The Effect of Linguistic Typology on Transfer Learning of Morphology

---

Conor Stuart Roe

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

December 9, 2019

Version: Final

Advisor: Jane Chandlee

https://github.com/cstuartroe/thesis

Abstract

In the SIGMORPHON 2019 shared task 1, multiple teams attempted for the first time to leverage transfer learning to build more accurate models of natural language morphology with small amounts of target domain data, with the intended goal of boosting modeling resources for low-resource languages. The performance of those models can be compared to the previous year's non-transfer learning task on the same data with the same goal, to find patterns in the efficacy of transfer learning to this computational problem. There is a highly robust relationship between similar verbal morphology and effective transfer learning outcomes between distantly related and unrelated source and target languages. Other relationships between the genealogical distance and data set similarities between source and target language also appear, but are less robust and have a less clear explanation. In order to ascertain the nature of these correlations suggested by existing data, a larger sampling of unrelated language pairs and a principled approach to isolating the effect of transfer learning should be undertaken. It may also be fruitful to measure lexical similarity and other types of typological similarities, such as the occurrence of morphological fusion and long-distance morphophonological processes, to investigate what effects they may have on transfer learning efficacy.

Acknowledgements

This thesis would not have been possible without the feedback of my student readers Tessa Pham and Anya Capps and my second faculty readers Amanda Payne and Steven Lindell, nor without the guidance of Sorelle Friedler, professor of the Haverford Computer Science department's thesis seminar, but most of all it has been enabled by the continued support and guidance of my advisor, Jane Chandlee.

Contents

1 Introduction and overview

In linguistics, morphology refers to alterations to words to reflect changes in meaning or grammatical category. For example, English verbs have differing morphological forms to indicate the simple present and simple past tenses, e.g., *show* → *showed*, *see* → *saw*, etc. (Dreyer et al., 2008). Grammatical inflection in particular has a tendency to be structured into **paradigms** - sets of all possible morphological forms that words of a certain type can take on, often shown arrayed in tables. The table below gives part of the paradigm for the verb *to see*:

	simple		progressive	
	3rd singular	3rd plural	3rd singular	3rd plural
present	(she) **sees**	(they) **see**	(she) **is seeing**	(they) **are seeing**
past	(she) **saw**	(they) **saw**	(she) **was seeing**	(they) **were seeing**

Historically, in language technologies and modeling, morphology has been somewhat under-emphasized. This is probably due at least in part to the dominance of English in language technology research, and its below-average morphological complexity (Cotterell, Kirov, Sylak-Glassman, et al., 2017). English lexemes tend to have few grammatically inflected forms compared to in most other languages. This means that in machine learning models which are given an English training corpus and then tested on new material, the occurrence of *out-of-vocabulary* (OOV) inflected forms - that is, forms in the test data that never occur in the training data - are less frequent.

Moreover, as can be seen above, many English grammatical forms are periphrastic, using several words to construct a single inflected form. This means that even with grammatical information removed via lemmatization, the process of reducing words to their citation form, an English sentence may still be fairly understandable. (The *citation form* of a lexeme is the form under which it would appear in a dictionary:

see, *sees*, *seeing*, and *saw* all belong to the same lexeme, with the citation form *see*.) For instance, given a sentence *she have see it* whose words have been lemmatized, a proficient reader of English can guess that the original sentence was *she has seen it* or *she had seen it*; not much information is lost.

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 (Kibrik, 1994). 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 of OOV forms 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 accuracy rates in language modeling and machine translation (Faruqui et al., 2015).

The state of the art for modeling morphology since 2016 has been variants of long short-term memory (LSTM) neural networks and the similar gated recurrent unit (GRU) architecture, operating over individual characters and grammatical category tags. These definitively overtook the field after their strong performance relative to traditional string transduction models in the SIGMORPHON 2016 shared task, a competition among several research teams on a morphology prediction problem. However, learning curve analysis in the SIGMORPHON 2017 task showed that such models, in particular encoder-decoder variants, perform well in high-data settings but, provided with lower volumes of training data, actually fare worse than string transduction models trained on similar amounts of data. To some degree, this poor performance could be ameliorated by pre-training on synthetic data, although this technique worked significantly better for languages with small inflectional paradigms (Cotterell, Kirov, Sylak-Glassman, et al., 2017). In the 2018 shared task, an identical morphology prediction task with data for more languages, the learning curve issue was addressed to some degree by ensembling with string transduction methods

(Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018), but for languages with low quantities of digital resources, there is still much room for improvement of computational morphology models.

One of the 2019 shared tasks focused on utilizing transfer learning to strengthen neural morphology models, using pretraining or combined models to leverage a high volume of data for one language to better model the morphology of another language for which a low volume of data was provided. The research teams tested their models on 100 transfer learning pairs, of which 80 were closely related languages. Transfer learning produced "modest" performance gains over models without access to transfer learning, with related pairs showing significantly more model improvement than unrelated pairs (McCarthy et al., 2019).

For some low-resource languages closely related to some high-resource language, the outcome of SIGMORPHON 2019 promises improved morphology modeling. But for the many low-resource languages may not closely related to other high-resource languages, it may be worthwhile to assess whether transfer learning can still be useful. It may be that pretraining on a typologically similar language, or simply on some other language(s) to understand the universal patterns of natural language morphology, can help neural models make more of sparse morphological data for low-resource languages. To this end, I'm planning on investigating what typological commonalities are predictive of more effective transfer learning of morphology between languages.

In order to accomplish this, I will need to compile a dataset of morphological typological characteristics of the languages I wish to work with, alongside their morphology data. SIGMORPHON 2018 published an annotated morphological data set for 103 languages, including various training data volumes as well as test data, which I will use for morphology learning; I'll compile typological information from an existing database, the World Atlas of Language Structures (WALS), and via string transduction analyses of the SIGMORPHON data. From there, I'll assess the efficacy of transfer learning between a large number of language pairs, and what relationship that has with various types of typological similarity between the languages. I'll also seek to answer some related questions, such as whether transfer learning efficacy

is typically reciprocal between two languages, how the efficacy of transfer learning interacts with learning curves, and potentially whether transferred knowledge of multiple languages can boost the performance of a morphology model even more.

2 Theoretical background

2.1 Natural language morphology and typology

2.2 LSTM and GRU neural modeling

Neural networks are a type of function approximation model inspired by the connection of neurons in animal brains. They have come to be implemented in a variety of forms, and underlie state of the art machine learning models in a variety of applications. The common feature of neural models is repeated matrix multiplication followed by the application of a non-linear "activation" function. Figure 2.4 illustrates a simple feed-forward network, in which a vector of three inputs is multiplied by some 3×4 matrix and activated to produce a vector of four intermediate values, which are again multiplied by some 4×2 matrix and activated to produce a vector of two outputs.

An RNN is a type of neural network that operates over sequences of inputs, typically with unknown length. Figure 2.5 illustrates the general structure: each input is represented by a vector, which is multiplied by a vector U to modify state, and the modified state is then multiplied by a vector W to produce an output vector and a matrix V to produce the next state. An LSTM network utilizes a specific means of using the inputs to modify state, represented by Figure 2.6, which includes the ability to modify the rate at which information is forgotten. LSTMs are intended to solve the forgetfulness of earlier RNN types, which tend to be unable to recall information from more than a few iterations prior (Hochreiter and Schmidhuber, 1997).

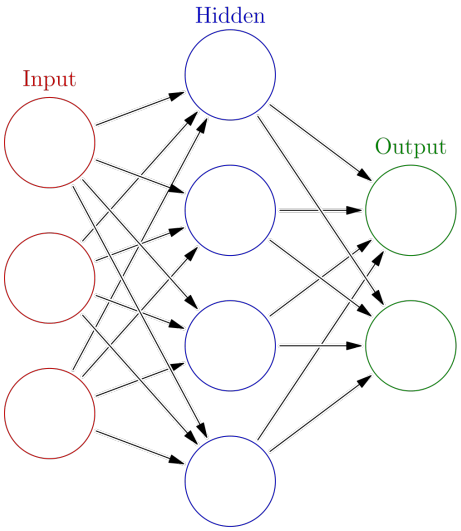


Fig. 2.1: The structure of a feed-forward neural network with three inputs, four hidden cells, and two outputs. (commons.wikimedia.org/wiki/File:Colored_neural_network.svg)

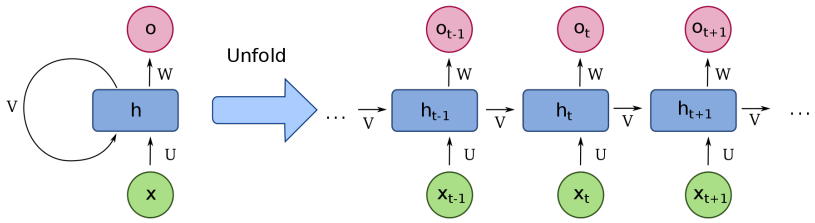


Fig. 2.2: The generalized structure of a recurrent neural network. (commons.wikimedia.org/wiki/File:Recurrent_neural_network_unfold.svg)

LSTMs have become the dominant model type in a variety of language tasks, including syntactic and morphological tasks. They significantly outperformed other types of models in SIGMORPHON 2016, since which time they have come to underlie nearly all morphology prediction models (Cotterell, Kirov, Sylak-Glassman, Yarowsky, et al., 2016, Cotterell, Kirov, Sylak-Glassman, et al., 2017).

The main differences between the LSTM architectures used in state of the art applications now, as evidenced by the four architectures used as baselines in the SIGMORPHON 2019 transfer learning task, are in their **attention mechanism**, the means by which hidden states are combined to generate sequential output (Cotterell and Schütze, 2019).

