Prospectus & Annotated Bibliography

Conor Stuart Roe

September 20, 2019

Prospectus

In highly inflected languages, a very high proportion of specific word forms appearing in a corpus may be hapax legomena, only appearing once. This presents a challenge for language technology which seeks to grammatically annotate texts, or linguists who wish to predict unattested forms. A single verb lemma in Archi, a Nakh-Daghestanian language, can be conservatively said to have 1,725 inflected forms [5]. Where thousands of inflected forms are possible for a given lemma or where a large body of vocabulary needs to be analyzed, manual annotation of inflected forms is impractical and rule-based approaches may be difficult to produce and of limited reliability, depending on morphophonological and orthographic complexity or morphological irregularity.

Neural networks have distinctly outshone other means of predicting inflections or grammatical tags that over broad scope of vocabulary [5], and supervised learning dominates this space [12]. Supervised deep learning approaches have achieved nearly perfect results in predicting part of speech [7] and considerable success at generating inflected forms given morphological category parameters [5]. Headway has even been made on unsupervised learning of morphology [12].

Specifically named gaps in research seem to include analysis of derivational morphology and reduplicating morphology [5]. Most of all though, a proliferation of word embedding strategies and neural network architectures have been used for different tasks, and it may be a ripe area of study to test the

limits of some particular design choice.

Word embeddings in a vector space, learned through methods such as Skip-Gram, have proven valuable in a variety of language tasks, including morphological analysis [11] and part of speech tagging [7]. However, character-level data has augmented results in both of these applications [3] [7].

In addition, both recurrent neural networks [5] [12] and convolutional networks [11] have proven useful for morphology learning tasks. A further exploration of the contrastive value of these architectures could be interesting. Because of the potential utility of a bidirectional morphology engine, I'm also interested in exploring the possibility of adapting current recurrent neural network models to be bidirectional. Research suggests that bidirectional, and thus non-forgetting, recurrent neural networks have particular limitations that may require some creativity to deal with [10].

Such a tool would have the greatest usefulness for highly inflected languages for which existing resources are limited. However, it can only be effectively tested on languages with sufficient data about grammatical paradigms of a large body of vocabulary, so that test data can be used to assess model competence. In the interest of ensuring that a cross-linguistically useful approach is produced, I propose supplying an algorithm with supervised data for several languages and assessing its competence on each. For identifying such test languages I propose three criteria: moderate to high morphological complexity, transparent orthography, and sufficient corpus and grammatical

data.

I have ascertained that sufficient supervised data can be taken from Wiktionary for Spanish, Russian, Turkish, and Hungarian, all of which have been used in related work [5] and fit my criteria. These languages all fit similar profiles of inflection style, all being heavily suffixing and avoiding reduplication. Candidate languages which may diversify my data set include Arabic, Maltese, Malay, and Navajo [5]. However, existing Arabic data on Wiktionary may present data consistency issues [5], and the others have much more sparse informaiton on Wiktionary.

The steps involved in building this approach include writing a literature review to better understand existing tools for morphological inference and identify current research gaps; the specification, identification, and formatting of training data; the proposition of an algorithmic design; and the implementation, assessment, and iterative revision of that algorithm.

I have already ascertained the existence of sufficient linguistic resources for Spanish, Russian, Turkish, and Hungarian on Wiktionary, and identified research precedent for using Wiktionary data [5]. I do not forsee the need for plaintext corpus data unless I follow the route of Wolf-Sonkin et. al. 2018. Future steps in preliminary data identification is deciding on whether to use more languages, and choosing which if so, in addition to potentially writing a proof-of-concept Wiktionary scraper.

The next research step for me is to continue to read papers, with a particular eye to contrasting neural network architectures, and perhaps use other materials to better understand them. I have already identified a number of papers to read next which have immediate relevance to me ([1], [3], [8], [12], [6], [4], [9]).

Annotated Bibliography

Radu Soricut and Franz Och. Unsupervised morphology induction using word embeddings. pages 1627–1637, 01 2015

- The authors present a means of induction of morphological alternations that uses only an unannotated monolingual corpus. They claim in addition that it works well across languages and language families, predicts unseen forms, discovers the semantic content of morphological categories, and identify which pairs of words are actually morphologically related, as opposed to being spuriously similar. My main points of skepticism are the simplistic manner of morphological modeling (simple one-to-one prefix and suffix alternation) and the claim that meaningful semantic information can be inferred.

- Perhaps the simplicity of the morphological model provides room for my project to build on. If I can build a more sophisticated morphological pattern recognizer, it would be interesting to try plugging it into their algorithm in some way.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James R. Glass. What do neural machine translation models learn about morphology? *CoRR*, abs/1704.03471, 2017

- Here, the authors present a means of assessing the inferences that neural machine translation systems make about morphology, and draw several conclusions. Least surprisingly, a convolutional neural network operating over character embeddings learns morphological patterns much

more effectively than a system using word embeddings. They also claim that lower network layers carry more morphological information, and that having a morphologically poor target language actually improves the system's learning of source language morphology.

- Some of the conclusions here are potentially very relevant to me, specifically the notes about which parts of a neural network architecture are best suited to learn morphology. I need to read more about the convolutional architecture they use to analyze character embeddings - dealing with character embeddings will be one of the most important tasks of my project.

