev, D. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *JAIR*, **22**, 457–479.

Fabri, R., Gasser, M., Habash, N., Kiraz, G., and Wintner, S. (2014). Linguistic introduction: The orthography, morphology and syntax of Semitic languages. In *Natural Language Processing of Semitic Languages*, pages 3–41. Springer.

Faruqui, M. and Dyer, C. (2014a). Community evaluation and exchange of word vectors at wordvectors.org. In *Proc. ACL (Demonstrations)*.

Faruqui, M. and Dyer, C. (2014b). Improving vector space word representations using multilingual correlation. In *Proc. EACL*.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proc. NAACL*.

Faruqui, M., Tsvetkov, Y., Neubig, G., and Dyer, C. (2016). Morphological inflection generation using character sequence to sequence learning. In *Proc. NAACL*.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: the concept revisited. In *Proc. WWW*.

Firat, O., Cho, K., and Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proc. NAACL*.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proc. EMNLP*, pages 451–459.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, **1**(3), 215–239.

Friesner, M. L. (2009). The adaptation of romanian loanwords from Turkish and French. *Calabrese & Wetzels (2009)*, pages 115–130.

Fügen, C., Stuker, S., Soltau, H., Metze, F., and Schultz, T. (2003). Efficient handling of multilingual language models. In *Proc. ASRU*, pages 441–446.

Ganchev, K., Gillenwater, J., and Taskar, B. (2009). Dependency grammar induction via bitext projection constraints. In *Proc. ACL*, pages 369–377.

Gandy, L., Allan, N., Atallah, M., Frieder, O., Howard, N., Kanareykin, S., Koppel, M., Last, M., Neuman, Y., and Argamon, S. (2013). Automatic identification of conceptual metaphors with limited knowledge. In *Proc. the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 328–334.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). Ppdb: The paraphrase database. In *Proc. HLT-NAACL*, pages 758–764.

Garley, M. and Hockenmaier, J. (2012). Beefmoves: dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In *Proc. ACL*, pages 135–139.

Garrette, D. and Baldridge, J. (2013). Learning a part-of-speech tagger from two hours of annotation. In *Proc. NAACL*, pages 138–147.

Gedigian, M., Bryant, J., Narayanan, S., and Ciric, B. (2006). Catching metaphors. In *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proc, ACL*, pages 42–47.

Gimpel, K., Batra, D., Dyer, C., and Shakhnarovich, G. (2013). A systematic exploration of diversity in machine translation. In *Proc. EMNLP*, pages 1100–1111.

Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proc. the Stockholm workshop on variation within Optimality Theory*, pages 111–120.

Grant, A. (2009). Loanwords in British English. In M. Haspelmath and U. Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 360–383. Max Planck Institute for Evolutionary Anthropology.

Graves, A. (2013). Generating sequences with recurrent neural networks. *CoRR*, **abs/1308.0850**.

Gross, D. and Miller, K. J. (1990). Adjectives in WordNet. *International Journal of Lexicography*, **3**(4), 265–277.

Guo, J., Che, W., Wang, H., and Liu, T. (2014). Revisiting embedding features for simple semi-supervised learning. In *Proc. EMNLP*.

Guy, G. R. (1990). The sociolinguistic types of language change. *Diachronica*, **7**(1), 47–67.

Habash, N. (2008). Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In *Proc. ACL*, pages 57–60.

Habash, N. and Hu, J. (2009). Improving Arabic-Chinese statistical machine translation using English as pivot language. In *Proc. WMT*, pages 173–181.

Habash, N., Rambow, O., and Roth, R. (2009). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proc. MEDAR*, pages 102–109.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proc. ACL*, pages 771–779.

Hajič, J., Hric, J., and Kuboň, V. (2000). Machine translation of very close languages. In *Proc. ANLP*, pages 7–12.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., *et al.* (2009). The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proc. CoNLL: Shared Task*, pages 1–18.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proc. EMNLP*, pages 363–371.

Hamp, B. and Feldweg, H. (1997). Germanet-a lexical-semantic net for German. In *Proc. ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a# twitter. In *Proc. ACL*, pages 368–378.

Harbeck, S., Nöth, E., and Niemann, H. (1997). Multilingual speech recognition. In *Proc. 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, pages 9–15.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, **16**(12), 2639–2664.

Haspelmath, M. (2009). Lexical borrowing: concepts and issues. *Loanwords in the World's Languages: a comparative handbook*, pages 35–54.