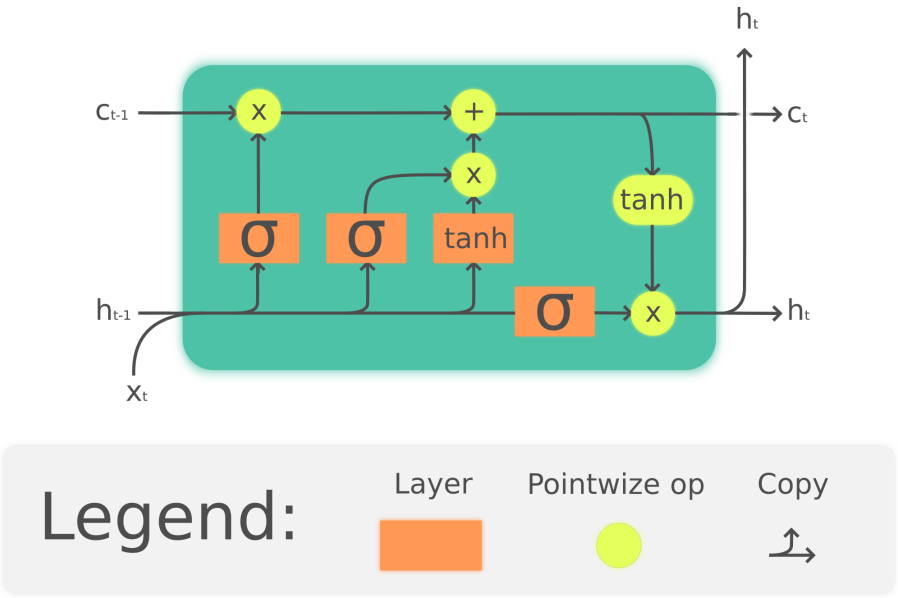


Fig. 2.3: The state cell of an LSTM model. (commons.wikimedia.org/wiki/File:The_LSTM_cell.png)

The main contrasting terminologies for attention are **hard** vs. **soft**, and **global** vs. **local**. In soft attention models, hidden states are all considered, weighted using an additional layer. In hard attention models, only a limited subset of hidden states are considered, the selection of hidden states may be chosen by the model at each stage or consist of a single sliding window of attention. Soft attention models are straightforward to apply backpropagation to, since each hidden state has a differentiable relationship to the output, while hard attention models that select which hidden states are used at a particular time step are not. A **monotonic** hard attention model is one in which the window of attention moves through the input at the same rate that the output is generated, which is applicable in scenarios when corresponding positions in input and output are expected to be strictly related. **Global** vs. **local** attention refers to whether all or only a narrow range of hidden states contribute to a hard attentional layer, as depicted in Figs. 2.8 and 2.9 (Luong et al., 2015).

2.3 Transfer learning

Transfer learning is a high-level term for any machine learning process for which either the domain or the distribution of some or all training data inputs is different

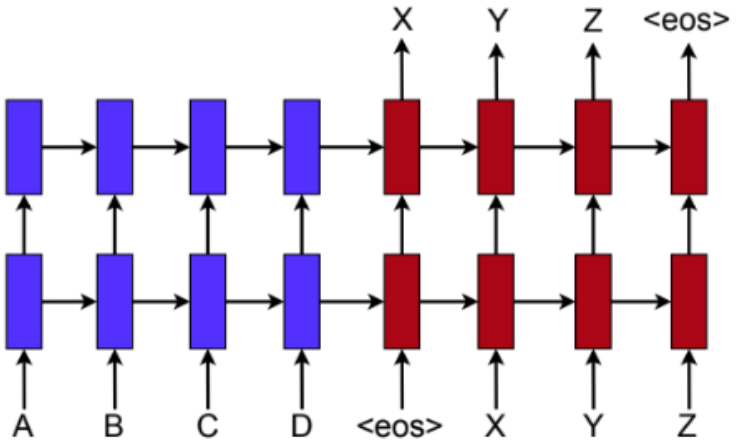


Fig. 2.4: A stacked RNN architecture, in which there are two layers of hidden state. In this architecture, hidden states in the second layer are exclusively derived from the aligned cell of the first layer. (Luong et al., 2015)

from that of the inputs to which a model will ultimately be applied. Such techniques arise in response to the conundrum that many real-world machine learning problems face: a lack of data that looks like the system that one wishes to predict or understand, either due to insufficient available quantity, or altogether lack in the case of seeking to predict the outcome of future events of which no past equivalents exist. The machine learning problem for which a model is initially trained is called the **source task**; the problem which the model is being constructed to solve is called the **target task**. Typically, the source model is not the only component of the eventual target model - manual output transformations or additional machine learning is usually applied (Pan and Yang, 2010).

Some examples are provided by Pan and Yang 2010. One is the problem of seeking to classify web documents by topic. If a significant portion of the documents that will need to be classified bear on topics rarely covered in an existing corpus of web documents, this is an example where the training and test domains are roughly similar but their distribution is different; it is perhaps a less obvious transfer learning problem. Another example is that of attempting to gauge sentiment of reviews of cameras, when the only available data is a corpus of reviews of other types of products. In this case, knowledge of many domains, none of which is equal to the target domain, must be transferred to attempt to make predictions about the target domain.

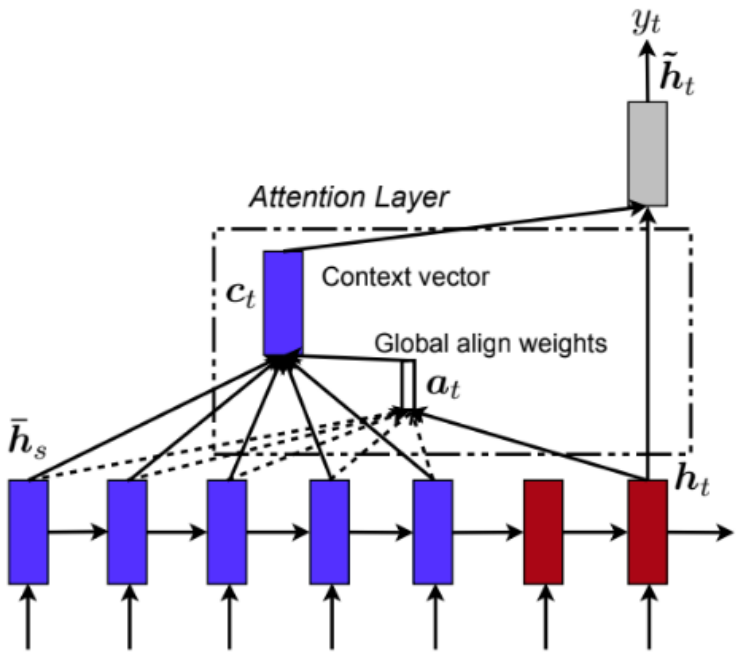


Fig. 2.5: An RNN architecture with global attention: all first-layer hidden states are used to construct an output. (Luong et al., 2015)

There are three broad categories of transfer learning that Pan and Yang identify, which differ in their intrinsic difficulty. **Inductive transfer learning** describes a modeling scenario in which the domains of source and target are identical, but the target task is different, that is, the codomain of the model differs. In inductive transfer learning, output from the source model is used to inform a model of the target task. **Transductive transfer learning** describes the scenario that the task is the same but the source and target domains are in some way different, either differing in feature space, or having the same feature space but differing in distribution over that space. The two examples above are both instances of transductive transfer learning; the web document categorization problem is an example of differing distribution while the review sentiment analysis problem is an example of differing feature space. **Unsupervised transfer learning** describes the scenario that source and target differ in both domain and codomain.

An example of unsupervised transfer learning is the morphology transfer learning problem that this thesis undertakes: labeled training inputs and outputs from the source (one language) are used to inform a model of the target task (a different language), and the feature spaces of both domain (lexemes and morphological

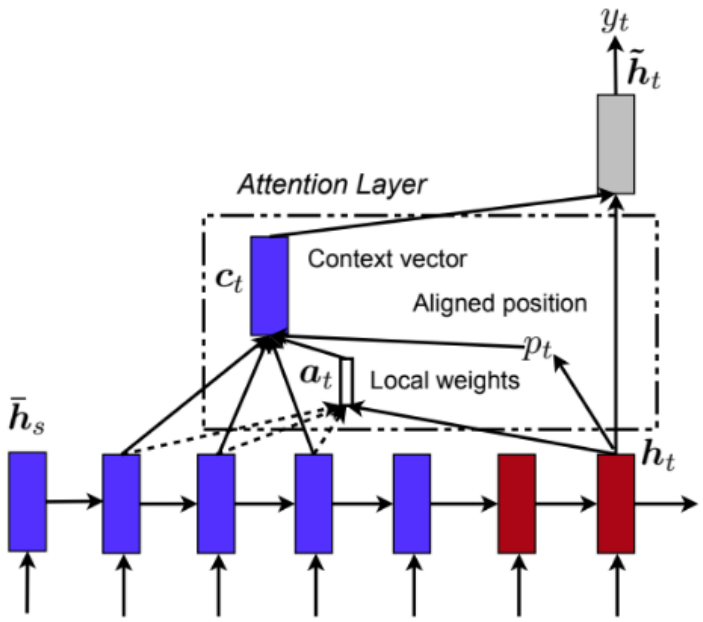


Fig. 2.6: An RNN architecture with local attention: only a subset of hidden states, not necessarily contiguous, are used to construct an output. (Luong et al., 2015)

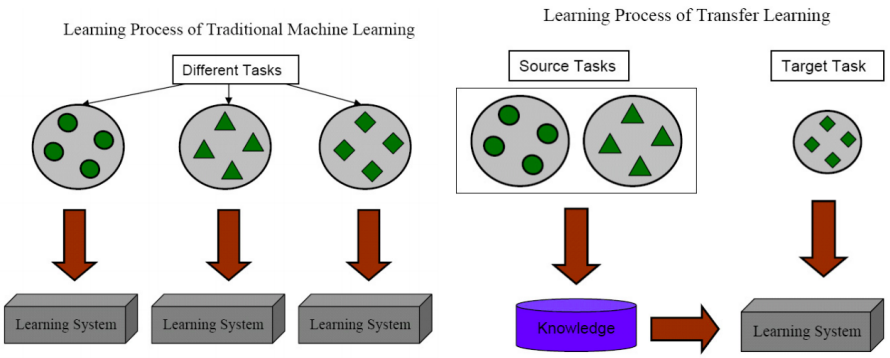


Fig. 2.7: An abstract characterization of traditional ML vs. transfer learning

categories) and codomain (inflected forms) differ between the two. After all, no two languages share the same set of words, and only in the case of very closely related languages or extreme coincidence will all grammatical paradigms inflect for the same set of morphological categories (one language is likely to have a grammatical gender, a verb tense, or some other category that the other language lacks).

