Morphological Paradigm Completion with LSTM Neural Models

Conor Stuart Roe

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Arts in Linguistics and Computer Science

October 11, 2019

Version: Draft 0

Advisor: Jane Chandlee

Contents

1 Literature Review

1.1 History and Motivation

Historically, in language technologies and modeling, morphology has been somewhat under-emphasized. This is probably due at least in part to the below-average morphological complexity of English. (Cotterell and Heigold, 2017) English lexemes have relatively few inflected forms, so occurrences of out-of-vocabulary inflected forms are less frequent. And even with grammatical information thrown out via lemmatization, English may still be fairly sensible.

However, in many languages grammatical inflection carries much more semantic burden, and the number of possible inflected forms can be much greater. For instance, a single verb in the Archi language can be inflected in 1,725 ways. (Cotterell, Kirov, Sylak-Glassman, Yarowsky, et al., 2016) For such highly inflected languages, data sparsity is a significant problem for language models naive to morphology. In languages with a high number of possible forms per lexeme, a much larger number out-of-vocabulary (OOV) are inevitably encountered in test data, requiring reliance on a model's representation of OOV words. Even inflected forms that do occur in training data may not appear often enough for a naive model to understand their semantic content. A model that considers grammatical categories separately is better able to understand the semantic content of inflected words. (Cotterell, Kirov, Sylak-Glassman, Yarowsky, et al., 2016) It has been empirically shown that comprehending grammatical categories and inflection improves outcomes in language modeling. (Faruqui et al., 2015)

1.2 Machine learning of morphology: sub-problems and related problems

Within the realm of machine understanding of morphology, there are many sub-problems and related problems. The most basic areas of research involve predicting the morphology of words in isolation - transforming a word into a specific morphological form, or the inverse, tagging a form with its morphological categories. Typically, this has been done with words in regular grammatical paradigms, such as number and case marking on nouns, or tense, aspect, mood, and/or argument agreement patterns on verbs.

1.2.1 Core supervised learning problems

Some of the earliest work in computational morphology involves making specific morphological transformations. That is, given a particular form of a lexeme (often, but not necessarily, a citation form), predicting another form. An example would be learning to pluralize English nouns, e.g., *show* $\rightarrow$ *showed*, *see* $\rightarrow$ *saw*, etc. (Dreyer et al., 2008)

The natural extension of this is aiming to be able to predict any inflected form given a one form and an arbitrary set of morphological categories. Generally speaking, a citation form has been used as input. (Durrett and DeNero, 2013) (Faruqui et al., 2015) (Cotterell and Heigold, 2017) The related "reinflection" problem involves being given any inflected form as input, and transforming it into any other. (Cotterell, Kirov, Sylak-Glassman, Yarowsky, et al., 2016)

A further extension of the morphology generation problem is the generation of complete inflection tables. The exact nature of this problem depends on the type of training and test data used. If a model is only trained on a sparse, random sampling of forms for each lexeme, then a task may consist of filling out the rest of an inflection table for those lexemes. If a model is trained using entire inflection tables, then test data must consist of new lexemes. Another dimension along which

paradigm completion tasks differ is whether only a single citation form is provided as a prompt, or whether any form or even multiple forms may be provided as a prompt for producing a single table. (Hulden et al., 2014) (Ahlberg et al., 2015) (Cotterell and Heigold, 2017)

Overall, a diverse set of inflection shapes have been worked with. SIGMORPHON 2018 included data from 103 typologically diverse languages, and paradigms using suffixing, prefixing, infixing, reduplication, and non-concatenative morphology. However, all work has been with well-defined, tabular paradigms of fairly moderate size (≤ 200 forms). (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018) Other types of morphology such as derivational morphology and cliticization, which have the potential to greatly expand the number of possible forms and in a less organized fashion, have been less well explored. Highly synthetic languages with more expansive types of word building, including incorporation and productive derivational morphology, are ripe for future work. (Cotterell, Kirov, Sylak-Glassman, Yarowsky, et al., 2016)

1.2.2 Related problems

Within only the last two or so years, there has been work on predicting morphology in context. In the 2018 and 2019 SIGMORPHON shared tasks, to which several teams of researchers submitted solutions, a sub-task was dedicated to cloze challenges in which one word in a sentence, given in citation form, was to be inflected based on context. (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018) (McCarthy et al., 2019) This work is essentially a synthesis of morphology generation and morphosyntactic modeling.