Haspelmath, M. and Tadmor, U., editors (2009). *Loanwords in the World's Languages: A Comparative Handbook*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, pages 210–231.

Hayes, B., Tesar, B., and Zuraw, K. (2013). OTSoft 2.3.2.

He, J., Chen, J., He, X., Gao, J., Li, L., Deng, L., and Ostendorf, M. (2016). Deep reinforcement learning with an action space defined by natural language. In *Proc. ACL*.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proc. WMT*.

Heilman, M. J., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proc. NAACL*, pages 460–467.

Hennig, P. and Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *The Journal of Machine Learning Research*, **13**(1), 1809–1837.

Hermann, K. M. and Blunsom, P. (2014). Multilingual Models for Compositional Distributional Semantics. In *Proc. ACL*.

Hermjakob, U., Knight, K., and Daumé III, H. (2008). Name translation in statistical machine translation-learning when to transliterate. In *Proc. ACL*, pages 389–397.

Herring, S. C. (2012). Grammar and electronic communication. *The Encyclopedia of Applied Linguistics*.

Hill, F., Reichart, R., and Korhonen, A. (2014). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, **abs/1408.3456**.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**(8), 1735–1780.

Hock, H. H. and Joseph, B. D. (2009). *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*, volume 218. Walter de Gruyter.

Hoffman, M. D., Brochu, E., and de Freitas, N. (2011). Portfolio allocation for Bayesian optimization. In *Proc. UAI*, pages 327–336.

Holden, K. (1976). Assimilation rates of borrowings and phonological productivity. *Language*, pages 131–147.

Hovy, D., Srivastava, S., Jauhar, S. K., Sachan, M., Goyal, K., Li, H., Sanders, W., and Hovy, E. (2013). Identifying metaphorical word use with tree kernels. In *Proc. the First Workshop on Metaphor in NLP*, page 52.

Hu, Y. and Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *Audio, Speech, & Language Processing*, **16**(1), 229–238.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proc. ACL*, pages 873–882.

Huang, K., Gardner, M., Papalexakis, E., Faloutsos, C., Sidiropoulos, N., Mitchell, T., Talukdar, P. P., and Fu, X. (2015). Translation invariant word embeddings. In *Proc. EMNLP*, pages 1084–1088.

Hurskainen, A. (2004a). HCS 2004–Helsinki corpus of Swahili. Technical report, Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC.

Hurskainen, A. (2004b). Loan words in Swahili. In K. Bromber and B. Smieja, editors, *Globalisation and African Languages*, pages 199–218. Walter de Gruyter.

Hutter, M. (2009). Feature reinforcement learning: Part i. unstructured mdps. *JAGI*, **1**(1), 3–24.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, **11**(3).

Jacobs, H. and Gussenhoven, C. (2000). Loan phonology: perception, salience, the lexicon and OT. *Optimality Theory: Phonology, syntax, and acquisition*, pages 193–209.

Jang, H., Jo, Y., Shen, Q., Moon, M. M. S., and Rosé, C. P. (2016). Metaphor detection with topic transition, emotion and cognition in context. In *Proc. ACL*.

Jiang, L., Meng, D., Yu, S.-I., Lan, Z., Shan, S., and Hauptmann, A. (2014). Self-paced learning with diversity. In *Proc. NIPS*, pages 2078–2086.

Jiang, L., Meng, D., Zhao, Q., Shan, S., and Hauptmann, A. G. (2015). Self-paced curriculum learning. In *Proc. AAAI*, volume 2, page 6.

Johnson, F. (1939). *Standard Swahili-English dictionary*. Oxford University Press.

Johnson, M. (2014). Trends of loanword origins in 20th century Finnish.

Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, **21**(4), 345–383.

Kager, R. (1999). *Optimality Theory*. Cambridge University Press.

Kail, R. (1990). *The development of memory in children*. W. H. Freeman and Company, 3rd edition.

Kallio, P. (2006). On the earliest Slavic loanwords in Finnic. *Slavica Helsingiensia*, **27**, 154–166.

Kang, Y. (2003). Perceptual similarity in loanword adaptation: English postvocalic word-final stops in Korean. *Phonology*, **20**(2), 219–274.

Kang, Y. (2011). Loanword phonology. In M. van Oostendorp, C. Ewen, E. Hume, and K. Rice, editors, *Companion to Phonology*. Wiley–Blackwell.