There is precedent for transfer learning on LSTMs for language technology tasks, typically with the explicit goal of dealing with low data volume in the target task due to sparse language resources. The baseline for SIGMORPHON 2019 was based on the LSTM transfer learning architecture introduced in Zoph et al. 2016 (McCarthy et al., 2019), which was applied there to a transductive machine translation task.

Their approach is relatively straightforward - they use an LSTM encoder-decoder model to train a machine translation system from French to English on a high data volume, then use that model as the initialization of an architecturally identical Uzbek-English model, holding the decoder fixed and simply allowing the encoder to learn encodings for Uzbek with quite a small dataset. Their results showed considerable improvements over similarly low-data machine translation techniques (Zoph et al., 2016).

3 Machine learning of morphology: existing work

3.1 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.

3.1.1 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 transform English verbs from present to past tense, e.g., *show* → *showed*, *see* → *saw*, etc. (Dreyer et al., 2008).

The natural extension of this is aiming to be able to predict any inflected form given one specific form of a lexeme and an arbitrary set of morphological categories. For instance, given a lexeme *see* and the categories `3rd person singular, simple present`, generating the correct form *sees*. Generally speaking, a citation form has been used as input (Durrett and DeNero, 2013, Faruqui et al., 2015, Cotterell, Kirov, Sylak-Glassman, et al., 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. For instance, a model may be given the forms *sees* and *seeing* among its training data, and be required to fill out the remaining forms of that paradigm, including *see* and *saw*.

If a model is instead trained using entire inflection tables, e.g., all forms of the verb *see*, 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, Kirov, Sylak-Glassman, et al., 2017).

3.1.2 Inflection types

Overall, a diverse set of inflection shapes have been worked with in the most recent efforts of this subfield. Since 2016, the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON), a research collective focused on computational morphology and related problems, has fielded "shared tasks" in which several research teams globally are given training data and a task definition, and attempt to create models which are subsequently compared. The SIGMORPHON shared task 2018 included training data from 103 typologically diverse languages, and paradigms using suffixing, prefixing, infixing, reduplication, and non-concatenative morphology.

However, all SIGMORPHON work has been with well-defined, tabular paradigms of grammatical inflection 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. What Mattissen, 2004 calls "compositionally polysynthetic" languages, those which make extensive use of 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). Such languages include Chukchi, Cherokee, Ainu, Nahuatl, and others (Mattissen, 2004).

The chief challenges with approaching these less paradigmatic types of inflection are data procurement and structuring. For grammatical inflection paradigms, the set of potential forms is simply combinatorial from the set of grammatical categories, and can be specifically enumerated. Concretely, the vast majority of grammatical paradigm data used in SIGMORPHON comes from Wiktionary, a free and collaborative online multilingual dictionary which provides full or partial tables of grammatical inflection along with lexeme definitions; these tables define the space of inflected forms for the task. In contrast, with derivational morphology, noun incorporation, and other less paradigmatic types of inflection, the possibility space is considerably less well-defined. How can a model know a priori that *likeable* is a valid word of English but *hateable* is not, or that the opposite of *accessible* is *inaccessible* and not *unaccessable*? There are no tables for the set of, e.g., derivational forms that a lexeme may take on, since it is constrained by semantics and idiosyncratic usage, and there is considerably more variability among lexemes of the same category (say, verbs) in which derivational categories they may take on than which grammatical categories. Consequently, labeled data is much harder to come by, and it is harder to demonstrate that a dataset of derivational forms is truly exhaustive.

3.1.3 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, a type of test 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. Evidence suggests that learning curve varies by model type; LSTM neural models, a type of model generally considered to be state of the art, despite being the most successful models with high-data training sets, often fare worse than more baseline string transduction models with low data training sets (Cotterell, Kirov, Sylak-Glassman, et al., 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).

3.2 Non-neural approaches

3.2.1 Vector embedding

A technique that has found success in a variety of computational linguistics tasks is that of representing words in relatively low-dimensional vector spaces. That is, words are represented as a vector, a series of numbers of fixed length; the length of the vector is typically much smaller than the number of total known words. This

has the intention of capturing semantic and syntactic content in a principled way - similarities between the numbers representing two words are expected to signify actual similarity in their meaning, and regular linear transformations between vectors should roughly correspond to specific semantic or grammatical changes. These vector representations can be generated via various unsupervised learning methods (Bilmes and Kirchhoff, 2003, Alexandrescu and Kirchhoff, 2006).

Regularities in the relative location of semantically related words have been exploited for semantic analysis tasks (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 (Mikolov et al., 2013, Soricut and Och, 2015, Dos Santos and Zadrozny, 2014). Figure 2.1 illustrates this idea in a slightly simplified way: once a vector embedding model has been trained on English, it can be discovered that semantic transformations (male to female) and grammatical transformations (singular to plural) roughly correspond to regular spatial translations in vector space.

Vector embedding has the limitation that it cannot extend to words for which a vector representation has not been trained, and so it cannot directly provide understanding of the many OOV forms encountered in test data of highly inflected languages (Soricut and Och, 2015, Cotterell and Schütze, 2019). However, it can be a means to discover relationships between words in an unsupervised manner, which may support labeling tasks in support of computational morphology and other tasks.

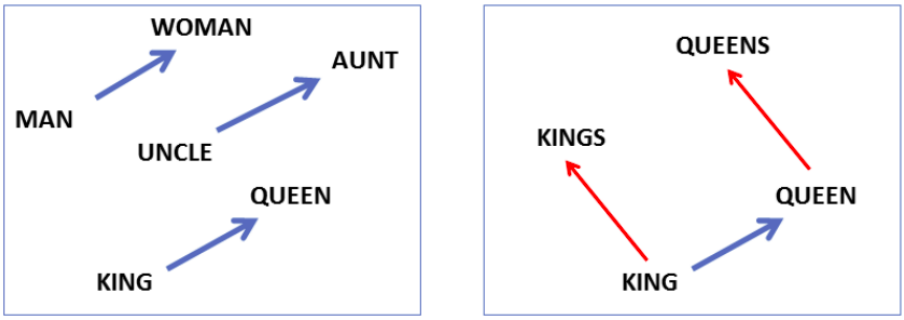


Fig. 3.1: Regular spatial transformations encode semantic or grammatical content (Mikolov et al., 2013)

3.2.2 String transduction

Earlier work specifically focused on the problem of morphology prediction made use of iteratively improving methods of string transduction - in essence, pattern matching on the written representations of words (Durrett and DeNero, 2013, Hulden et al., 2014, Nicolai et al., 2015, Ahlberg et al., 2015). Typical steps of string transduction methods include character alignment (depicted in Figure 2.2), identification of characters that are inserted or deleted based on grammatical form, and generalization of lexemes which are inflected by the same sets of insertions or deletions (depicted in Figure 2.3).

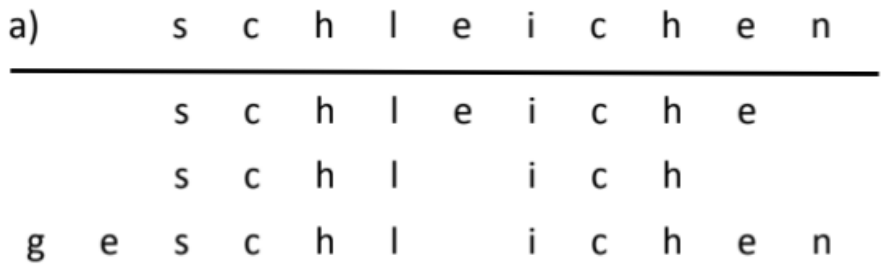


Fig. 3.2: Character alignment for various forms of the German verb *schleichen* (Nicolai et al., 2015).

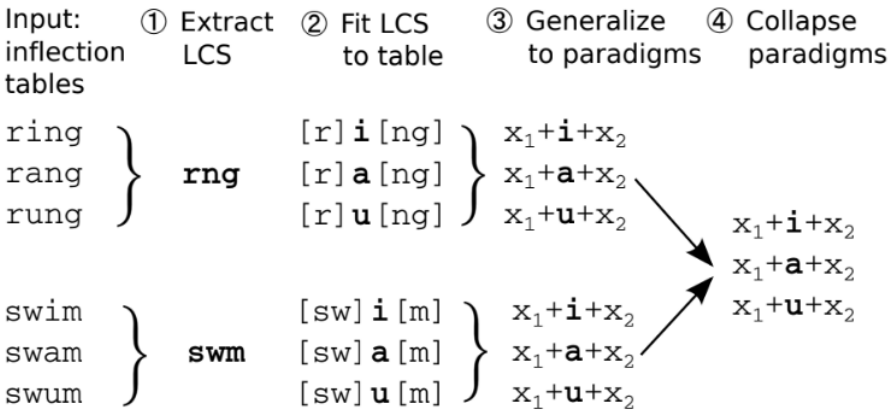


Fig. 3.3: Conceptual depiction of a typical example of a string transduction method of morphology learning (Hulden et al., 2014).

A crucial limitation of string transduction methods are their general assumption that most lexemes have exactly the same set of characterwise transformations as a large group of other lexemes, and that a manageably small number of such inflection classes exist. There are paradigms with such a limited set of inflection classes, such as Spanish *-ar*, *-er*, and *-ir* verbs. However, when multiple morpholonological

processes are at play, individual lexemes may be nearly unique in their exact set of transformations.

