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

January 4, 2021

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 language data, with the intended goal of boosting modeling resources for low-resource languages. It was found that transfer learning could aid the development of computational models for low resource languages and that transfer learning was most effective between genealogically related languages. This study expounds on those findings by testing a much larger number of unrelated language pairs, systematically comparing two model architectures, and examining relationships between model performance and selected linguistic typological similarities of source and target languages. It was found that transfer learning can still afford substantial benefits when source and target language are unrelated, and that transfer learning is most beneficial when source and target language have similar sets of morphologically inflected categories and similar patterns of fusion between those categories, while similarities in inflection shape are not predictive of transfer learning efficacy. This information can be used to select source languages when leveraging transfer learning to improve computational resources for low-resource target languages, especially those without closely related high-resource languages.

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 Spanish adjective *pequeño* "large":

	singular	plural
masculine	*pequeño*	*pequeños*
feminine	*pequeña*	*pequeñas*

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 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 than in some other languages.

In a more morphologically complex language, a text may contain many inflected forms that are individually rarer but nonetheless perfectly intelligible to a person or model that understands the morphology. For example, a person learning Spanish may have never encountered the form *pequeñas* before, but if they have seen all three other forms of the word and are familiar with Spanish adjectival morphology, they will have no difficulty in fully understanding the meaning of the word. If language models can behave similarly rather than treating a new word like *pequeñas* as OOV - that is, fundamentally unknown - their understanding of new material in morphologically complex languages may be substantially enhanced (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 a model type called **long short-term memory (LSTM) neural networks**, briefly described in 2.2, operating over individual characters and grammatical category tags. Their use for computational morphology has been pioneered by SIGMORPHON, a research group that holds annual *shared tasks*, competitions among several research teams on a morphology prediction problem. LSTMs definitively overtook the field after their strong performance relative to other model types in the SIGMORPHON 2016 shared task. However, learning curve analysis in the SIGMORPHON 2017 task showed that LSTMs perform well in high-data settings but, provided with lower volumes of training data, actually fare worse than simpler model types trained on similar amounts of data (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** - a means of using the output of several models at once - with other methods (Cotterell, Kirov, Sylak-Glassman, Walther, et al. 2018). For languages with low quantities of digital resources, though, there is still much room for improvement of computational morphology models.

One of the SIGMORPHON 2019 shared tasks focused on **transfer learning**, leveraging 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. They claim that this use of data about other languages produced "modest" model performance gains, with data transferred from related languages being more conducive to strengthening models (McCarthy et al. 2019). This finding produces one potentially useful strategy for leveraging transfer learning to improve modeling of a low-resource language. Unfortunately, there are plenty of low-resource languages not closely related to other high-resource languages. It may be worthwhile to assess what other linguistic properties of a language may predict its usefulness for improving modeling of other languages.

This study seeks to conduct a more in-depth investigation of the relationships between linguistic typological properties and ultimate transfer learning outcomes for low-resource languages. The computational model distributed as a baseline in SIGMORPHON 2019 was modified to fulfill this task, and 242 pairs of languages were tested as transfer learning pairs, with target languages artificially limited to smaller training data sets to mimic a low availability of structured training data. Performance (as measured by overall accuracy and Levenshtein distance) of these transfer learned models was compared to performance of models lacking a transfer learning component, and absolute differences in performance metrics were taken to represent the efficacy of transfer learning. Alongside this modeling, multiple types of typology information was computationally generated about each language in the study. Finally, statistical relationships were explored between transfer learning efficacy and similarities in these typological characteristics between source and target languages, to understand which types of typological information should be taken into account when designing transfer learning models for low-resource languages.

2 Theoretical background

2.1 Natural language morphology and typology

2.1.1 What is natural language morphology?

One way that human languages express grammatical or semantic meaning is by altering individual words; this process is called **morphology**. Linguistic morphology is split into two major categories: **inflectional** and **derivational**. Inflectional morphology comprises alterations to words that reflect *grammatical* meaning, that is, it indicates which of a determined subset of grammatical categories is at work in a particular phrase. For example, the verb *paint* is grammatically inflected in the sentence "*He paints.*" to indicate agreement with a third-person singular subject. Derivational morphology applies *semantic* alterations, that is, it shifts the fundamental meaning of a word, often switching its part of speech. For example, the verb *paint* can become the noun *painter*, denoting a person characterized by performing the action of the verb (Hogan 2010).