Kann, K. and Schütze, H. (2016). Morphological reinflection with encoder-decoder models and edit tree. In *Proc. ACL*.

Kawahara, S. (2008). Phonetic naturalness and unnaturalness in Japanese loanword phonology. *Journal of East Asian Linguistics*, **17**(4), 317–330.

Kay, G. (1995). English loanwords in Japanese. *World Englishes*, **14**(1), 67–76.

Kemeny, J. G. (1959). Mathematics without numbers. *j-DAEDALUS*, **88**(4), 577–591.

Kenstowicz, M. (2005). The phonetics and phonology of Korean loanword adaptation. In *Proceedings of the first European conference on Korean linguistics*, volume 1, pages 17–32.

Kenstowicz, M. (2006). Tone loans: The adaptation of English loanwords into Yoruba. In *Selected proceedings of the 35th annual conference on African Linguistics*, pages 136–146.

Kenstowicz, M. (2007). Salience and similarity in loanword adaptation: a case study from Fijian. *Language Sciences*, **29**(2), 316–340.

Kenstowicz, M. and Suchato, A. (2006). Issues in loanword adaptation: A case study from Thai. *Lingua*, **116**(7), 921–949.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, **abs/1412.6980**.

Kiros, R. and Salakhutdinov, R. (2013). Multimodal neural language models. In *Proc. NIPS Deep Learning Workshop*.

Kiros, R., Salakhutdinov, R., and Zemel, R. (2015). Unifying visual-semantic embeddings with multimodal neural language models. *TACL*.

Klakow, D. and Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication*, **38**(1), 19–28.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proc. ICASSP*, volume 1, pages 181–184.

Knight, K. and Graehl, J. (1998). Machine transliteration. *Computational Linguistics*, **24**(4), 599–612.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proc. NAACL-HLT*, pages 48–54.

Kominek, J., Schultz, T., and Black, A. W. (2008). Synthesizer voice quality of new languages calibrated with mean Mel Cepstral Distortion. In *Proc. SLTU*, pages 63–68.

Kondrak, G. (2001). Identifying cognates by phonetic and semantic similarity. In *Proc. NAACL*, pages 1–8. Association for Computational Linguistics.

Kondrak, G. and Sherif, T. (2006). Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proc. the Workshop on Linguistic Distances*, pages 43–50.

Kondrak, G., Marcu, D., and Knight, K. (2003). Cognates can improve statistical translation models. In *Proc. HLT-NAACL*, pages 46–48.

Kozhenikov, M. and Titov, I. (2013). Cross-lingual transfer of semantic role labeling models. In *Proc. ACL*, pages 1190–1200.

Kozhevnikov, M. and Titov, I. (2013). Cross-lingual transfer of semantic role labeling models. In *Proc. ACL*, pages 1190–1200.

Krishnakumaran, S. and Zhu, X. (2007). Hunting elusive metaphors using lexical resources. In *Proc. the Workshop on Computational approaches to Figurative Language*, pages 13–20.

Kuhn, J. (2004). Experiments in parallel-text based grammar induction. In *Proc. ACL*, page 470.

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proc. WWW*, pages 625–635.

Kumar, M. P., Packer, B., and Koller, D. (2010). Self-paced learning for latent variable models. In *Proc. NIPS*, pages 1189–1197.

Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, **44**(4), 978–990.

Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, **86**(1), 97–106.

Kwiatkowski, T., Goldwater, S., Zettlemoyer, L., and Steedman, M. (2012). A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proc. EACL*, pages 234–244.

Labov, W. (1966). *The social stratification of English in New York city*. Cambridge University Press.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, pages 282–289.

Lakoff, G. and Johnson, M. (1980). Conceptual metaphor in everyday language. *The Journal of Philosophy*, pages 453–486.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proc. NAACL*.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*.

Lazaridou, A., Vecchi, E. M., and Baroni, M. (2013). Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proc. EMNLP*.

Lee, Y. J. and Grauman, K. (2011). Learning the easy things first: Self-paced visual category discovery. In *Proc. CVPR*, pages 1721–1728.

Lembersky, G., Ordan, N., and Wintner, S. (2012). Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, **38**(4), 799–825.

Lembersky, G., Ordan, N., and Wintner, S. (2013). Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, **39**(4), 999–1023.

Levin, L., Mitamura, T., Fromm, D., MacWhinney, B., Carbonell, J., Feely, W., Frederking, R., Gershman, A., and Ramirez, C. (2014). Resources for the detection of conventionalized metaphors in four languages. In *Proc. LREC*.

Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proc. ACL*, pages 302–308.

Lewis, M. P., Simons, G. F., and Fennig, C. D. (2015). *Ethnologue: Languages of the world*. Texas: SIL International. `http://www.ethnologue.com`.

Li, J., Monroe, W., Ritter, A., and Jurafsky, D. (2016). Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Li, S., Graça, J. V., and Taskar, B. (2012). Wiki-ly supervised part-of-speech tagging. In *Proc. EMNLP*, pages 1389–1398.

Ling, W., Luís, T., Marujo, L., Astudillo, R. F., Amir, S., Dyer, C., Black, A. W., and Trancoso, I. (2015a). Finding function in form: Compositional character models for open vocabulary word representation. In *Proc. NAACL*.

Ling, W., Chu-Cheng, L., Tsvetkov, Y., Amir, S., Fermandez, R., Dyer, C., Black, A. W., and Trancoso, I. (2015b). Not all contexts are created equal: Better word representations with variable attention. In *Proc. EMNLP*.

Ling, W., Dyer, C., Black, A., and Trancoso, I. (2015c). Two/too simple adaptations of `word2vec` for syntax problems. In *Proc. NAACL*.

List, J.-M. and Moran, S. (2013). An open source toolkit for quantitative historical linguistics. In *Proc. ACL (System Demonstrations)*, pages 13–18.

Littell, P., Price, K., and Levin, L. (2014). Morphological parsing of Swahili using crowdsourced lexical resources. In *Proc. LREC*.

Littell, P., Mortensen, D., Goyal, K., Dyer, C., and Levin, L. (2016). Bridge-language capitalization inference in Western Iranian: Sorani, Kurmanji, Zazaki, and Tajik. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'16)*.

Lu, A., Wang, W., Bansal, M., Gimpel, K., and Livescu, K. (2015). Deep multilingual correlation for improved word embeddings. In *Proc. NAACL*.

Lü, Y., Huang, J., and Liu, Q. (2007). Improving statistical machine translation performance by training data selection and optimization. In *Proc. EMNLP*, pages 343–350.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2016). Multi-task sequence to sequence learning. In *Proc. ICLR*.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., and Kulick, S. (2010). LDC Standard Arabic morphological analyzer (SAMA) v. 3.1.

MacWhinney, B. (1987). The competition model. *Mechanisms of language acquisition*, pages 249–308.

Magurran, A. E. (2013). *Measuring biological diversity*. John Wiley & Sons.

Mann, G. S. and Yarowsky, D. (2001). Multipath translation lexicon induction via bridge languages. In *Proc. HLT-NAACL*, pages 1–8.

Manning, C. D. (2016). Computational linguistics and deep learning. *Computational Linguistics*.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, **19**(2), 313–330.

Martínez Alonso, H., Johannsen, A., Olsen, S., Nimb, S., Sørensen, N. H., Braasch, A., Søgaard, A., and Pedersen, B. S. (2015). Supersense tagging for Danish. In *Proc. NODALIDA*, page 21.

Martins, A. F. T., Smith, N. A., Xing, E. P., Aguiar, P. M. Q., and Figueiredo, M. A. T. (2010). Turbo parsers: dependency parsing by approximate variational inference. In *Proc. ENMLP*, pages 34–44.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. In *Proc. EMNLP*, pages 381–390.

Mason, Z. J. (2004). CorMet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, **30**(1), 23–44.

McCarthy, J. J. (1985). *Formal problems in Semitic phonology and morphology*. Ph.D. thesis, MIT.

McCarthy, J. J. (2009). *Doing Optimality Theory: Applying theory to data*. John Wiley & Sons.

McCarthy, J. J. and Prince, A. (1995). Faithfulness and reduplicative identity. *Beckman et al. (Eds.)*, pages 249–384.

McDonald, R., Petrov, S., and Hall, K. (2011a). Multi-source transfer of delexicalized dependency parsers. In *Proc. EMNLP*.

McDonald, R., Petrov, S., and Hall, K. (2011b). Multi-source transfer of delexicalized dependency parsers. In *Proc. EMNLP*, pages 62–72.

McDonald, R. T., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K. B., Petrov, S., Zhang, H., Täckström, O., *et al.* (2013). Universal dependency annotation for multilingual parsing. In *Proc. ACL*, pages 92–97.