For example, Finnish noun declension has processes of vowel harmony, consonant gradation, and vowel alternation and lengthening operating to produce final inflected forms (Ranta, 2008). As an illustration, consider the Finnish nouns *puku* "suit" and *kenkä* "shoe", which have the inessive singular forms *puvussa* "in the suit" and *kengässä* "in the shoe", respectively. In both forms, the letter *k* is transformed via consonant gradation, but the letter it becomes depends on the surrounding letters. The final vowel of the forms may be *a* or *ä*, depending on vowel harmony. In other inflected forms, the final vowel of the words may be doubled (*English Wiktionary* n.d.). A model that naively seeks to match these words with other words using the same set of character transformations across the paradigm may need to assign nearly every word to its own category, failing to generalize the patterns at work.

The poorer performance of string transduction relative to neural models has led to a move of the field away from string transduction since about 2016 (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018).

3.3 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, Kirov, Sylak-Glassman, et al., 2017, Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018, McCarthy et al., 2019). An LSTM network is a variation on a recurrent neural network (RNN), a variant of neural network (Hochreiter and Schmidhuber, 1997).

The SIGMORPHON 2019 transfer learning task used a diverse set of architectures as baselines, including a soft attention, a non-monotonic hard attention, and two monotonic hard attention models, reflecting a diversity of strategies employed by the best current models. Soft attention has dominated prior morphology learning work, but Wu and Cotterell (2019) demonstrate that hard monotonic models may be more

appropriate for morphology tasks, where string transductions are mostly monotonic - that is, (except for in instances of reduplication or metathesis) characters in an input word correspond to characters in the same order in the output (McCarthy et al., 2019, Wu and Cotterell, 2019).

4 Data

4.1 Grammatical inflection data

In order to assess grammatical similarities between languages, the data set published for the SIGMORPHON first shared task 2019, a subset of the data set for the first SIGMORPHON 2018 shared task, was used. Both sets are publicly available on GitHub (McCarthy et al., 2019, Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018).

The first 2018 shared task was simply a morphology learning problem. The data consists of triples - lemma, grammatical categories, and appropriately inflected form - for 103 languages, partitioned into training, development, and test sets. An example of some Finnish triples is shown in Fig. 3.2. The training sets are further partitioned into low, medium, and high-resource sets; the low-resource training sets contain about 100 forms, medium about 1,000 forms, and high about 10,000 forms, with the sets being nested so that the smaller training sets are subsets of the larger ones. These data levels are used to simulate different resource settings, e.g., low-resource training sets represent data available for a poorly-resources language; the data levels can also be used to assess model learning curve (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018).

The 2019 shared task was a transfer learning task, which sought to use a large volume of data about a source language to inform a model with access to a low volume of data for the target language. The data available for the task was a subset of the 2018 data. The task consisted of 100 pairs across 79 languages. For each training pair, some data sets from 2018 were used: the high-resource training data for the source language and the low-resource training set and development and test data sets for the target language (McCarthy et al., 2019).

Inflection of *tukehtua* (Kotus type 52/sanoa, *t-d* gradation)		
indicative mood		
present tense		
person	positive	negative
1st sing.	tukehdun	en tukehdu
2nd sing.	tukehdut	et tukehdu
3rd sing.	tukehtuu	ei tukehdu
1st plur.	tukehdumme	emme tukehdu
2nd plur.	tukehdutte	ette tukehdu
3rd plur.	tukehtuvat	eivät tukehdu
passive	tukehdutaan	ei tukehduta

Fig. 4.1: The English Wiktionary partial inflection table for the Finnish word *tukehtua*.

```
keisarillinen   keisarillisitta ADJ;PRIV;PL
tukehtua        tukehdutaan     V;PASS;PRS;POS;IND
juhtaeläin      juhtaeläimille  N;AT+ALL;PL
vastaava        vastaavatta     ADJ;PRIV;SG
```

Fig. 4.2: A sample of the SIGMORPHON 2018 data for Finnish, scraped from Wiktionary and provided on GitHub (https://github.com/sigmorphon/conll2018/blob/master/task1/all/finnish-train-high).

4.1.1 Language diversity

The languages represented in the 2018 data cover a wide range of families and typological categories. Although over half of the languages are from the Indo-European family, a grouping that includes most languages of Europe, Greater Iran, and the northern part of the Indian subcontinent, one or more languages each of the Athabaskan, Bantu, Causasian, Kartvelian, Quechua, Semitic, Sino-Tibetan, Turkic, and Uralic families, as well as two isolates (Haida and Basque), are represented. Diverse inflection strategies are also represented, including suffixing, prefixing, infixing, ablaut, and introflexion, and long-distance processes like vowel harmony and consonant harmony (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018).

Among the 2019 pairs, all the aforementioned families except Athabaskan, Kartvelian, and Sino-Tibetan are represented (McCarthy et al., 2019). The pairs are not chosen randomly from among the 2018 languages; many languages are chosen more often

as source than target languages and vice versa, and the Turkic language family in particular was exhaustively paired. I conjecture that many seemingly difficult-to-explain trends in the data are related to confounding factors in the choices of language pairs, and in the conclusion section I will discuss what some of these effects may be and how future work might overcome these confounds.

4.1.2 Sourcing and sampling

The English Wiktionary, a collaborative online dictionary, has become something of a standard source of supervised morphological data. It provides full or partial inflection tables alongside lexeme definitions; the structure of tables is consistent for a given language and part of speech. An example table is given in figure 3.1. For some highly inflected languages (e.g., Navajo), Wiktionary only provides a fixed subset of forms. For some relationships between words that could be considered grammatical, it may simply offer them as separate lexical entries; for example, Russian perfect and imperfect forms are given as separate entries, as are Navajo verb forms that vary by aspect or thematic classifier (*English Wiktionary* n.d.).

For most of the languages in the SIGMORPHON 2018 data, forms were gathered via scraping from Wiktionary. Multiple parts of speech are represented for most languages, but only parts of speech with a significant number of entries relative to all entries in a given language. Inflected forms were sampled for inclusion according to their estimated distribution in the text of Wikipedia for each respective language. For languages with sufficient data, 12,000 forms were sampled, and from these 1,000 were randomly selected for the development set and 1,000 for the testing set; the remaining 10,000 became the high-resource training set, of which 1,000 were randomly chosen as the medium-resource set and 100 of those as the low-resource set. For languages with less available data, sets might be smaller and the high-resource training set might be omitted; 17 languages lack high-resource sets (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018).

4.1.3 Representation of morphology

The SIGMORPHON data set uses the UniMorph format to indicate grammatical categories (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018). The UniMorph project, initially published in 2015, is a format for encoding morphological categories uniformly cross-linguistically. It uses a universal label set to encode morphological categories across languages. An example of the annotation can be seen in Fig. 3.2, where words are marked for, e.g., part of speech (`ADJ, V, N`), voice (PASS), tense (`PRS`), number (`SG, PL`), and other categories (Sylak-Glassman, Kirov, Post, et al., 2015, Sylak-Glassman, Kirov, Yarowsky, et al., 2015, Sylak-Glassman, 2016). These annotations were key in generating important metrics of language pairs for this study: category overlap for each part of speech, and part of speech distribution overlap, both discussed in section 3.2.3.

4.2 Language typology data

Data about language typology was drawn from the World Atlas of Language Structures (WALS), Ethnologue, and generated from the UniMorph tags in the SIGMORPHON 2019 data.

4.2.1 Typology data from WALS

The World Atlas of Language Structures is "a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors." It contains information about twelve morphology features, such as "Reduplication," "Prefixing vs. Suffixing in Inflectional Morphology," and "Inflectional Synthesis of the Verb" (*The World Atlas of Language Structures Online* n.d.). WALS labels for twelve morphological features were scraped and collected for all languages for which they are currently available. Among the 79 languages present in the SIGMORPHON 2019 transfer pairs, as many as 47 had labels available for a particular feature ("Prefixing vs. Suffixing in Inflectional Morphology"), while as few as 17 had labels available for

other features ("Fusion of Selected Inflectional Formatives", "Exponence of Selected Inflectional Formatives", "Exponence of Tense-Aspect-Mood Inflection"). The full set of scraped data can be found on the GitHub repository for this paper; explanations of the categories can be found on WALS.

Abkhaz	Productive full and partial reduplication	Hewitt 1979: 265
Aghul	No productive reduplication	
Agta (Central)	Productive full and partial reduplication	Healey 1960
Ainu	Full reduplication only	Refsing 1986

Fig. 4.3: A sample of WALS data on reduplication, available at https://wals.info/feature/27A, accessed 19 Nov 2019.

4.2.2 Genealogy from Ethnologue

SIGMORPHON 2018 supplied a basic "Family" designation for each of the 103 languages in its data set, but these do not correspond to any particular taxonomic level; some, such as Indo-European and Uralic, are language families, the most broad designation of language genealogy, while others, like Semitic, Slavic, and Romance, are subfamilies of various size. Language genealogy data from Ethnologue was used to supplement the SIGMORPHON 2018 labels and provide a more fine-grained measure of distance of relation between languages; my designations can be found on the GitHub repository for this paper (*Ethnologue* n.d.).

4.2.3 Part of speech category sets

As a simple measure of language structure, the SIGMORPHON 2019 data was scraped for UniMorph tags to identify the total set of morphological tags present for each language and each part of speech. For instance, the set of tags found on German nouns is `ACC;DAT;GEN;NOM;PL;SG`. This measure should not be taken as a set of all inflectional categories actually used in a particular language; the SIGMORPHON data can skew toward particular parts of speech or particular inflection types. The generated category sets were simply used as a rudimentary measure of structural similarity between languages. For instance, if a language A marks nouns for case but not definiteness, then another language B that marks nouns for case as well is in

some sense more similar to language A than a language C that does not mark case but does mark definiteness on nouns.

5 Methods and Results

5.1 Assessing the impact of transfer learning

There are no results available from SIGMORPHON that directly measure the impact of transfer learning while controlling for other model parameters.

