Don't Forget the Long Tail! A Comprehensive Analysis of Morphological Generalization in Bilingual Lexicon Induction

Abstract

Human translators routinely have to translate rare inflections of words - due to the Zipfian distribution of words in a language. When translating from Spanish, a good translator would have no problem identifying the proper translation of a statistically rare inflection such as hablaramos. Note the lexeme itself, hablar, is relatively common. In this work, we investigate whether state-of-the-art bilingual lexicon inducers are capable of learning this kind of generalization. We introduce 40 morphologically complete dictionaries in 10 languages and evaluate three of the state-of-the-art models on the task of translation of less frequent morphological forms. We demonstrate that the performance of state-of-the-art models drops considerably when evaluated on infrequent morphological inflections and then show that adding a simple morphological constraint at training time improves the performance, proving that the bilingual lexicon inducers can benefit from better encoding of morphology.

Introduction

Human translators exhibit remarkable generalization capabilities and are able to translate even rare inflections they may have never seen before. Indeed, this skill is necessary for translation since language follows a Zipfian distribution BIBREF0: a large number of the tokens in a translated text will come from rare types, including rare inflections of common lexemes. For instance, a Spanish translator will most certainly know the verb hablar “to speak”, but they will only have seen the less frequent, first-person plural future form hablarámos a few times. Nevertheless, they would have no problem translating the latter. In this paper we ask whether current methods for bilingual lexicon induction (BLI) generalize morphologically as humans do. Generalization to rare and novel words is arguably the main point of BLI as a task—most frequent translation pairs are already contained in digital dictionaries. Modern word embeddings encode character-level knowledge BIBREF1, which should—in principle—enable the models to learn this behaviour; but morphological generalization has never been directly tested.

Most existing dictionaries used for BLI evaluation do not account for the full spectrum of linguistic properties of language. Specifically, as we demonstrate in sec:dictionaries, they omit most morphological inflections of even common lexemes. To enable a more thorough evaluation we introduce a new resource: 40 morphologically complete dictionaries for 5 Slavic and 5 Romance languages, which contain the inflectional paradigm of every word they hold. Much like with a human translator, we expect a BLI model to competently translate full paradigms of lexical items. Throughout this work we place our focus on genetically-related language pairs. This not only allows us to cleanly map one morphological inflection onto another, but also provides an upper bound for the performance on the generalization task; if the models are not able to generalize for closely related languages they would most certainly be unable to generalize when translating between unrelated languages.

We use our dictionaries to train and evaluate three of the best performing BLI models BIBREF3, BIBREF4, BIBREF5 on all 40 language pairs. To paint a complete picture of the models' generalization ability we propose a new experimental paradigm in which we independently control for four different variables: the word form's frequency, morphology, the lexeme frequency and the lexeme (a total of 480 experiments). Our comprehensive analysis reveals that BLI models can generalize for frequent morphosyntactic categories, even of infrequent lexemes, but fail to generalize for the more rare categories. This yields a more nuanced picture of the known deficiency of word embeddings to underperform on infrequent words BIBREF6. Our findings also contradict the strong empirical claims made elsewhere in the literature BIBREF4, BIBREF2, BIBREF5, BIBREF7, as we observe that performance severely degrades when the evaluation includes rare morphological variants of a word and infrequent lexemes. We picture this general trend in Figure FIGREF2, which also highlights the skew of existing dictionaries towards more frequent words. As our final contribution, we demonstrate that better encoding of morphology is indeed beneficial: enforcing a simple morphological constraint yields consistent performance improvements for all Romance language pairs and many of the Slavic language pairs.z

Morphological Dictionaries ::: Existing Dictionaries

Frequent word forms can often be found in human-curated dictionaries. Thus, the practical purpose of training a BLI model should be to create translations of new and less common forms, not present in the existing resources. In spite of this, most ground truth lexica used for BLI evaluation contain mainly frequent word forms. Many available resources are restricted to the top 200k most frequent words; this applies to the English–Italian dictionary of BIBREF8, the English–German and English–Finnish dictionaries of BIBREF4, and BIBREF9's English–Spanish resource. The dictionaries of BIBREF10 contain only the top most frequent 10k words for each language. BIBREF11 extracted their Spanish–English and Italian–English lexica from Open Multilingual WordNet BIBREF12, a resource which only yields high frequency, lemma level mappings. Another example is the recent MUSE dataset BIBREF2, which was generated using an “internal translation tool”, and in which the majority of word pairs consist of forms ranked in the top 10k of the vocabularies of their respective languages.