McMahon, A. M. S. (1994). *Understanding language change*. Cambridge University Press.

Metze, F., Hsiao, R., Jin, Q., Nallasamy, U., and Schultz, T. (2010). The 2010 CMU GALE speech-to-text system. In *Proc. INTERSPEECH*, pages 1501–1504.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proc. Interspeech*, pages 1045–1048.

Mikolov, T., Kombrink, S., Burget, L., Černockỳ, J. H., and Khudanpur, S. (2011). Extensions of recurrent neural network language model. In *Proc. ICASSP*, pages 5528–5531.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pages 3111–3119.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013b). Efficient estimation of word representations in vector space. In *Proc. ICLR*.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013c). Exploiting similarities among languages for Machine Translation. *CoRR*, **abs/1309.4168**.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Proc. HLT*, pages 303–308.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. In *Proc. ICWSM*, pages 554–557.

Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohammad, T., Nakashole, N., Platanios, E., Ritterk, A., Samadi, M., Settles, B., Wang, R., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., and Welling, J. (2015). Never-ending learning. In *Proc. AAAI*.

Mnih, A. and Hinton, G. (2007). Three new graphical models for statistical language modelling. In *Proc, ICML*, pages 641–648.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., *et al.* (2015). Human-level control through deep reinforcement learning. *Nature*, **518**(7540), 529–533.

Močkus, J., Tiesis, V., and Žilinskas, A. (1978). On Bayesian methods for seeking the extremum. *Towards global optimization*, **2**(117-129), 2.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proc. ACL*, pages 220–224.

Moran, S., McCloy, D., and Wright, R., editors (2014). *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology. `http://phoible.org/`.

Moravcsik, E. (1978). Language contact. *Universals of human language*, **1**, 93–122.

Mwita, L. C. (2009). The adaptation of Swahili loanwords from Arabic: A constraint-based analysis. *Journal of Pan African Studies*.

Myers-Scotton, C. (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press Oxford.

Nakov, P. and Ng, H. T. (2012). Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, pages 179–222.

Narasimhan, K., Kulkarni, T., and Barzilay, R. (2015). Language understanding for text-based games using deep reinforcement learning. In *Proc. EMNLP*.

Naseem, T., Barzilay, R., and Globerson, A. (2012). Selective sharing for multilingual dependency parsing. In *Proc. ACL*, pages 629–637.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, **41**(2), 10:1–10:69.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer journal*, **7**(4), 308–313.

Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Howard, N., and Frieder, O. (2013). Metaphor identification in large texts corpora. *PloS one*, **8**(4), e62343.

Nguyen, D., Smith, N. A., and Rosé, C. P. (2011). Author age prediction from text using linear regression. In *Proc. the ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Niculae, V. and Danescu-Niculescu-Mizil, C. (2014). Brighter than gold: Figurative language in user generated comparisons. In *Proc. EMNLP*.

Niehues, J., Herrmann, T., Vogel, S., and Waibel, A. (2011). Wider context by using bilingual language models in machine translation. In *Proc. WMT*, pages 198–206.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proc. ACL*, pages 160–167.

O'Connor, B., Balasubramanyan, R., Routledge, B., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proc. ICWSM*.

Ojo, V. (1977). *English-Yoruba language contact in Nigeria*. Universität Tübingen.

Orešnik, B. (1982). On the adaptation of English loanwords into Finnish. *The English Element in European Languages*, **2**, 180–212.

Padó, S. and Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, **36**(1), 307–340.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2009). English Gigaword fourth edition.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proc. EMNLP*.

Perlich, C., Provost, F., and Simonoff, J. S. (2003). Tree induction vs. logistic regression: a learning-curve analysis. *Journal of Machine Learning Research*, **4**, 211–255.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proc. LREC*.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proc. EMNLP*, pages 186–195.

Polomé, E. C. (1967). *Swahili Language Handbook*. ERIC.

Prahallad, K., Kumar, E. N., Keri, V., Rajendran, S., and Black, A. W. (2012). The IIIT-H Indic speech databases. In *Proc. Interspeech*.

Preotiuc-Pietro, D. and Cohn, T. (2013). A temporal model of text periodicities using gaussian processes. In *Proc. EMNLP*, pages 977–988.

Prince, A. and Smolensky, P. (2008). *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Qian, P., Qiu, X., and Huang, X. (2016). Investigating language universal and specific in word embedding. In *Proc. ACL*.