SIGMORPHON 2018 task 1 used a string transduction model as a baseline, and received a variety of submitted models which all used a neural component but many of which took advantage of ensembling with string transduction methods. It did not explicitly make transfer learning data available, though one team did use multilingual training data to build its model (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018).

The goal of the SIGMORPHON 2019 task 1 was the same as SIGMORPHON 2018 - morphological inflection - but it was specifically a transfer learning task. For the source language, a high-resource data set was provided, and for the target language a low-resource data set; these data sets are identical in size and format to the 2018 high- and low-resource training sets, but they were resampled and so don't contain exactly the same data. SIGMORPHON 2019 provided four baseline models for task 1, all of which were purely neural and differed by attention mechanism, and all submissions were neural models (McCarthy et al., 2019).

Despite the many confounding factors between the 2018 and 2019 best model performances, some suggestive results are yielded by comparing the best performance of any submission on each pair in 2019 to the best performance of any 2018 submission on the comparable task of modeling the pair's target language in a low-resource setting. Barring differences in modeling technique and data sampling, the two problems are essentially equivalent - using about 100 examples of a language to

predict any other inflected form from that language - with the major difference being the transfer learning data, about 10,000 examples of a different language, that was made available in 2019.

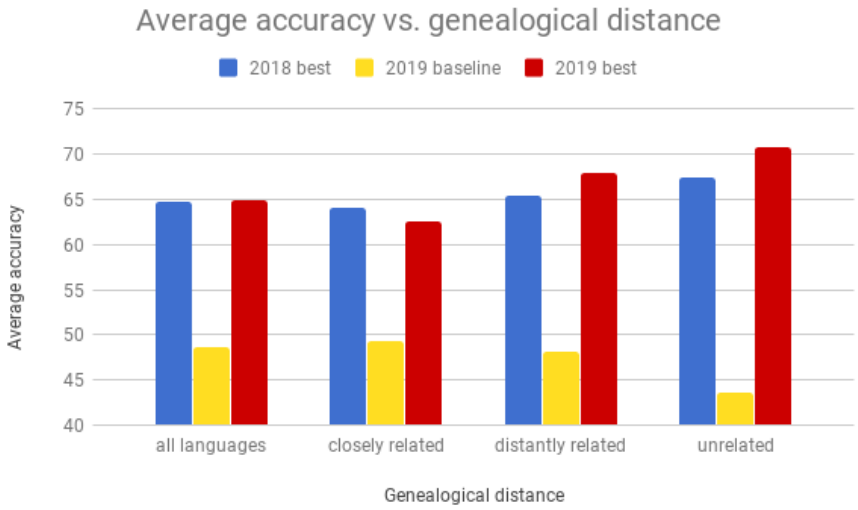
I looked at rudimentary typological, genealogical, and sampling metrics of the languages and training pairs to find correlations with model improvements between 2018 and 2019. The methods of generation of this data, and statistical analysis of it, are presented in this section; potential interpretations and implications are discussed in the next section.

5.2 Genealogical distance

SIGMORPHON 2019 stated that 20 of its 100 language pairs were distantly or unrelated, based on the language "family" designations in SIGMORPHON 2018. I used the Ethnologue language genealogy designations to generate a three-tiered distance designation: "Closely related" (in the same language family and subfamily), "Distantly related" (in different subfamilies of the same family), and "Unrelated" (in different families). By my designation, there are 62 closely related pairs, 31 distantly related pairs, and 7 unrelated pairs.

I conjecture that most distantly related pairs have dissimilar enough forms as to behave much more like unrelated languages with perhaps coincidental structural similarities than like closely related languages. However, my designations are not a fine-grained or necessarily consistent metric of actual similarity; some language families are much more internally diverse than others, and subfamily divisions were chosen somewhat arbitrarily.

On average, more distantly related language pairs actually perform better and show more improvement between 2018 and 2019, perhaps suggesting that transfer learning between closely related languages confuses the model, although this result is not statistically significant.



I conducted all statistical tests on both the set of all language pairs and the set of all distantly related and unrelated language pairs, but no statistically significant results were achieved over the set of all language pairs; perhaps because genealogical relationship is too strong a confounding variable.

5.3 Part of speech category overlap

As discussed in the data section, the UniMorph annotations in the training data were scraped to discover the full set of morphological categories for which each part of speech could be inflected in each language. These category sets were used to generate a metric of structural similarity I call category overlap.

5.3.1 Calculating overlap

For each part speech in each of two languages, given the category sets $C_{POS,language\ A}$ and $C_{POS,language\ B}$, the overlap was calculated as

$$overlap(POS, language\ A, language\ B) = \frac{|C_{POS,language\ A} \cap C_{POS,language\ B}|}{|C_{POS,language\ A} \cup C_{POS,language\ B}|}$$

Take as an example the category overlap of nouns in German and Greek. The German category set for nouns is $C_{N,German} = \{ACC, DAT, GEN, NOM, PL, SG\}$ - German inflects nouns for singular and plural number and four grammatical cases. The Greek nominal category set is similar, except that it lacks a dative case and has a vocative case: $C_{N,Greek} = \{ACC, VOC, GEN, NOM, PL, SG\}$. The nominal category overlap is

$$
\begin{aligned}
&overlap(N, German, Greek) \\
&= \frac{|\{ACC,DAT,GEN,NOM,PL,SG\} \cap \{ACC,VOC,GEN,NOM,PL,SG\}|}{|\{ACC,DAT,GEN,NOM,PL,SG\} \cup \{ACC,VOC,GEN,NOM,PL,SG\}|} \\
&= \frac{|\{ACC,GEN,NOM,PL,SG\}|}{|\{ACC,DAT,VOC,GEN,NOM,PL,SG\}|} \\
&= \frac{5}{7} \approx .71
\end{aligned}
$$

It is not uncommon for one or both languages in a pair to lack any morphological categories for a particular part of speech; many languages do not have training data for all parts of speech. If only one language has an empty tagset for a part of speech, then by the above formula the category overlap is 0. If both languages have empty tagsets, the above formula would yield $\frac{0}{0}$, not a real number; in such a case the pair is excluded from analysis, which is why n is less than the total number of language pairs in the findings below.

5.3.2 Correlation with model performance

Overall, verbal category overlap was much more predictive than nominal category overlap of performance changes between 2018 and 2019. In particular, the verbal category overlap of a transfer learning pair had a highly significant ($p<.01$) relationship with year over year accuracy improvement, with an additional .1 verbal category overlap score predicting a 3.7% jump in absolute model accuracy between the two years. I used randomized permutation testing to approximate the significance level of the correlation.

Many of the pairs concentrated in the lower left of the above plot are pairs of Turkic languages, raising the possibility that language pair sampling is confounding the

outcome. However, the same result is present and statistically significant when all Turkic pairs are removed.

The relationship of nominal category overlap to model improvement did not rise to any significance threshold.

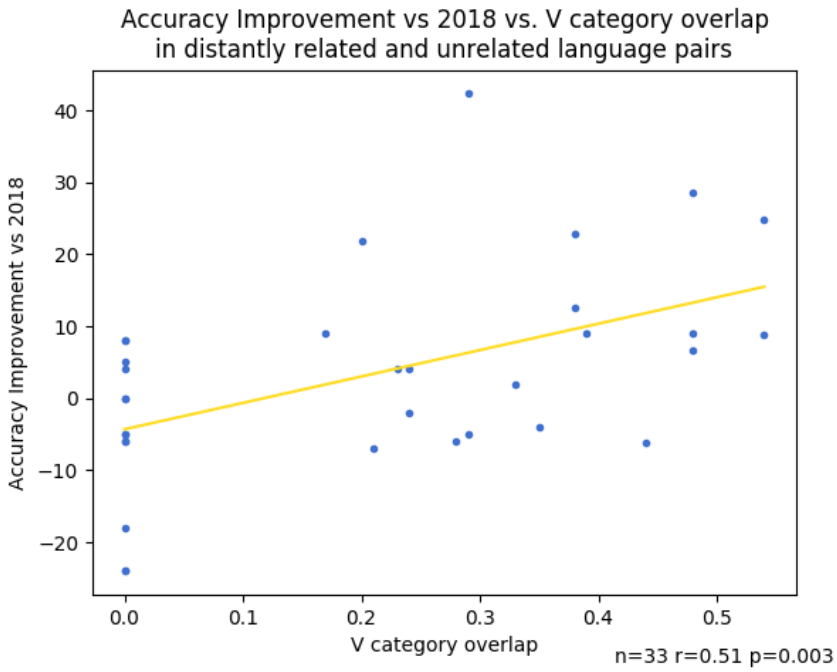


Fig. 5.1: The relationship between a pair's verbal category overlap and the best team's performance in 2019 relative to 2018.

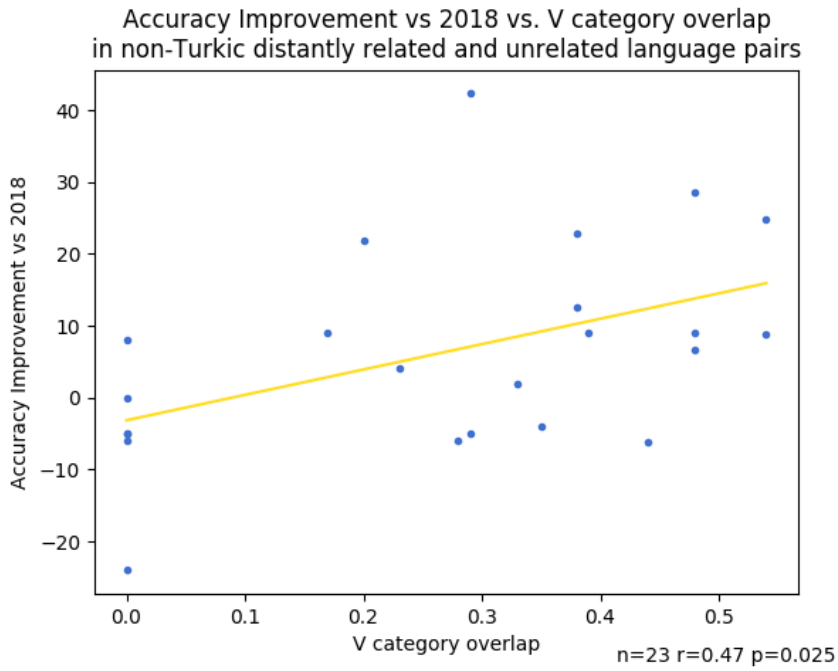


Fig. 5.2: The relationship between a non-Turkic pair's verbal category overlap and the best team's performance in 2019 relative to 2018.