The concept of **part of speech** refers to high-level groupings of words within a language according to what inflectional morphology they may undergo. Common parts of speech include nouns, verbs, and adjectives, although the exact division of parts of speech is language-specific. Within a given language and part of speech, there is typically a well-defined, finite set of **forms** that a given word may take; that abstract set of forms is called a **paradigm**. In particular, there is typically a set of **morphological categories** associated with a given part of speech in a given language. Common examples of morphological categories include number (e.g., singular vs. plural), verb tense, and noun or pronoun case (English exhibits case only in pronouns: I, me, my). Inflection is then used to assign a value to some or all of the morphological categories for a given word (Hogan 2010). For example, Spanish adjectives such as *pequeño* "small" inflect for gender and number:

	singular	plural
masculine	*pequeño*	*pequeños*
feminine	*pequeña*	*pequeñas*

The term "word" can be ambiguous when speaking of linguistic morphology. To be more specific, a **form** is a word with a specific spelling or pronunciation and corresponding to a specific set of grammatical values, e.g., a single cell in the above table. The set of forms across an entire inflectional paradigm are said to belong to the same **lexeme**. A lexeme can be identified by its **lemma** - a citation or dictionary form, a particular cell in the paradigm that is chosen to represent the entire lexeme. Spanish adjectives are usually cited in the masculine singular, so the above forms may be said to belong to the Spanish lexeme that has the lemma *pequeño* (Hogan 2010).

In contrast, derivational morphology is applied in a less systematic manner, and constraints on its application are as often semantic or idiosyncratic as determined by part of speech or other formal categories. For example, the English verb *like* may become the common derived term *likeable*, but an analogous term *hateable* is much less often seen. For this reason, derivational morphology is more difficult to systematically study; this paper is only concerned with inflectional morphology.

An important concept in morphology is that of **inflection classes** - subgroupings of parts of speech in a given language that exhibit similar patterns of inflection to one another. An example is the division of Spanish verbs into -*ar*, -*er*, and -*ir* classes:

infinitive	3rd person sg. imperfect indicative	gerund
bailar	*bailaba*	*bailando*
comer	*comía*	*comiendo*
partir	*partía*	*partiendo*

Many types of grammatical categories commonly appear cross-linguistically, such as number, gender, animacy, case, politeness, tense, aspect, mood, and definiteness. The values that these categories may take often differ somewhat in meaning from language to language (for instance, English only has one inflected past tense while Spanish has two), but the parallels between them are typically enough that it's possible to make general characterizations of the degree of similarity of inflection systems between two languages. English and Spanish both inflect nouns for singular and plural number, and neither inflects nouns for case, but Spanish can inflect some human nouns for gender as well, and inflects adjectives for number and gender where English adjectives are not (Hogan 2010). A related notion is that of paradigm size. English nouns only ever take two forms - singular and plural - while Finnish nouns may take about 28 (*English Wiktionary* n.d.).

2.1.2 Types of inflection

There are many **inflection shapes** seen in natural languages. The most basic is **affixation**: the addition of sounds to the beginning, middle, or end of a word. Affixation can be broadly divided into prefixation (addition of an affix to the beginning of a word), suffixation (to the end), infixation (in the middle), circumfixation (at both ends). Other designations for affixation shapes include ablaut (alteration or replacement of a sound in the middle of a word) and templatic morphology. Templatic morphology is a complex process by which certain elements of a stem are distributed throughout a form, interwoven with elements indicating grammatical value.

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").

	1st person singular	3rd person singular
present	*hablo*	*habla*
preterite	*hablé*	*habló*

Reduplication is the repetition of all or part of a word. Reduplication may be full, as in Indonesian *orang* "person" vs. *orang-orang* "people", or partial, as in the formation of the contemplative aspect of the Tagalog word *bili* "buy" by repeating the first syllable: *bibili* (examples from Wiktionary).

Long-distance morphophonological 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. For instance, Finnish exhibits vowel harmony, a process by which the vowels present in affixes depend on the set of vowels in the stem. The Finnish inessive singular suffix *-ssa/-ssä* looks different on the nouns *puku* "suit" and *kenkä* "shoe", with the final vowel of the suffix dependent on the set of vowels in the rest of the word.

nominative singular	inessive singular
puku	*puvussa*
kenkä	*kengässä*

(*English Wiktionary* n.d.)