Qian, T., Hollingshead, K., Yoon, S.-y., Kim, K.-y., Sproat, R., and LREC, M. (2010). A Python toolkit for universal transliteration. In *Proc. LREC*.

Rabinovich, E. and Wintner, S. (2015). Unsupervised identification of translationese. *TACL*, **3**, 419–432.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2016). Sequence level training with recurrent neural networks. In *Proc. ICLR*.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, **21**(1), 24–43.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Pro. ACL*, pages 320–322.

Rasmussen, C. E. (2006). Gaussian processes for machine learning.

Razmara, M., Siahbani, M., Haffari, R., and Sarkar, A. (2013). Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proc. ACL*, pages 1105–1115.

Rendón, J. G. (2008). Spanish lexical borrowing in Imbabura Quichua: In search of constraints on language contact. *Empirical Approaches to Language Typology*, **39**, 95.

Rendón, J. G. and Adelaar, W. (2009). Loanwords in Imbabura Quichua. In M. Haspelmath and U. Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 944–967. Max Planck Institute for Evolutionary Anthropology.

Repetti, L. (2006). The emergence of marked structures in the integration of loans in Italian. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, **274**, 209.

Rosch, E. (1978). Principles of categorization. In E. Rosch and B. B. Lloyd, editors, *Cognition and categorization*, pages 28–71.

Rose, Y. and Demuth, K. (2006). Vowel epenthesis in loanword adaptation: Representational and phonetic considerations. *Lingua*, **116**(7), 1112–1139.

Rosenzweig, M. L. (1995). *Species diversity in space and time*. Cambridge University Press.

Rothman, N. C. (2002). Indian Ocean trading links: The Swahili experience. *Comparative Civilizations Review*, **46**, 79–90.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proc. EMNLP*.

Saluja, A., Hassan, H., Toutanova, K., and Quirk, C. (2014). Graph-based semi-supervised learning of translation models from monolingual data. In *Proc. ACL*, pages 676–686.

Sankoff, G. (2002). Linguistic outcomes of language contact. In J. Chambers, P. Trudgill, and N. Schilling-Estes, editors, *Handbook of Sociolinguistics*, pages 638–668. Blackwell.

Schadeberg, T. C. (2009). Loanwords in Swahili. In M. Haspelmath and U. Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 76–102. Max Planck Institute for Evolutionary Anthropology.

Schlinger, E., Chahuneau, V., and Dyer, C. (2013). `morphogen`: Translation into morphologically rich languages with synthetic phrases. *The Prague Bulletin of Mathematical Linguistics*, **100**, 51–62.

Schmidt, C. K. (2009). Loanwords in Japanese. In M. Haspelmath and U. Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 545–574. Max Planck Institute for Evolutionary Anthropology.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proc. EMNLP*.

Schneider, N. (2014). *Lexical Semantic Analysis in Natural Language Text*. Ph.D. thesis, Carnegie Mellon University.

Schneider, N., Mohit, B., Dyer, C., Oflazer, K., and Smith, N. A. (2013). Supersense tagging for Arabic: the MT-in-the-middle attack. In *Proc. NAACL-HLT*, pages 661–667.

Schulte, K. (2009). Loanwords in Romanian. In M. Haspelmath and U. Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 230–259. Max Planck Institute for Evolutionary Anthropology.

Schultz, T. and Schlippe, T. (2014). GlobalPhone: Pronunciation dictionaries in 20 languages. In *Proc. LREC*.

Schwarm, S. E. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proc. ACL*, pages 523–530.

Schwarzwald, O. (1998). Word foreignness in modern Hebrew. *Hebrew Studies*, pages 115–142.

Schwenk, H. (2007). Continuous space language models. *Computer Speech & Language*, **21**(3), 492–518.

Sedoc, J., Gallier, J., Ungar, L., and Foster, D. (2016). Semantic word clusters using signed normalized graph cuts. *arXiv preprint arXiv:1601.05403*.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE*, **104**(1), 148–175.

Shutova, E. (2010). Models of metaphor in NLP. In *Proc. ACL*, pages 688–697.

Shutova, E. and Sun, L. (2013). Unsupervised metaphor identification using hierarchical graph factorization clustering. In *Proc. NAACL-HLT*, pages 978–988.

Shutova, E. and Teufel, S. (2010). Metaphor corpus annotated for source-target domain mappings. In *Proc. LREC*, pages 3255–3261.