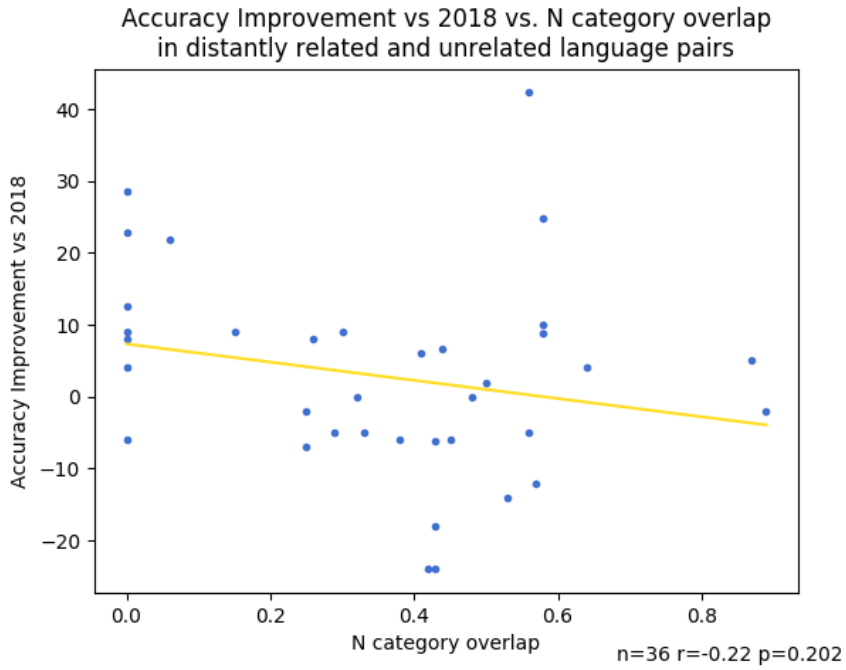


Fig. 5.3: The relationship between a pair's nominal category overlap and the best team's performance in 2019 relative to 2018.

5.4 Part of speech distribution similarity

The UniMorph tags identify four broad parts of speech cross-linguistically in the SIGMORPHON data: nouns, verbs, adjectives, and determiners. However, there is only one language among the 79 in the SIGMORPHON 2019 data that has all of these parts of speech represented; 64 languages have verb data, 55 have noun data, 40 have adjective data, and only 2 have determiner data.

I use similarity between the relative distributions of parts of speech in source and target training sets as a metric of data set similarity. Part of speech distribution in the SIGMORPHON data is not necessarily an indication of actual linguistic typology; while some languages lack inflection on some parts of speech, there are also omissions in the SIGMORPHON data due to data sparsity (Cotterell, Kirov, Sylak-Glassman, Walther, et al., 2018).

5.4.1 Calculating similarity

My part of speech distribution similarity statistic is simply the statistical distance between the part of speech distributions of the two languages, calculated by summing the differences of the proportions of each part of speech between the two languages. That is, if $f_{POS,language}$ is the number of training forms for a given language and part of speech and $f_{language}$ is the total number of training forms for a language, the part of speech distribution similarity between language A and language B is

$$POSDS(language\ A, language\ B) = \sum_{POS} \left| \frac{f_{POS,language\ A}}{f_{language\ A}} - \frac{f_{POS,language\ B}}{f_{language\ B}} \right|$$

where $POS = \{N, V, ADJ, DET\}$.

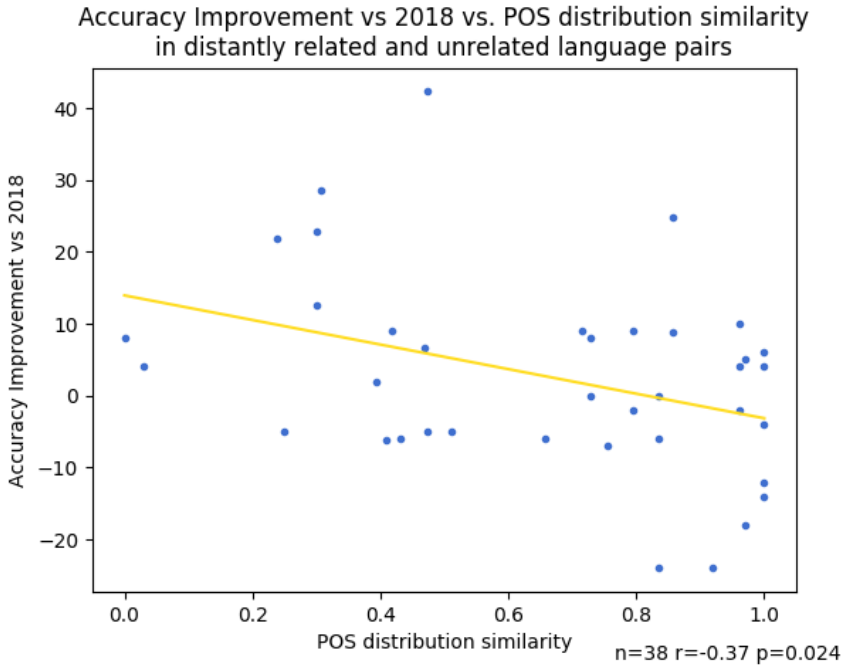


Fig. 5.4: The relationship between a pair's part of speech distribution similarity and the best team's performance in 2019 relative to 2018.

5.4.2 Correlation with model performance

Another significant ($p<.05$) negative relationship was discovered between part of speech distribution similarity and model improvement between 2018 and 2019. That is, the more that the data sets of source and target language for a given pair shared the same parts of speech, the worse a transfer learning model could be expected to perform relative to a 2018 non-transfer model. If models over transfer learning pairs with similar part of speech distributions actually performed worse overall, that would constitute evidence that a similar source and target domain somehow confused or worsened the model, and that transfer learning was thus counterproductive. However, there is no evidence of any relationship between data set similarity and *overall* model performance:

Since transfer learning pairs with similar part of speech distributions could be modeled as well as other pairs in 2019, but showed less *improvement* from transfer learning, the target languages of those pairs must have been more effectively modeled in 2018. Since transfer learning was not an available strategy in 2018, and only target language data was available, similarities between source and target

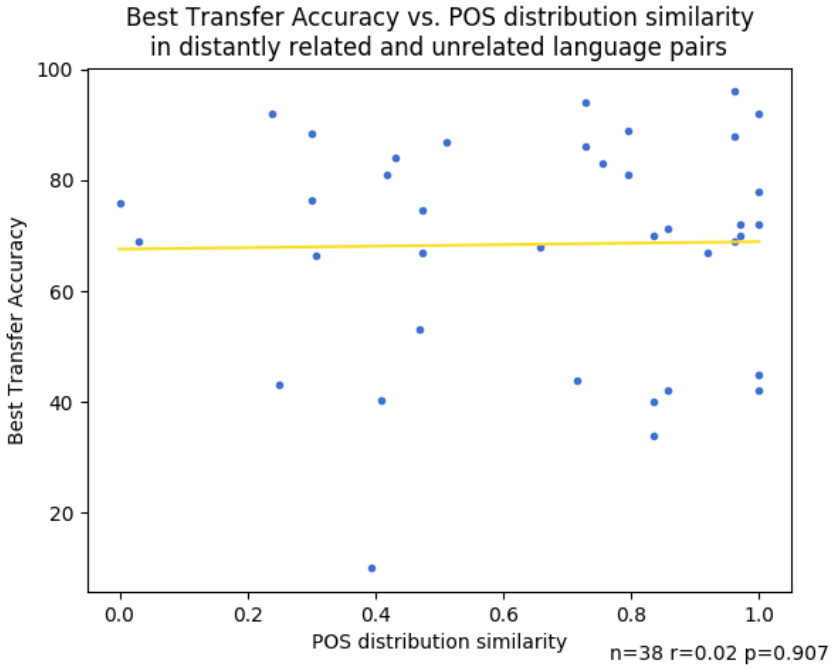


Fig. 5.5: The relationship between a pair's part of speech distribution similarity and the best team's performance in 2019.

data cannot directly explain this better performance in 2018. There must be some confounding factor accounted for the by selection of transfer learning pairs.

6 Conclusions and Discussion

6.1 Overall model accuracy disparities between languages

One trend that plainly jumps out in the data from SIGMORPHON 2018 and 2019 is that some languages are persistently more difficult to model than others.

In SIGMORPHON 2018 task 1 best achieved scores, the SIGMORPHON 2019 task 1 baseline models, and the SIGMORPHON 2019 best achieved scores, models of Old Irish always have the lowest accuracy rate, never higher than 10%, typically followed in second or third place by Latin.

On the other end of the spectrum, the Turkic languages included in the SIGMORPHON data (Azeri, Bashkir, Crimean Tatar, Kazakh, Khakas, Tatar, Turkish, Turkmen, Uzbek) have among the highest-accuracy models. In the SIGMORPHON 2018 task 1 with a low data setting, 6 of 9 featured Turkic languages achieved accuracies above 85%, and the Turkic languages averaged 80% accuracy while the overall mean accuracy was 62%. In SIGMORPHON 2019 task 1, language pairs with a Turkic target language averaged 74% baseline accuracy while the overall was 49%, 81% best accuracy where the average was 65%. Curiously, the Turkic languages actually had slightly worse models on average in SIGMORPHON 2019 than 2018, while overall, the accuracy of best models for a given language held constant on average. (The reason that the average best score for Turkic languages was 81% in 2019 compared to 80% for 2018 is that higher-accuracy Turkic languages happened to be chosen more often as target languages for transfer pairs in 2019. Best model accuracy for individual Turkic languages decreased by an average of 3% from 2018 to 2019.)