Lastly, **irregularity** is a concept with familarity to anyone who's ever tried to learn a foreign language. Forms or entire lexemes are said to be irregular if they are aberrant from the normal patterns of morphology. For example, almost all English verbs only mark a third person singular subject differently from others, with a suffix *-s*: *I run* but *she runs*. However, *to be* marks the first person singular subject differently as well, and the third person singular subject marking bears no resemblance to the normal affixation process: *I am*, *you are*, *she is*. It can be said that *be* is a highly irregular verb.

2.1.3 Linguistic typology

Linguistic typology refers to the classification of languages into different categories depending on their structure. The major categories of morphological typology are **isolating**, **agglutinating**, and **fusional**. Languages are considered to be more isolating if they have little grammatical inflection, smaller inflection paradigms, and a smaller number of affixes per word on average (Hogan 2010). English is a fairly isolating language, likely a factor in the lack of computational morphology research (Cotterell and Heigold 2017). Languages are considered to be more agglutinating if they exhibit more extensive inflection but do not typically exhibit fusion; agglutination is also associated with greater morphological regularity. Fusional languages are simply those that exhibit relatively more fusion. Like many typological distinctions, there is no strict assignment of languages into one or another of these categories; rather, they are ends of a spectrum that languages may be placed along (Hogan 2010).

Other typological labels may be applied depending on the set and shape of morphological inflection. Languages may be considered **highly inflecting** or **polysynthetic** if they have very large morphological paradigms. Languages may be typified by the presence of grammatical gender, vowel harmony, or other morphological characteristics. In general, this study is not concerned with defining specific labels for languages so much as computationally measuring the structural similarity of their morphological systems.

2.1.4 Linguistic genealogy

Languages change and diverge over time, and their development tends to follow genealogical patterns analogous to species descent and divergence. They are commonly grouped into **language families**, described with genealogical trees, and genealogical distance can be measured between them.

Language relationships are ascertained via historical comparison, and language families are simply the largest possible groupings that can be supported with certainty by available evidence. As such, they are not alike in historical timescale or internal diversity; more well-studied or attested families, and those with more historical written material, are likely to be larger because there is information available with which to make inferences about language descent in the distant past.

For this reason, a simple measure of genealogical separation cannot easily be come by. Relative location in a language family tree may not be informative, since the

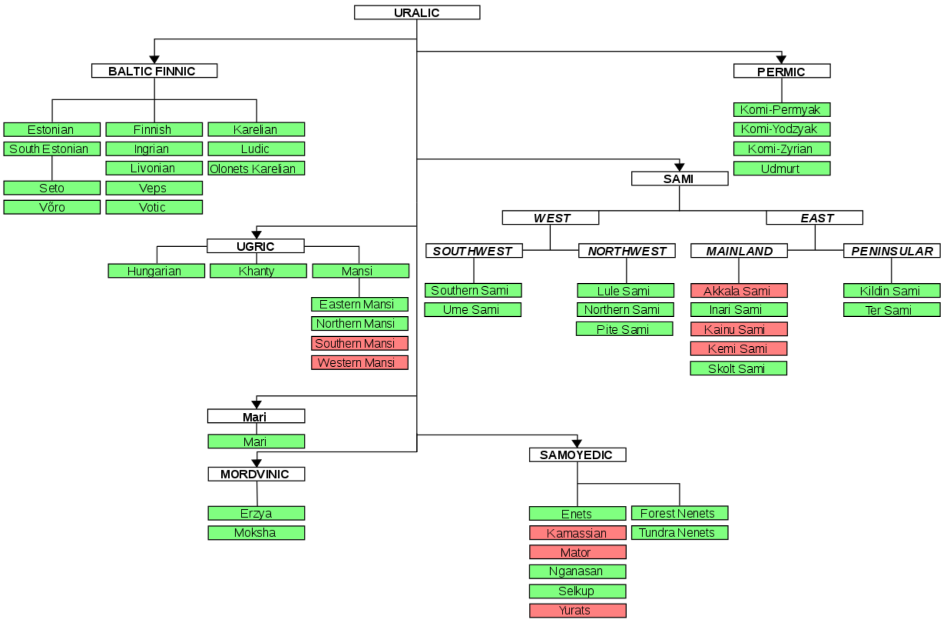


Fig. 2.1: The entire Uralic language family.
(commons.wikimedia.org/wiki/File:UralicTree.svg)

number of members and the density of branching differs between families. Another strategy is to measure **lexical similarity**: the proportion of words between two languages which are obviously similar due to shared origin.