Since 2017, there has been some work done on learning curves for computational morphology. The datasets published for SIGMORPHON 2017 and 2018 include partitions into low (∼100 forms), medium (∼1000 forms), and high (∼10,000 forms) data training sets for the express purpose of assessing the learning curve of different models. The most successful models with high-data training sets are LSTM neural models, generally considered the state of the art model type, yet these often fare worse than baseline string transduction models with low data training

sets. (Cotterell and Heigold, 2017) (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018) Improving performance with small training sets is of interest, as much of the applicability of computational morphology models is to languages which don't already have high-quality technical tools or datasets.

The most recent new challenge that SIGMORPHON has tried to address is that of transfer learning of morphology, in the shared task earlier this year. Given a state of the art model trained on a language with a high volume of training data, teams were asked to alter it into a model that would perform well on a new language, given a smaller amount of training data for that language. 80% of the pairs of languages were closely related, while 20% were distantly or not at all related. Gains of transfer learning models between closely related languages generally performed better than transfer learning models between more distant languages. (McCarthy et al., 2019)

1.3 Non-neural approaches

1.3.1 Vector embedding

Representing words in relatively low-dimensional vector spaces, with the intention of capturing semantic and syntactic content in a principled way, has found success in a variety of computational linguistics tasks. Regularities in the relative location of semantically related words have been exploited for semantic analysis tasks. (Bilmes and Kirchhoff, 2003) (Alexandrescu and Kirchhoff, 2006).

Similarly, morphological changes may appear as spatial transformations in vector space, and work has been done on discovering morphological relationships between in-vocabulary words based on their relative spatial locations. This method has the limitation that it cannot extend to words for which a vector representation has not been trained, and so it cannot provide understanding of the many OOV forms encountered in test data of highly inflected languages. (Mikolov et al., 2013) (Soricut and Och, 2015) (Dos Santos and Zadrozny, 2014) (Cotterell and Schütze, 2019)

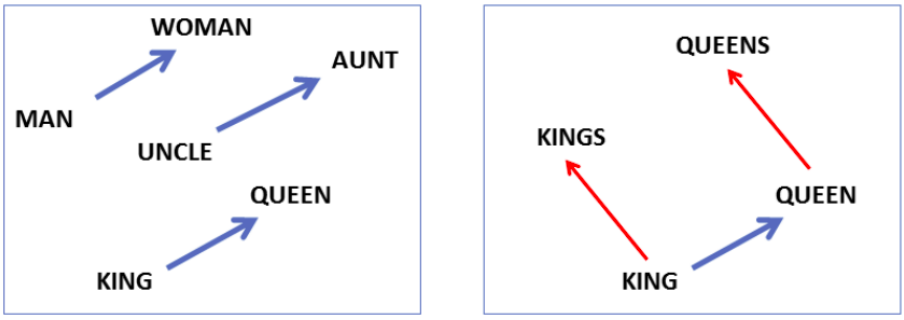

Fig. 1.1: Regular spatial transformations encode semantic or grammatical content (Mikolov et al., 2013)

1.3.2 String transduction

1.4 LSTM and other neural approaches

Since 2016, almost all work on paradigm completion has made use of long short-term memory (LSTM) or related gated recurrent network (GRU) models. (Faruqui et al., 2015) (Cotterell, Kirov, Sylak-Glassman, Yarowsky, et al., 2016) (Cotterell and Heigold, 2017) (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018) (McCarthy et al., 2019) In high-data training settings, encoder-decoder LSTMs have proven over the course of annual SIGMORPHON shared tasks to be the state of the art method of computational morphology. However, they have worse performance with sparse training data, an issue which has seen attempts to address.