To some degree, these outcomes correspond with probably expected intuitions or general perceptions of the relevant languages. Turkic languages, for instance, are known for straightforward agglutinative morphology and almost entirely lacking declension classes and irregular forms; for a given grammatical meaning, there is typically a single suffix that is applied to all words (Johanson, 1998).

6.1.1 CMU-03

There are two language pairs that stand out as seeming outliers in the SIGMORPHON 2019 data, in different ways: Bengali → Greek and Swahili → Quechua.

In 2018, the University of Zurich team achieved 32% accuracy in modeling Greek with a low volume data, the best team on that problem. Greek → Bengali and Bengali → Greek pairs appeared in SIGMORPHON 2019. The Greek → Bengali pair scored fairly typically: the best 2019 model was slightly better than the 2019 baselines and slightly worse than the best 2018 low-resource Greek model. But the Bengali → Greek scored only 18% accuracy from the best 2019 baseline model, yet the Carnegie Mellon team was able to achieve 75% accuracy on that pair, by far the largest difference between a 2019 baseline and best team performance, and the second-highest absolute improvement between 2018 and 2019. The case of Swahili → Quechua is even more drastic: while the best 2019 baseline could only score 14% accuracy, the same Carnegie Mellon model achieved 92% accuracy, the single largest disparity between a 2019 baseline and best score.

That model, denoted `CMU-03` in McCarthy et al., 2019, had the highest overall accuracy of all submitted models, and was the best performer on 61 of the 100 language pairs, but in particular performed comparatively well on pairs that had done poorly in 2018 or with the 2019 baseline models. McCarthy et al., 2019 et. al. also note that the success of `CMU-03` is correlated with linguistic similarity of the two languages, though these examples in particular do not provide evidence for that idea. Bengali → Greek is a quite distantly related pair, with the two languages belonging to different primary branches of the large and diverse Indo-European family, and Swahili → Quechua are completely unrelated. Bengali → Greek has fairly typical part of speech distribution overlap and nominal and verbal category

overlap, while Swahili → Quechua scores substantially below average on all three similarity metrics.

CMU-03 is unique in that it employs techniques to attend over morphological tags before ingesting the input lemma, and to bias toward character copying. The effect of these techniques is not clear, but provides an interesting area of future research. Unfortunately, there is no separately published paper or codebase for CMU-03.

6.2 Overall outlook for transfer learning as a morphology learning strategy

McCarthy et al., 2019 states that "gains from cross-lingual training were generally modest, with gains positively correlating with the linguistic similarity of the two languages." The claim that using transferred knowledge boosts model performance, if slightly, seems to come from individual reports from teams about their models. Unfortunately, code or full results for the individual 2019 SIGMORPHON task 1 submissions were not published, so to verify the claim, a comparison of models that differ only in their use of transfer knowledge must be conducted. According to comparison of 2018 and 2019 best models, however, average model performance with or without access to transfer learning was essentially the same, while best model performance varied widely between different pairs.

Given the similarity of data and goals between SIGMORPHON 2018 and 2019, it may be surprising that, for 50 of the 100 training pairs, higher accuracy was achieved on the target language in low-resource settings by the best model of 2018 than by any submission in 2019. It might have been expected that with access to additional data, as well as information from the previous year about which models were successful, models in 2019 should have avoided declining performance. The reason for the performance declines may be that in 2018, the models that performed best in low-resource settings were actually quite different than those that scored well in high-resource settings. Models adapted to low-resource settings tended to avoid pure reliance on neural encoder-decoder models and used techniques such as using

string transduction to learn edit sequences and string alignments, and biasing toward copying characters. With the task refocused on neural transfer learning in 2019, many of these techniques might have been dropped. Since all team performances on SIGMORPHON 2019 task 1 were not published, it is not possible to assess the impact of particular model parameters on performance with different types of languages.

It also may be the case that, since training sets were resampled from 2018 to 2019, the languages that saw worsened performance in 2019 simply had by random chance less useful training sets that year. The low-resource training sets in 2018 and 2019 were relatively small with at most 100 examples, making for greater likelihood of chance sampling effects.

As shown in section 5, from comparison of 2018 and 2019 data, it is far from clear that genealogical relationship between languages leads to effective transfer learning, while verbal category overlap does seem to be significantly predictive of better transfer learning. The statistically significant negative relationship between part of speech distribution similarity and model performance, as well as the suggestive but not statistically significant negative relationship between genealogical closeness and model performance, suggests that transfer knowledge may actually be capable of *confusing* a model.

6.3 Potential language pair sampling confounds

6.3.1 Language pair sampling

Language pairs for SIGMORPHON 2019 were not selected at random from an even distribution. Many languages appear only as target languages because they lack high-volume data, and some of the languages from the SIGMORPHON 2018 data appear in as many as 5 pairs while others were not included. Some groups of languages, such as Germanic and Turkic groups, appear to have been near-exhaustively paired between a subset of source and a subset of target languages, while the number of fully unrelated pairs is just 7. Genealogical relationship between a language pair appears to be a substantial confound - many results appear statistically significant

among non-closely related languages, while no statistically significant inferences could be drawn about the entire pool of language pairs.

6.3.2 Turkic languages

Particular language families also differ from one another, both in internal diversity and in typological trends. In particular, the Turkic family stood out as an outlier in the SIGMORPHON 2019 data. As noted in 6.1, Turkic languages share a set of morphological characteristics that make them relatively easy to model accurately, as evidenced by the higher-than-average accuracy of models over the Turkic languages. The Turkic language family may also be an instance of a family with less internal diversity than, say, the Indo-European family.

	Singular
Nominative	**kitap**
Definite accusative	kitabı
Dative	kitaba
Locative	kitapta
Ablative	kitaptan
Genitive	kitabın

Fig. 6.1: Nominal declension in Turkmen, an Oghuz Turkic language (Wiktionary).

Declension of *kitap*	[hide ▲]
nominative	kitap
genitive	kitapnıñ
dative	kitapqa
accusative	kitapnı
locative	kitapta
ablative	kitaptan

Fig. 6.2: Nominal declension in Crimean Tatar, a Kipchak Turkic language (Wiktionary).

For instance, the above declension tables for the word *kitap* "book" show that three languages from different primary branches of the Turkic family all share the same set of noun cases, with cognate suffixes.

Contrast this with inflection of the respective words for "book" in Swedish and Czech, two Indo-European languages of different subfamilies.

declension of كتاب	
	singular بىرلىك (birlik)
nominative باش (bash)	كتاب (kitab)
genitive ئىگىلىك (igilik)	كتابنىڭ (kitabning)
dative يۆبىلىش (yöbilish)	كتابغا (kitabgha)
accusative چۈشۈم (chüshüm)	كتابنى (kitabni)
locative ئورۇن (orun)	كتابدا (kitabda)
ablative چىقىش (chiqish)	كتابدىن (kitabdin)

Fig. 6.3: Nominal declension in Uyghur, a Karluk Turkic language (Wiktionary).

	Singular		Plural	
	Indefinite	Definite	Indefinite	Definite
Nominative	bok	boken	böcker	böckerna
Genitive	boks	bokens	böckers	böckernas

Fig. 6.4: Nominal declension in Swedish, a Germanic Indo-European language (Wiktionary).

Declension [hide ▲]		
	singular	plural
nominative	kniha	knihy
genitive	knihy	knih
dative	knize	knihám
accusative	knihu	knihy
vocative	kniho	knihy
locative	knize	knihách
instrumental	knihou	knihami

Fig. 6.5: Nominal declension in Czech, a Slavic Indo-European language (Wiktionary).

Given that removing closely related languages from consideration was necessary to generate any statistically significant results, if Turkic languages are designated as "Distantly related" while being quite similar, this might obscure otherwise significant results. And indeed, that seems to have occurred in at least one place: a statistically significant ($p<.05$) relationship emerges between genealogical distance and model performance. If all language pairs with Turkic target languages are removed from the data set, and genealogical distance is mapped onto a continuous space from 0 (different language family) to 1 (different subfamilies of the same language family) to 2 (same family and subfamily), the previously non-significant negative relationship between genealogical closeness and model performance appears to be statistically confirmed. This result implies that transferred knowledge about a genealogically related source language actually somehow hinders or confuses a model of a target language, contrary to the conclusions of McCarthy et al., 2019.

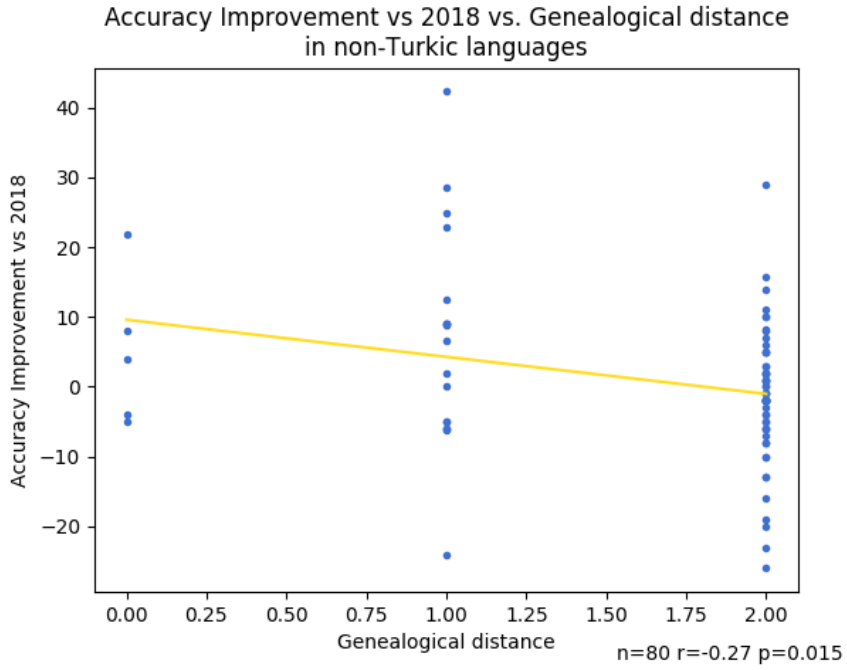


Fig. 6.6: Language pair genealogical similarity has a significant negative relationship with model performance once Turkic languages are removed from the data set.