However, lexical similarity is a measure of the similarity of vocabulary items and is not a good proxy for the grammatical similarity of languages. English has borrowed a very large amount of vocabulary from various historical varieties of French, yet their morphological systems remain fairly faithful to their respective origins in different branches of the Indo-European family. French in particular bears substantial morphological similarity to other Western Romance languages like Spanish and Portuguese. Even languages with no known genealogical relationship and extremely different grammatical systems may have non-zero lexical similarity: Turkish and Urdu have both borrowed heavily from Arabic, yet the grammatical systems of the three languages are quite distinct.

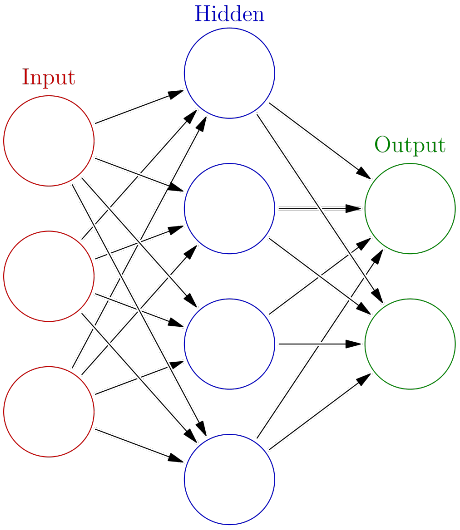


Fig. 2.2: The structure of a feed-forward neural network with three inputs, four hidden cells, and two outputs.
(commons.wikimedia.org/wiki/File:Colored_neural_network.svg)

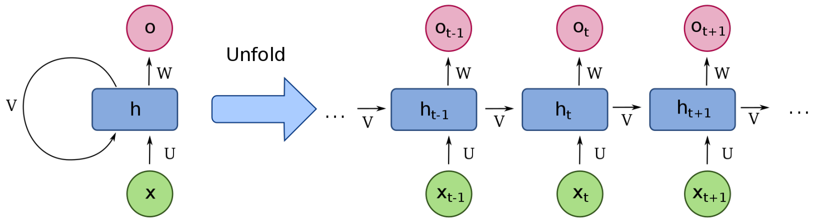


Fig. 2.3: The generalized structure of a recurrent neural network.
(commons.wikimedia.org/wiki/File:Recurrent_neural_network_unfold.svg)

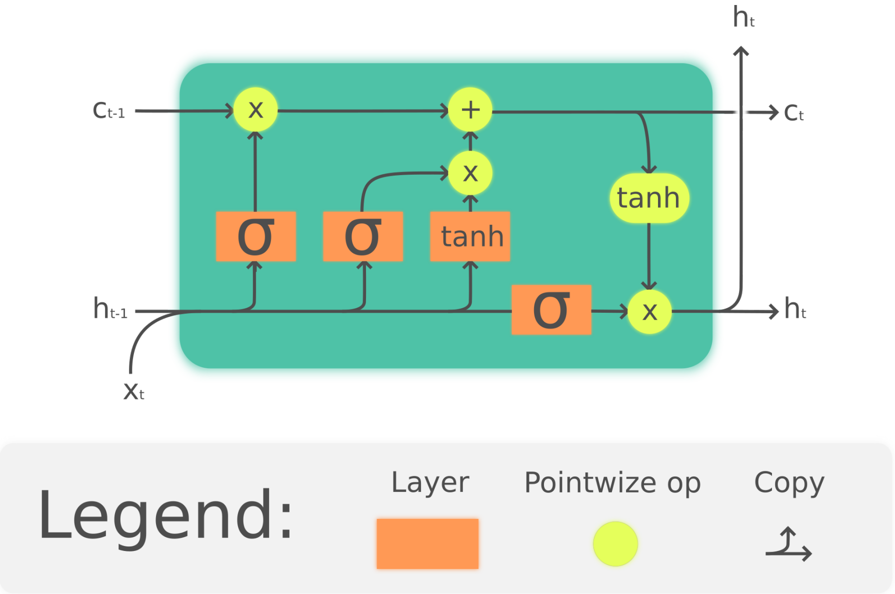


Fig. 2.4: The state cell of an LSTM model.
(commons.wikimedia.org/wiki/File:The_LSTM_cell.png)

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. Fig. 2.2 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.