2 Data

2.1 Structured paradigm data

The English Wiktionary, a collaborative online dictionary, has become something of a standard source of supervised morphological data. Durrett and DeNero, 2013 published a Wiktionary dataset of five paradigms from three languages (Finnish, Spanish, and German) which have been used as a benchmark in much future work. (Hulden et al., 2014) (Nicolai et al., 2015) (Ahlberg et al., 2015) (Faruqui et al., 2015) SIGMORPHON 2017 published a dataset partitioned into three training levels (100, 1000, and 10,000 tables), containing both sparse and full inflection tables for one or more parts of speech for each of 52 languages. Most of that data was derived from a January 2017 Wiktionary dump; data for four languages came from the Alexina project and data for Haida was prepared by Jordan Lachler. (Cotterell and Heigold, 2017)

The most complete structured dataset to date was published for the SIGMORPHON shared task 2018, a superset of the SIGMORPHON 2017 data, which is partitioned in the same manner and includes data from 103 languages. For most of the languages, data was scraped from Wiktionary.

2.2 Text corpora

2.3 Representation of morphology

In earlier work, morphology is encoded in a language-specific and model-specific way.

(Luong et al., 2013) The UniMorph project, initially published in 2015, makes an effort to encode morphological categories uniformly cross-linguistically. (Sylak-Glassman, Kirov, Post, et al., 2015) (Sylak-Glassman, Kirov, Yarowsky, et al., 2015) (Sylak-Glassman, 2016) The SIGMORPHON data since 2018 has been published in UniMorph format.

3 Potential research directions

3.1 Research gaps

The research so far has been conducted on a fairly comprehensive set of languages, covering many language families and types of grammatical inflection. However, there are areas that stand out as gaps, many of which have been noted as potential research directions for the field.

3.1.1 Highly synthetic languages and large paradigms

Although certain highly synthetic languages such as Navajo have been modeled, the grammatical paradigms of those languages have not been modeled in their entirety. The data sources used, primarily Wiktionary, segment paradigms, and certain grammatical dimensions are either treated as lexical or simply not listed. For instance, in Wiktionary's Navajo data, only subject agreement patterns are explicitly given; tables may be listed for a few aspects, but most aspectual categories, as well as object agreement, thematic and classifier distinctions are not given. The paradigms trained in existing SIGMORPHON tasks are limited to 100-200 forms at the most, even though there are paradigms in some languages an order of magnitude or more greater. The chief challenges in using larger paradigms include producing them manually and handling the much larger sets of individual forms they entail.

3.1.2 Less paradigmatic morphology

There has also been no real exploration into less paradigmatic types of morphology, including cliticization and more compositional or derivational morphology. Highly

synthetic languages such as the Inuit languages, of a type that rely more on compositional morphology and other strategies like noun incorporation, would have no way of being effectively modeled given the current scope of research.

The major hurdle in modeling less paradigmatic morphology, similar to very large paradigms, is acquiring and structuring data. Existing datasets of these types are much harder to find, and some reliance on unannotated text corpora is probably necessary. Thus, any research in these directions must concern itself with new strategies for data acquisition and labeling.

3.1.3 Transfer learning pairs

A last area is more pointedly investigating the effectiveness of transfer learning pairs. Only 79 transfer pairs have been tested, while the possibility space of transfer learning pairs is much larger than that of individual languages. A further advantage of this research question is that data already exists - it's only a matter of combining pairs of the many existing structured paradigm training sets. Since the 2019 shared task explicitly focused on related language pairs, there is room to conduct testing of unrelated pairs, and hopefully learn more about what factors other than genetic relation can make for effective transfer learning pairs. Such a research question has a real-world justification as well - many currently under-resourced languages are not closely related to major world languages, and if modeling of those languages can benefit from large existing datasets, the language technology toolkits for those language communities could see considerable improvement. Such an investigation would be aided by a dataset of typological information about the languages for which training data exists, so that regular patterns in transfer learning effectiveness can be identified. Much of this information can probably be derived from typological databases such as WALS, though the dataset may need to be manually constructed to some degree. Fortunately, it would be much smaller than most NLP datasets!

3.2 Research proposal

My choice of research direction will ultimately depend on the nature of the data that I can locate, but investigating transfer learning is the most straightforward direction from a data acquisition perspective. In the coming week, I will conduct a concerted search for useful datasets, but in the meantime I tentatively propose constructing a typological dataset specific to the SIGMORPHON 2018 languages, and using it to identify gaps in the transfer learning tests to date and explore which types of such pairs are most effective.