7 Future Work

7.1 Language Features

The typological features I'm interested in considering include inflection shape (prefixing, suffixing, infixing, introflexion), set of inflected categories and overall paradigm size by part of speech, degree of fusion, and presence of long-distance phonological processes.

7.1.1 Morphological typology features

Inflection shape may be important in determining, for instance, how soft attention models choose to focus on various parts of a word, so that a model trained on a language with a particular profile of inflection shapes may be more likely to attend to the correct parts of words when readapted to model a language with similar inflection shapes.

Fusion refers to the marking of multiple categories at once with a single indivisible morpheme, e.g., the Spanish verb *hablé* "I spoke" marks tense and subject person and number with the suffix *-é*, lacking separable morphemes to indicate the preterite tense and first person singular subject (cf. *hablo* "I speak", *habló* "he/she spoke"). Transfer learning may be beneficial between languages that both exhibit or lack fusion, or more specifically between languages that exhibit fusion between the same categories; since fusion operates over combinations of two or more categories, though, the space of possible fusion behaviors is quite large and it may be difficult to find unrelated languages with similar fusion behavior.

Long-distance phonological processes are those by which the surface form of an affix depends on the phonological properties of a segment at some distance from the

affixation site; consider my earlier example of vowel harmony in Finnish: the nouns *puku* "suit" and *kenkä* "shoe" have the inessive singular forms *puvussa* "in the suit" and *kengässä* "in the shoe", respectively, with final vowel of the suffix dependent on the set of vowels in the rest of the word. Given that long-distance processes present considerable difficulty for non-neural morphology models, neural models may have to be intensively trained to attend to such processes, presenting an opportunity for transfer learning to leverage existing knowledge.

7.1.2 Language relatedness metrics

I'd also like to take into account a more fine-grained measure of language relatedness, so that incidental typological similarities can be successfully statistically blocked against similarities arising from common origin. Two possible strategies are using detailed language genealogical trees and counting degrees of separation between languages, or finding some way to assess lexical similarity, the proportion of words between two languages which have both similar forms and meanings due to shared origin. Lexical similarity may be a more relevant measure - if word stems are similar between two languages, it is likely that grammatical affixes are as well. However, lexical similarity may also be due to shared areal loanwords, such as the profusion of Arabic loanwords into Turkish, Persian, and Urdu. Areal effects can also cause grammatical similarity to spontaneously arise between geographically collocated languages (Ponti et al., 2018), though grammatical similarities due to shared ancestry are probably stronger. Ultimately, lexical similarity and genealogical closeness measure different types of linguistic relationships, and both should be taken into account if good data is obtainable.

Consistent and quite fine-grained information about language genealogy can be found on Ethnologue, a database of basic typological and socioliguistic information about all recognized languages (*Ethnologue* n.d.). Finding good data about lexical similarity looks to be significantly more challenging - certainly, no database exists of pairwise lexical similarity between all languages, and automatically calculating lexical similarity from the SIGMORPHON 2018 data would require semantic or translation information about the words, to identify cognate words with corresponding

meanings between languages. Such information may be obtainable via the Google Cloud Translation API. An easier mode of estimation may be via the "Translations" tab of entries on English Wiktionary, which provides translations of a word into a potentially large number of other languages. Average Levenshtein distance between translations for entries into a pair of languages might be a good indicator of overall lexical distance - such a method could be attempted on pairs of languages with known lexical distance to assess its utility.

Some of these may also be measurable via analysis of the SIGMORPHON data as well, e.g., "Exponence of Tense-Aspect-Mood Inflection," "Prefixing vs. Suffixing in Inflectional Morphology," and "Case Syncretism," while others, such as "Reduplication" and "Locus of Marking in the Clause" would probably be harder to measure in such a way. For the features which can also be generated by looking at the SIGMORPHON data, the WALS data can at least be used as a gold standard to calibrate and assess the quality of generated metrics. WALS will not be useful in assessing overlap between sets of inflectional categories which are present or exhibit fusion in languages - its categorical tagging is simply not granular enough - but fortunately these will be relatively straightforward measures to generate from the SIGMORPHON 2018 data.

To generate the other metrics, string alignment and transduction methods adopted from pre-neural morphology models will be necessary. Most clear is that to assess inflection shape, I will need to perform character alignment as in Figure 2.2 and count how many string changes take place before, after, or within the stem.

To measure fusion between a pair of grammatical categories, average Levenshtein or LCS distance can be taken between forms that differ along both categories and compared to distance between forms that only differ along one category. For example, recall the Spanish verbs *hablé* "I spoke", *hablo* "I speak", *habló* "he/she spoke". *Hablé* differs from each of the other forms along one category - it has a different tense from *hablo* and a different subject person than *habló* - and its LCS distance from each is 2. *Hablo* and *habló* differ in both tense and subject person but also have a LCS distance of 2; the fact that forms that differ along both categories are no more dissimilar than forms that differ along one is an indicator of the fact that tense and

subject agreement are fused in Spanish verbs. In contrast, consider the Finnish verbs *puhuin* "I spoke", *puhun* "I speak", and *puhui* "he/she spoke" - *puhuin* has an LCS distance of 1 from both other forms while they have an LCS distance of 2 from one another, indicating that in this case Finnish does not fuse tense and subject marking - past tense is constructed with a suffix *-i* and first person subject with a subsequent suffix *-n*.

Presence of long-distance phonological processes will be the most difficult to measure by analyzing the SIGMORPHON data. Fortunately, the presence of vowel and consonant harmony is typically quite binary and pervasive throughout a language's morphology, so rather than attempting to generate a measurement of it I may simply hand-annotate languages with a binary indication of whether or not they possess some long-distance process, and perhaps secondarily with a more specific indication of the type (e.g., frontness vowel harmony, sibilant harmony, etc.).

7.2 Transfer learning experiments

7.2.1 Pair selection

7.2.2 Neural model changes and comparisons

Bibliography

Ahlberg, Malin, Markus Forsberg, and Mans Hulden (Jan. 2015). „Paradigm classification in supervised learning of morphology“. In: pp. 1024–1029 (cit. on pp. 13, 17).

Alexandrescu, Andrei and Katrin Kirchhoff (Jan. 2006). „Factored Neural Language Models.“ In: (cit. on p. 16).

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. 16).

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, 13–15, 18, 20–23, 26, 32).

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, et al. (Aug. 2016). „The SIGMORPHON 2016 Shared Task—Morphological Reinflection“. In: (cit. on pp. 2, 6, 13, 14, 18).

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, et al. (2017). „CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages“. In: *CoRR* abs/1706.09031. arXiv: 1706.09031 (cit. on pp. 1, 2, 6, 12, 13, 15, 18).

Cotterell, Ryan and Hinrich Schütze (2019). „Morphological Word Embeddings“. In: *CoRR* abs/1907.02423. arXiv: 1907.02423 (cit. on pp. 6, 16).

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. 16).

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 pp. 1, 12).

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. 12, 17).

English Wiktionary (n.d.). https://en.wiktionary.org. Accessed: 2019-10-25 (cit. on pp. 18, 22).

Ethnologue (n.d.). https://www.ethnologue.com/. Accessed: 2019-11-17 (cit. on pp. 24, 43).

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. 2, 12, 18).

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). „Long Short-Term Memory“. In: *Neural Comput.* 9.8, pp. 1735–1780 (cit. on pp. 5, 18).

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. 13, 17).

Johanson, Lars (1998). „The Structure of Turkic“. In: *The Turkic Languages*. Routledge (cit. on p. 36).

Kibrik, Aleksandr E (1994). „Archi“. In: *The indigenous languages of the Caucasus* 4.part 2, pp. 297–365 (cit. on p. 2).

Luong, Thang, Hieu Pham, and Christopher D. Manning (Sept. 2015). „Effective Approaches to Attention-based Neural Machine Translation“. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1412–1421 (cit. on pp. 7–10).

Mattissen, Johanna (2004). „A structural typology of polysynthesis“. In: *WORD* 55.2, pp. 189–216. eprint: https://doi.org/10.1080/00437956.2004.11432546 (cit. on p. 14).

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, 10, 15, 18–21, 26, 36, 37, 41).

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 p. 16).

Nicolai, Garrett, Colin Cherry, and Grzegorz Kondrak (Jan. 2015). „Inflection Generation as Discriminative String Transduction“. In: pp. 922–931 (cit. on p. 17).

Pan, S. J. and Q. Yang (Oct. 2010). „A Survey on Transfer Learning“. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10, pp. 1345–1359 (cit. on p. 8).

Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, et al. (2018). „Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing“. In: *CoRR* abs/1807.00914. arXiv: 1807.00914 (cit. on p. 43).

Ranta, Aarne (2008). „How predictable is Finnish morphology? an experiment on lexicon construction“. In: *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, pp. 130–148 (cit. on p. 18).

Soricut, Radu and Franz Och (Jan. 2015). „Unsupervised Morphology Induction Using Word Embeddings“. In: pp. 1627–1637 (cit. on p. 16).

Sylak-Glassman, John (2016). „The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema)“. In: (cit. on p. 23).

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. 23).

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. 23).

The World Atlas of Language Structures Online (n.d.). https://wals.info/. Accessed: 2019-11-17 (cit. on p. 23).

Wu, Shijie and Ryan Cotterell (July 2019). „Exact Hard Monotonic Attention for Character-Level Transduction“. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1530–1537 (cit. on p. 19).

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight (Nov. 2016). „Transfer Learning for Low-Resource Neural Machine Translation“. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1568–1575 (cit. on p. 11).