A **recurrent neural network (RNN)** is a type of neural network that operates over sequences of inputs, typically with unknown length. Fig. 2.3 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.

Long short-term memory (LSTM) neural networks are a variant of RNNs which use a more complex sequence of computations to update state, and maintain a separate piece of state *c* that controls rate of forgetting. LSTMs are intended to solve the forgetfulness of more basic RNN types, which tend to be unable to recall information from more than a few iterations prior (Hochreiter and Schmidhuber 1997). LSTM architecture is described in Fig. 2.4: the input x_i and the previous state h_{i-1} are linearly transformed, added, and activated, as in a simple RNN, before interacting via a series of operations with the previous cell c_{i-1}, producing the next state h_i and cell c_i. **Gated recurrent units** (GRU) are slightly simplified models that do not include separate parameterization of the output gate (not pictured).

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).

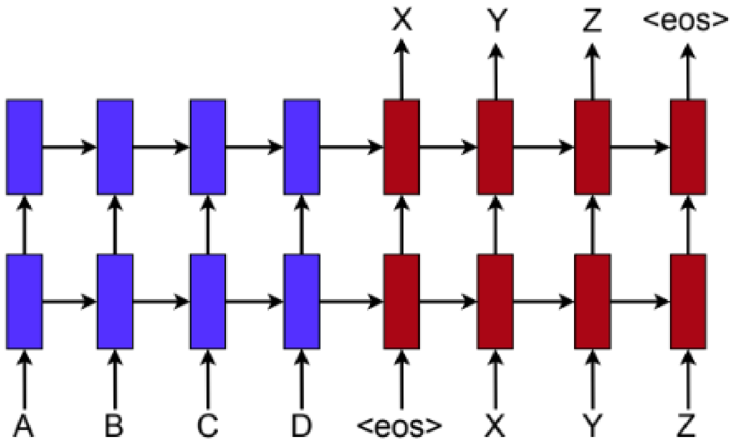


Fig. 2.5: A stacked RNN architecture, in which there are two layers of hidden state. In this architecture, hidden states in the second layer are exclusively derived from the aligned cell of the first layer. (Luong et al. 2015)

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).

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. 2.6 and 2.7 (Luong et al. 2015).

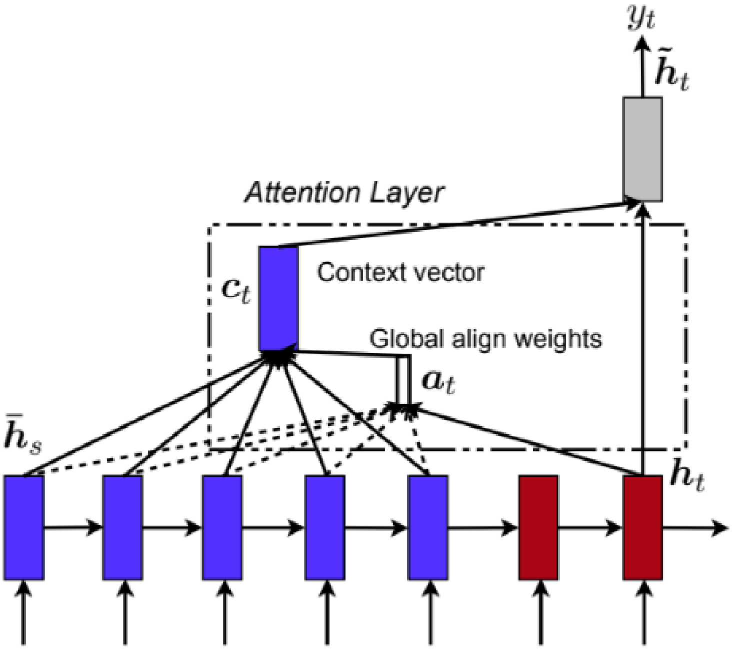


Fig. 2.6: An RNN architecture with global attention: all first-layer hidden states are used to construct an output. (Luong et al. 2015)