Another problem associated with existing resources is `semantic leakage' between train and evaluation sets. As we demonstrate in §SECREF14, it is common for a single lexeme to appear in both train and test dictionary—in the form of different word inflections. This circumstance is undesirable in evaluation settings as it can lead to performance overstatements—a model can `memorize' the corresponding target lemma, which ultimately reduces the translation task to a much easier task of finding the most appropriate inflection. Finally, most of the available BLI resources include English in each language pair and, given how morphologically impoverished English is, those resources are unsuitable for analysis of morphological generalization.

Morphological Dictionaries ::: Our Dictionaries

To address the shortcomings of the existing evaluation, we built 40 new morphologically complete dictionaries, which contain most of the inflectional paradigm of every word they contain. This enables a more thorough evaluation and makes the task much more challenging than traditional evaluation sets. In contrast to the existing resources our dictionaries consist of many rare forms, some of which are out-of-vocabulary for large-scale word embeddings such as fastText. Notably, this makes them the only resource of this kind that enables evaluating open-vocabulary BLI.

We focus on pairs of genetically-related languages for which we can cleanly map one morphological inflection onto another. We selected 5 languages from the Slavic family: Polish, Czech, Russian, Slovak and Ukrainian, and 5 Romance languages: French, Spanish, Italian, Portuguese and Catalan. Table TABREF5 presents an example extract from our resource; every source–target pair is followed by their corresponding lemmata and a shared tag.

We generated our dictionaries automatically based on openly available resources: Open Multilingual WordNet BIBREF12 and Extended Open Multilingual WordNet BIBREF13, both of which are collections of lexical databases which group words into sets of synonyms (synsets), and UniMorph BIBREF14—a resource comprised of inflectional word paradigms for 107 languages, extracted from Wiktionary and annotated according to the UniMorph schema BIBREF15. For each language pair $(L1, L2)$ we first generated lemma translation pairs by mapping all $L1$ lemmata to all $L2$ lemmata for each synset that appeared in both $L1$ and $L2$ WordNets. We then filtered out the pairs which contained lemmata not present in UniMorph and generated inflected entries from the remaining pairs: one entry for each tag that appears in the UniMorph paradigms of both lemmata. The sizes of dictionaries vary across different language pairs and so does the POS distribution. In particular, while Slavic dictionaries are dominated by nouns and adjectives, verbs constitute the majority of pairs in Romance dictionaries. We report the sizes of the dictionaries in Table TABREF6. In order to prevent semantic leakage, discussed in §SECREF4, for each language pair we split the initial dictionary into train, development and test splits so that each sub-dictionary has its own, independent set of lemmata. In our split, the train dictionary contains 60% of all lemmata, while the development and test dictionaries each have 20% of the lemmata.

Morphological Dictionaries ::: Comparison with MUSE

In this section we briefly outline important differences between our resource and the MUSE dictionaries BIBREF2 for Portuguese, Italian, Spanish, and French (12 dictionaries in total). We focus on MUSE as it is one of the few openly available resources that covers genetically-related language pairs.

Morphological Dictionaries ::: Comparison with MUSE ::: Word Frequency

The first and most prominent difference lies in the skew towards frequent word forms in MUSE evaluation. While our test dictionaries contain a representative sample of forms in lower frequency bins, the majority of forms present in MUSE are ranked in the top 10k in their respective language vocabularies. This is clearly presented in Figure FIGREF2 for the French–Spanish resource and also holds for the remaining 11 dictionaries.

Morphological Dictionaries ::: Comparison with MUSE ::: Morphological Diversity

Another difference lies in the morphological diversity of both dictionaries. The average proportion of paradigm covered for lemmata present in MUSE test dictionaries is 53% for nouns, 37% for adjectives and only 3% for verbs. We generally observe that for most lemmata the dictionaries contain only one inflection. In contrast, for our test dictionaries we get 97% coverage for nouns, 98% for adjectives and 67% for verbs. Note that we do not get 100% coverage as we are limited by the compatibility of source language and target language UniMorph resources.