Matthew MacKay, Paul Vicol, Jimmy Ba, and Roger B. Grosse. Reversible recurrent neural networks. *CoRR*, abs/1810.10999, 2018

- To achieve the goal of producing a system which can both provide a morphological analysis given a word and vice versa, I'd need a deep learning system which can be used bidirectionally. That is, such an architecture would need to be able to accept either end as input and produce output from the other, without loss of information.

- Here, a class of reversible, non-forgetting RNN architectures is presented, which has been designed for another goal - to avoid the memory-intensive backpropagation step of standard LSTM and GRU RNNs. The authors claim to demonstrate severe limitations of this class - inability to forget information means that reversible RNNs simply become overloaded on too long an input. These concerns do not apply to

my project because my input sizes will be limited, reducing the memory burden of backpropagation or a no-forgetting architecture, so this model may be a good fit.

- The authors also cite Gomez et. al. 2017 [8] in saying of residual neural networks: "Because this mapping is reversible with an easily computable Jacobian determinant, maximum likelihood training is efficient." Reversible residual networks sounds like a more developed area; if I can accomplish my goals with a residual neural network, that may be simpler.

Cícero Nogueira Dos Santos and Bianca Zadrozny. Learning character-level representations for part-of-speech tagging. pages II–1818–II–1826, 2014

- This paper is one of the closest I've found to my own project idea. They use a convolutional neural architecture on character embeddings, in additional to more typical word vector embeddings, to extract a piece of morphological category information, namely part of speech. They claim to have surpassed other means of part of speech tagging, attaining up to 97% accuracy. At the moment I don't fully understand their convolutional neural architecture, but it appears to be a promising method for extracting morphological information from words.

- I have two main concerns in adapting their methodlogy. The first is whether comparable results can be achieved with character embeddings alone, as opposed to the addition of word embeddings, which would require conducting SkipGram or a similar method on a plaintext corpus.

The second is whether their convolutional architecture can be made bidirectional without introducing new limitations.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. The sigmorphon 2016 shared task—morphological reinflection. August 2016

- The task in this paper is very similar to what I'd like to accomplish - given any inflected form of a word, "re-inflect" it in a specified way. That is, replace all current inflection with inflection for a new set of categories. Their approach is very similar to what I initially imagined: a recurrent neural network architecture design to work cross-linguistically (they tested their system on ten languages, including all four I am considering using).

- Their system does a similar task to one direction of my idea, transforming a set of morphological categories and a lemma into a correct inflected form. It does not, however, do the reverse, and MacKay et. al. 2018 among others make me think that adapting their system to be bidirectional might not be straightforward. Tackling the challenge of bidirectionality with this type of system might be an interesting research niche for me to follow.

- Other interesting directions they name for future research include generating entire morphological paradigms, analyzing reduplicating morphology, and analyzing derivational morphology. The last idea might tie in nicely with Dos Santos and Zadrozny 2014.

- One thing I found very interesting about their approach is that their output is a sequence of string edits, which transform the input word to the output, rather than directly producing an output word.

- A valuable piece of information they provide is a statistical analysis, separate from their main work, of how frequently their subject languages were prefixing, suffixing, and apophonic. According to their data, the four languages I named as possible subjects (Spanish, Turkish, Russian, and Hungarian), are all extremely suffixing. It might be valuable for me to incorporate less purely suffixing languages, such as Maltese or Navajo.

Bibliography

[1] Andrei Alexandrescu and Katrin Kirchhoff. Factored neural language models. 01 2006.

[2] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James R. Glass. What do neural machine translation models learn about morphology? *CoRR*, abs/1704.03471, 2017.

[3] Ryan Cotterell and Georg Heigold. Cross-lingual, character-level neural morphological tagging. *CoRR*, abs/1708.09157, 2017.

[4] Ryan Cotterell, Christo Kirov, Sebastian J. Mielke, and Jason Eisner. Unsupervised disambiguation of syncretism in inflected lexicons. *CoRR*, abs/1806.03740, 2018.

[5] Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. The sigmorphon 2016 shared task—morphological reinflection. August 2016.

[6] Ryan Cotterell and Julia Kreutzer. Explaining and generalizing back-translation through wake-sleep. *CoRR*, abs/1806.04402, 2018.

[7] Cícero Nogueira Dos Santos and Bianca Zadrozny. Learning character-level representations for part-of-speech tagging. pages II–1818–II–1826, 2014.

[8] Aidan N. Gomez, Mengye Ren, Raquel Urtasun, and Roger B. Grosse. The reversible residual network: Backpropagation without storing activations. *CoRR*, abs/1707.04585, 2017.

[9] Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. pages 104–113, August 2013.

[10] Matthew MacKay, Paul Vicol, Jimmy Ba, and Roger B. Grosse. Reversible recurrent neural networks. *CoRR*, abs/1810.10999, 2018.

[11] Radu Soricut and Franz Och. Unsupervised morphology induction using word embeddings. pages 1627–1637, 01 2015.

[12] Lawrence Wolf-Sonkin, Jason Naradowsky, Sebastian J. Mielke, and Ryan Cotterell. A structured variational autoencoder for contextual morphological inflection. *CoRR*, abs/1806.03746, 2018.