Shutova, E., Sun, L., and Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In *Proc. COLING*, pages 1002–1010.

Shutova, E., Teufel, S., and Korhonen, A. (2013). Statistical metaphor processing. *Computational Linguistics*, **39**(2), 301–353.

Shutova, E., Kiela, D., and Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. In *Proc. NAACL*.

Simpson, E. H. (1949). Measurement of diversity. *Nature*.

Skinner, B. F. (1938). The behavior of organisms: an experimental analysis. *An Experimental Analysis*.

Smith, D. A. and Smith, N. A. (2004). Bilingual parsing with factored estimation: Using english to parse korean. In *Proc. EMNLP*, pages 49–56.

Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the Common Crawl. In *Proc. ACL*, pages 1374–1383.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Proc. NIPS*, pages 2951–2959.

Snyder, B., Naseem, T., and Barzilay, R. (2009). Unsupervised multilingual grammar induction. In *Proc. ACL/AFNLP*, pages 73–81.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013a). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. EMNLP*.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013b). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. EMNLP*.

Spitkovsky, V. I., Alshawi, H., and Jurafsky, D. (2010). From baby steps to leapfrog: How less is more in unsupervised dependency parsing. In *Proc. NAACL*, pages 751–759.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. ICML*, pages 1015–1022.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., and Krennmayr, T. (2010). Metaphor in usage. *Cognitive Linguistics*, **21**(4), 765–796.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, **4**(15), 707–719.

Strzalkowski, T., Broadwell, G. A., Taylor, S., Feldman, L., Yamrom, B., Shaikh, S., Liu, T., Cho, K., Boz, U., Cases, I., *et al.* (2013). Robust extraction of metaphors from novel data. In *Proc. the First Workshop on Metaphor in NLP*, page 67.

Styblo Jr, M. (2007). *English loanwords in modern Russian language*. Master's thesis, University of North Carolina.

Sundermeyer, M., Schlüter, R., and Ney, H. (2012). LSTM neural networks for language modeling. In *Proc. Interspeech*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proc. NIPS*, pages 3104–3112.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press Cambridge.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *Proc. ICLR*.

Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *TACL*, **1**, 1–12.

Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *TACL*, **1**, 1–12.

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. In M. Haspelmath and U. Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 55–75. Max Planck Institute for Evolutionary Anthropology.

Tang, E. K., Suganthan, P. N., and Yao, X. (2006). An analysis of diversity measures. *Machine Learning*, **65**(1), 247–271.

Thibodeau, P. H. and Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PLoS One*, **6**(2), e16782.

Thomason, S. G. and Kaufman, T. (2001). *Language contact*. Edinburgh University Press Edinburgh.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**(3/4), 285–294.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proc. LREC*, pages 2214–2218.

Tiedemann, J. (2014). Rediscovering annotation projection for cross-lingual parser induction. In *Proc. COLING*.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proc. CoNLL*.

Tsvetkov, Y. and Dyer, C. (2015). Lexicon stratification for translating out-of-vocabulary words. In *Proc. ACL*, pages 125–131.

Tsvetkov, Y. and Dyer, C. (2016). Cross-lingual bridges with models of lexical borrowing. *JAIR*, **55**, 63–93.

Tsvetkov, Y., Mukomel, E., and Gershman, A. (2013a). Cross-lingual metaphor detection using common semantic features. In *The 1st Workshop on Metaphor in NLP 2013*, page 45.

Tsvetkov, Y., Dyer, C., Levin, L., and Bhatia, A. (2013b). Generating English determiners in phrase-based translation with synthetic translation options. In *Proc. WMT*.

Tsvetkov, Y., Schneider, N., Hovy, D., Bhatia, A., Faruqui, M., and Dyer, C. (2014a). Augmenting English adjective senses with supersenses. In *Proc. LREC*, pages 4359–4365.

Tsvetkov, Y., Metze, F., and Dyer, C. (2014b). Augmenting translation models with simulated acoustic confusions for improved spoken language translation. In *Proc. EACL*, pages 616–625.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014c). Metaphor detection with cross-lingual model transfer. In *Proc. ACL*, pages 248–258.

Tsvetkov, Y., Ammar, W., and Dyer, C. (2015a). Constraint-based models of lexical borrowing. In *Proc. NAACL*, pages 598–608.

Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., and Dyer, C. (2015b). Evaluation of word vector representations by subspace alignment. In *Proc. EMNLP*, pages 2049–2054. `https://github.com/ytsvetko/qvec`.