Morphological Dictionaries ::: Comparison with MUSE ::: Train–test Paradigm Leakage

Finally, we carefully analyze the magnitude of the train–test paradigm leakage. We found that, on average 20% (299 out of 1500) of source words in MUSE test dictionaries share their lemma with a word in the corresponding train dictionary. E.g. the French–Spanish test set includes the form perdent—a third-person plural present indicative of perdre (to lose) which is present in the train set. Note that the splits we provide for our dictionaries do not suffer from any leakage as we ensure that each dictionary contains the full paradigm of every lemma.

Bilingual Lexicon Induction

The task of bilingual lexicon induction is well established in the community BIBREF16, BIBREF17 and is the current standard choice for evaluation of cross-lingual word embedding models. Given a list of N source language word forms $x_1, \ldots , x_N$, the goal is to determine the most appropriate translation t_i, for each query form x_i. In the context of cross-lingual embeddings, this is commonly accomplished by finding a target language word that is most similar to x_i in the shared semantic space, where words' similarity is usually computed using a cosine between their embeddings. The resulting set of (x_i, t_i) pairs is then compared to the gold standard and evaluated using the precision at k (P@k) metric, where k is typically set to 1, 5 or 10. Throughout our evaluation we use P@1, which is equivalent to accuracy.

In our work, we focus on the supervised and semi-supervised settings in which the goal is to automatically generate a dictionary given only monolingual word embeddings and some initial, seed translations. For our experiments we selected the models of BIBREF3, BIBREF4 and BIBREF5—three of the best performing BLI models, which induce a shared cross-lingual embedding space by learning an orthogonal transformation from one monolingual space to another (model descriptions are given in the supplementary material). In particular, the last two employ a self-learning method in which they alternate between a mapping step and a word alignment (dictionary induction) step in an iterative manner. As we observed the same general trends across all models, in the body of the paper we only report the results for the best performing model of BIBREF5. We present the complete set of results in the supplementary material.

Bilingual Lexicon Induction ::: Experimental Setup

We trained and evaluated all models using the Wikipedia fastText embeddings BIBREF19. Following the existing work, for training we only used the most frequent 200k words in both source and target vocabularies. To allow for evaluation on less frequent words, in all our experiments the models search through the whole target embedding matrix at evaluation (not just the top 200k words, as is common in the literature). This makes the task more challenging, but also gives a more accurate picture of performance. To enable evaluation on the unseen word forms we generated a fastText embedding for every out-of-vocabulary (OOV) inflection of every word in WordNet that also appears in UniMorph. We built those embeddings by summing the vectors of all n-grams that constitute an OOV form. In the OOV evaluation we append the resulting vectors to the original embedding matrices.

Morphological Generalization

We propose a novel quadripartite analysis of the BLI models, in which we independently control for four different variables: (i) word form frequency, (ii) morphology, (iii) lexeme frequency and (iv) lexeme. We provide detailed descriptions for each of those conditions in the following sections. For each condition, we analyzed all 40 language pairs for each of our selected models—a total of 480 experiments. In the body of the paper we only present a small representative subset of our results.

Morphological Generalization ::: Controlling for Word Frequency

For highly inflectional languages, many of the infrequent types are rare forms of otherwise common lexemes and, given the morphological regularity of less frequent forms, a model that generalizes well should be able to translate those capably. Thus, to gain insight into the models' generalization ability we first examine the relation between their performance and the frequency of words in the test set.

We split each test dictionary into 9 frequency bins, based on the relative frequencies of words in the original training corpus for the word embeddings (Wikipedia in the case of fastText). More specifically, a pair appears in a frequency bin if its source word belongs to that bin, according to its rank in the respective vocabulary. We also considered unseen words that appear in the test portion of our dictionaries, but do not occur in the training corpus for the embeddings. This is a fair experimental setting since most of those OOV words are associated with known lemmata. Note that it bears a resemblance to the classic Wug Test BIBREF20 in which a child is introduced to a single instance of a fictitious object—`a wug'—and is asked to name two instances of the same object—`wugs'. However, in contrast to the original setup, we are interested in making sure the unseen inflection of a known lexeme is properly translated.