Bibliography

Ahlberg, Malin, Markus Forsberg, and Mans Hulden (Jan. 2015). „Paradigm classification in supervised learning of morphology“. In: pp. 1024–1029 (cit. on pp. 3, 6).

Alexandrescu, Andrei and Katrin Kirchhoff (Jan. 2006). „Factored Neural Language Models.“ In: (cit. on p. 4).

Bilmes, Jeff A. and Katrin Kirchhoff (2003). „Factored Language Models and Generalized Parallel Backoff“. In: *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pp. 4–6 (cit. on p. 4).

Cotterell, Ryan and Georg Heigold (2017). „Cross-lingual, Character-Level Neural Morphological Tagging“. In: *CoRR* abs/1708.09157. arXiv: 1708.09157 (cit. on pp. 1–6).

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Géraldine Walther, et al. (Oct. 2018). „The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection“. In: *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*. Brussels: Association for Computational Linguistics, pp. 1–27 (cit. on pp. 3–5).

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, et al. (Aug. 2016). „The SIGMORPHON 2016 Shared Task—Morphological Reinflection“. In: (cit. on pp. 1–3, 5).

Cotterell, Ryan and Hinrich Schütze (2019). „Morphological Word Embeddings“. In: *CoRR* abs/1907.02423. arXiv: 1907.02423 (cit. on p. 4).

Dos Santos, Cıcero Nogueira and Bianca Zadrozny (2014). „Learning Character-level Representations for Part-of-speech Tagging“. In: ICML'14, pp. II-1818–II-1826 (cit. on p. 4).

Dreyer, Markus, Jason Smith, and Jason Eisner (Jan. 2008). „Latent-Variable Modeling of String Transductions with Finite-State Methods.“ In: pp. 1080–1089 (cit. on p. 2).

Durrett, Greg and John DeNero (2013). „Supervised Learning of Complete Morphological Paradigms“. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics* (cit. on pp. 2, 6).

Faruqui, Manaal, Yulia Tsvetkov, Graham Neubig, and Chris Dyer (2015). „Morphological Inflection Generation Using Character Sequence to Sequence Learning“. In: *CoRR* abs/1512.06110. arXiv: 1512.06110 (cit. on pp. 1, 2, 5, 6).

Hulden, Mans, Markus Forsberg, and Malin Ahlberg (Apr. 2014). „Semi-supervised learning of morphological paradigms and lexicons“. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 569–578 (cit. on pp. 3, 6).

Luong, Thang, Richard Socher, and Christopher Manning (Aug. 2013). „Better Word Representations with Recursive Neural Networks for Morphology“. In: pp. 104–113 (cit. on p. 7).

McCarthy, Arya D., Ekaterina Vylomova, Shijie Wu, et al. (Aug. 2019). „The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection“. In: *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Florence, Italy: Association for Computational Linguistics, pp. 229–244 (cit. on pp. 3–5).

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (June 2013). „Linguistic Regularities in Continuous Space Word Representations“. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751 (cit. on pp. 4, 5).

Nicolai, Garrett, Colin Cherry, and Grzegorz Kondrak (Jan. 2015). „Inflection Generation as Discriminative String Transduction“. In: pp. 922–931 (cit. on p. 6).

Soricut, Radu and Franz Och (Jan. 2015). „Unsupervised Morphology Induction Using Word Embeddings“. In: pp. 1627–1637 (cit. on p. 4).

Sylak-Glassman, John (2016). „The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema)“. In: (cit. on p. 7).

Sylak-Glassman, John, Christo Kirov, Matt Post, Roger Que, and David Yarowsky (2015). „A Universal Feature Schema for Rich Morphological Annotation and Fine-Grained Cross-Lingual Part-of-Speech Tagging“. In: *Systems and Frameworks for Computational Morphology*. Ed. by Cerstin Mahlow and Michael Piotrowski. Cham: Springer International Publishing, pp. 72–93 (cit. on p. 7).

Sylak-Glassman, John, Christo Kirov, David Yarowsky, and Roger Que (July 2015). „A Language-Independent Feature Schema for Inflectional Morphology“. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, pp. 674–680 (cit. on p. 7).