Tsvetkov, Y., Faruqui, M., and Dyer, C. (2016a). Correlation-based intrinsic evaluation of word vector representations. In *Poc. RepEval*.

Tsvetkov, Y., Faruqui, M., Ling, W., and Dyer, C. (2016b). Learning the curriculum with Bayesian optimization for task-specific word representation learning. In *Proc. ACL*.

Tsvetkov, Y., Sitaram, S., Faruqui, M., Lample, G., Littell, P., Mortensen, D., Black, A. W., Levin, L., and Dyer, C. (2016c). Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proc. NAACL*.

Turian, J., Ratinov, L., Bengio, Y., and Roth, D. (2009). A preliminary evaluation of word representations for named-entity recognition. In *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*, pages 1–8.

Turian, J., Ratinov, L., and Bengio, Y. (2010a). Word representations: a simple and general method for semi-supervised learning. In *Proc. ACL*, pages 384–394.

Turian, J., Ratinov, L., and Bengio, Y. (2010b). Word representations: a simple and general method for semi-supervised learning. In *Proc. ACL*.

Turney, P. D., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proc. EMNL*, pages 680–690.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proc. BEA*, pages 163–173.

Van Coetsem, F. (1988). *Loan phonology and the two transfer types in language contact*. Walter de Gruyter.

Veale, T., Shutova, E., and Klebanov, B. B. (2016). *Metaphor: A Computational Perspective*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Wang, P., Nakov, P., and Ng, H. T. (2012). Source language adaptation for resource-poor machine translation. In *Proc. EMNLP*, pages 286–296.

Wang, X., Liu, Y., Sun, C., Wang, B., and Wang, X. (2015). Predicting polarities of tweets by composing word embeddings with long short-term memory. In *Proc. ACL*, pages 1343–1353.

Wang, Z., Topkara, U., Schultz, T., and Waibel, A. (2002). Towards universal speech recognition. In *Proc. ICMI*, page 247.

Wang, Z., Zoghi, M., Hutter, F., Matheson, D., Freitas, N., *et al.* (2013). Bayesian optimization in high dimensions via random embeddings. In *Proc. IJCAI*.

Wang, Z., Hutter, F., Zoghi, M., Matheson, D., and de Feitas, N. (2016). Bayesian optimization in a billion dimensions via random embeddings. *JAIR*, **55**, 361–387.

Ward, T., Roukos, S., Neti, C., Gros, J., Epstein, M., and Dharanipragada, S. (1998). Towards speech understanding across multiple languages. In *Proc. ICSLP*.

Watanabe, T. and Sumita, E. (2015). Transition-based neural constituent parsing. In *Proc. ACL*.

Watts, O., Wu, Z., and King, S. (2015). Sentence-level control vectors for deep neural network speech synthesis. In *Proc. Interspeech*.

Weinreich, U. (1979). *Languages in contact: Findings and problems*. Walter de Gruyter.

Weng, F., Bratt, H., Neumeyer, L., and Stolcke, A. (1997). A study of multilingual speech recognition. In *Proc. EUROSPEECH*, pages 359–362.

Whitney, W. D. (1881). On mixture in language. *Transactions of the American Philological Association (1870)*, pages 5–26.

Wilks, Y. (1978). Making preferences more active. *Artificial Intelligence*, **11**(3), 197–223.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, **20**(1), 6–10.

Wintner, S. (2009). What science underlies natural language engineering? *Computational Linguistics*, **35**(4), 641–644.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, **23**(3), 377–403.

Xi, C. and Hwa, R. (2005). A backoff model for bootstrapping resources for non-English languages. In *Proc. EMNLP*, pages 851–858.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proc. HLT*, pages 1–8.

Yip, M. (1993). Cantonese loanword phonology and Optimality Theory. *Journal of East Asian Linguistics*, **2**(3), 261–291.

Yogatama, D. and Smith, N. A. (2014). Linguistic structured sparsity in text categorization. In *Proc. ACL*.

Yogatama, D., Kong, L., and Smith, N. A. (2015). Bayesian optimization of text representations. In *Proc. EMNLP*, pages 2100–2105.

Zawawi, S. (1979). *Loan words and their effect on the classification of Swahili nominals*. Leiden: E.J. Brill.

Zhang, Y. and Clark, S. (2011). Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, **37**(1), 105–151.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*.