Figure FIGREF18 presents the results on the BLI task for four example language pairs: two from the Slavic and two from the Romance language family. The left-hand side of the plots shows the performance for the full dictionaries (with and without OOVs), while the right-hand side demonstrates how the performance changes as the words in the evaluation set become less frequent. The general trend we observe across all language pairs is an acute drop in accuracy for infrequent word forms—e.g. for Catalan–Portuguese the performance falls from 83% for pairs containing only the top 10k most frequent words to 40% for pairs, which contain source words ranked between 200k and 300k.

Morphological Generalization ::: Controlling for Morphology

From the results of the previous section, it is not clear whether the models perform badly on inflections of generally infrequent lemmata or whether they fail on infrequent morphosyntactic categories, independently of the lexeme frequency. Indeed, the frequency of different morphosyntactic categories is far from uniform. To shed more light on the underlying cause of the performance drop in sec:freqcontrol, we first analyze the differences in the models' performance as they translate forms belonging to different categories and, next, look at the distribution of these categories across the frequency bins.

In Table TABREF26 we present our findings for a representative sample of morphosyntactic categories for one Slavic and one Romance language pair (we present the results for all models and all language pairs in the supplementary material). It illustrates the great variability across different paradigm slots—both in terms of their frequency and the difficulty of their translation.

As expected, the performance is best for the slots belonging to the highest frequency bins and forms residing in the rarer slots prove to be more challenging. For example, for French–Spanish the performance on , and is notably lower than that for the remaining categories. For both language pairs, the accuracy for the second-person plural present imperative () is particularly low: 1.5% accuracy for French–Spanish and 11.1% for Polish–Czech in the in-vocabulary setting. Note that it is unsurprising for an imperative form, expressing an order or command, to be infrequent in the Wikipedia corpora (the resource our monolingual embeddings were trained on). The complex distribution of the French across the frequency bins is likely due to syncretism—the paradigm slot shares a form with 2nd person present plural slot, . Our hypothesis is that syncretism may have an effect on the quality of the monolingual embeddings. To our knowledge, the effect of syncretism on embeddings has not yet been systematically investigated.

Morphological Generalization ::: Controlling for Lexeme Frequency

To get an even more complete picture, we inspect how the performance on translating inflections of common lemmata differs to translating forms coming from less frequent paradigms by controlling for the frequency of the lexeme. We separated our dictionaries into two bins based on the relative frequency of the source lexeme. We approximated frequency of the lexemes by using ranks of their most common inflections: our first bin contained lexemes whose most common inflection is ranked in the top 20k forms in its respective vocabulary, while the second bin consisted of lexemes with most common inflection ranked lower than 60k. We present the results for the same morphosyntactic categories as in §SECREF27 on the left side of the graphs in Figure FIGREF30. As anticipated, in the case of less frequent lexemes the performance is generally worse than for frequent ones. However, perhaps more surprisingly, we discover that some morphosyntactic categories prove to be problematic even for the most frequent lexemes. Some examples include the previously mentioned imperative verb form or, for Slavic languages, singular dative nouns ().

Morphological Generalization ::: Controlling for Lexeme

We are, in principle, interested in the ability of the models to generalize morphologically. In the preceding sections we focused on the standard BLI evaluation, which given our objective is somewhat unfair to the models—they are additionally punished for not capturing lexical semantics. To gain more direct insight into the models' generalization abilities we develop a novel experiment in which the lexeme is controlled for. At test time, the BLI model is given a set of candidate translations, all of which belong to the same paradigm, and is asked to select the most suitable form. Note that the model only requires morphological knowledge to successfully complete the task—no lexical semantics is required. When mapping between closely related languages this task is particularly straightforward, and especially so in the case of fastText where a single n-gram, e.g. the suffix -ing in English as in the noun running, can be highly indicative of the inflectional morphology of the word.

We present results on 8 representative language pairs in Table TABREF35 (column Lexeme). We report the accuracy on the in-vocabulary pairs as well as all the pairs in the dictionary, including OOVs. As expected, compared to standard BLI this task is much easier for the models—the performance is generally high. For Slavic languages numbers remain high even in the open-vocabulary setup, which suggests that the models can generalize morphologically. On the other hand, for Romance languages we observe a visible drop in performance. We hypothesize that this difference is due to the large quantities of verbs in Romance dictionaries; in both Slavic and Romance languages verbs have substantial paradigms, often of more than 60 forms, which makes identifying the correct form more difficult. In contrast, most words in our Slavic dictionaries are nouns and adjectives with much smaller paradigms.

Following our analysis in sec:parcontrol, we also examine how the performance on this new task differs for less and more frequent paradigms, as well as across different morphosyntactic categories. Here, we exhibit an unexpected result, which we present in the two right-hand side graphs of Figure FIGREF30: the state-of-the-art BLI models do generalize morphologically for frequent slots, but do not generalize for infrequent slots. For instance, for the Polish–Czech pair, the models achieve 100% accuracy on identifying the correct inflection when this inflection is , , or for frequent and, for the first two categories, also the infrequent lexemes; all of which are common morphosyntactic categories (see Table TABREF26). The results from Figure FIGREF30 also demonstrate that the worst performing forms for the French–Spanish language pair are indeed the infrequent verbal inflections.

Morphological Generalization ::: Experiments on an Unrelated Language Pair

So far, in our evaluation we have focused on pairs of genetically-related languages, which provided an upper bound for morphological generalization in BLI. But our experimental paradigm is not limited to related language pairs. We demonstrate this by experimenting on two example pairs of one Slavic and one Romance language: Polish–Spanish and Spanish–Polish. To construct the dictionaries we followed the procedure discussed in §SECREF2, but matched the tags based only on the features exhibited in both languages (e.g. Polish can be mapped to in Spanish, as Spanish nouns are not declined for case). Note that mapping between morphosyntactic categories of two unrelated languages is a challenging task BIBREF21, but we did our best to address the issues specific to translation between Polish and Spanish. E.g. we ensured that Spanish imperfective/perfective verb forms can only be translated to Polish forms of imperfective/perfective verbs.

The results of our experiments are presented in the last two rows of Table TABREF35 and, for Polish–Spanish, also in Figure FIGREF39. As expected, the BLI results on unrelated languages are generally, but not uniformly, worse than those on related language pairs. The accuracy for Spanish–Polish is particularly low, at 28% (for in vocabulary pairs). We see large variation in performance across morphosyntactic categories and more and less frequent lexemes, similar to that observed for related language pairs. In particular, we observe that —the category difficult for Polish–Czech BLI is also among the most challenging for Polish–Spanish. However, one of the highest performing categories for Polish–Czech, , yields much worse accuracy for Polish–Spanish.

Morphological Generalization ::: Adding a Morphological Constraint

In our final experiment we demonstrate that improving morphological generalization has the potential to improve BLI results. We show that enforcing a simple, hard morphological constraint at training time can lead to performance improvements at test time—both on the standard BLI task and the controlled for lexeme BLI. We adapt the self-learning models of BIBREF4 and BIBREF5 so that at each iteration they can align two words only if they share the same morphosyntactic category. Note that this limits the training data only to word forms present in UniMorph, as those are the only ones for which we have a gold tag. The results, a subset of which we present in Table TABREF35, show that the constraint, despite its simplicity and being trained on less data, leads to performance improvements for every Romance language pair and many of the Slavic language pairs. We take this as evidence that properly modelling morphology will have a role to play in BLI.

Discussion and Conclusion

We conducted a large-scale evaluation of the generalization ability of the state-of-the-art bilingual lexicon inducers. To enable our analysis we created 40 morphologically complete dictionaries for 5 Slavic and 5 Romance languages and proposed a novel experimental paradigm in which we independently control for four different variables.

Our study is the first to examine morphological generalization in BLI and it reveals a nuanced picture of the interplay between performance, the word's frequency and morphology. We observe that the performance degrades when models are evaluated on less common words—even for the infrequent forms of common lexemes. Our results from the controlled for lexeme experiments suggest that models are able to generalize well for more frequent morphosyntactic categories and for part-of-speech with smaller paradigms. However, their ability to generalize decreases as the slots get less frequent and/or the paradigms get larger. Finally, we proposed a simple method to inject morphological knowledge and demonstrated that making models more morphologically aware can lead to general performance improvements.