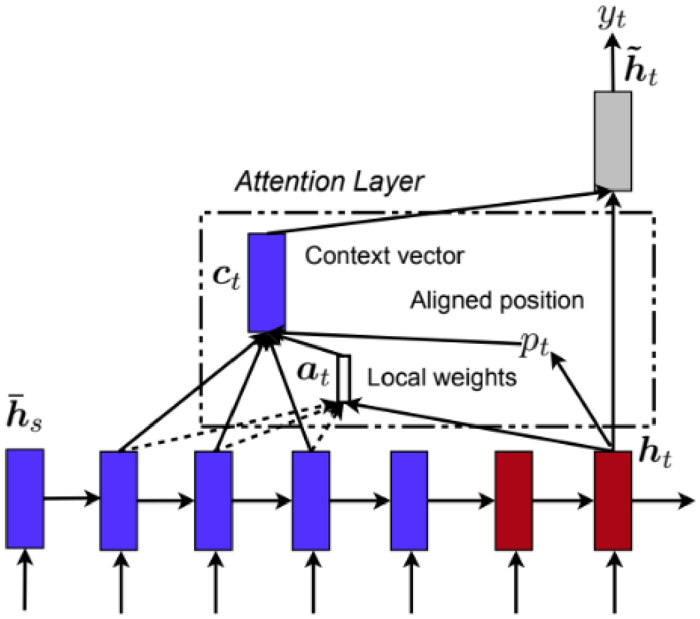


Fig. 2.7: An RNN architecture with local attention: only a subset of hidden states, not necessarily contiguous, are used to construct an output. (Luong et al. 2015)

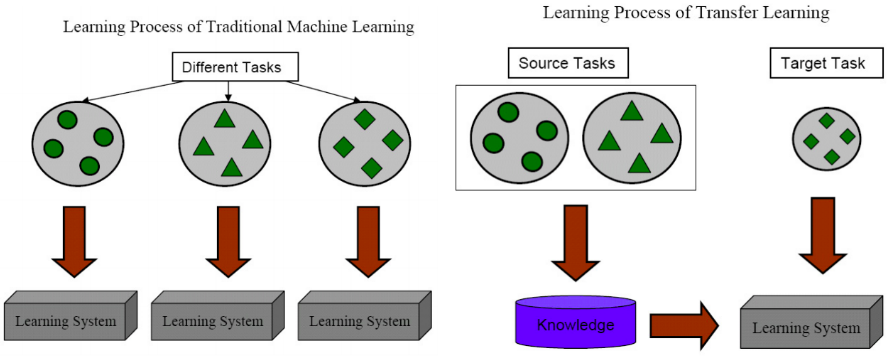


Fig. 2.8: An abstract characterization of traditional ML vs. transfer learning

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 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.

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 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), suggesting that transfer learning may be a promising strategy for computational linguistics tasks.

2.4 Ensembling

Ensembling is another core ML concept that arises in explaining the differences in structure and performance of computational morphology models. It is a generic term for combining the outputs of multiple models which operate over the same domain and codomain. That is, ensembled models receive the same input, and produce outputs of the same form; their outputs are typically combined by vote in classification tasks or weighted or unweighted averaging in regression tasks. Ensembling among neural and non-neural models has been used to improve outcomes in a variety of tasks (Krogh and Vedelsby 1995).

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 inflection morphology of words in isolation - transforming a word into a specific morphological form, or the inverse, tagging a form with its morphological categories.

In this section, I used the machine learning terms **supervised** and **unsupervised**. Supervised learning is that conducted with labeled input-output sets, in which a model attempts to produce outputs similar to those it's seen. Unsupervised learning is that which attempts to find patterns in data sets with no output, such as finding clusters in a scatter plot of points. Note that this definition of **unsupervised** differs from the sense specific to transfer learning. The transfer learning task in SIGMORPHON 2019 is an example of an *unsupervised* transfer learning problem in that the source domain and codomain are both different from target domain and codomain, yet it is *supervised* in a general ML sense in that the triples included in the training data include output values which the model attempts to mimic.

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 (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.

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. LSTMs are generally the most accurate morphology models with high-data training sets and are considered state of the art. However, LSTMs and related neural model types often fare worse than more baseline string transduction models with small training sets, likely due to the very gradual gradient descent process used to fine-tune them (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 were said to have generally performed better than transfer learning models between more distant languages (McCarthy et al. 2019), although that conclusion is examined more closely in this study.

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).

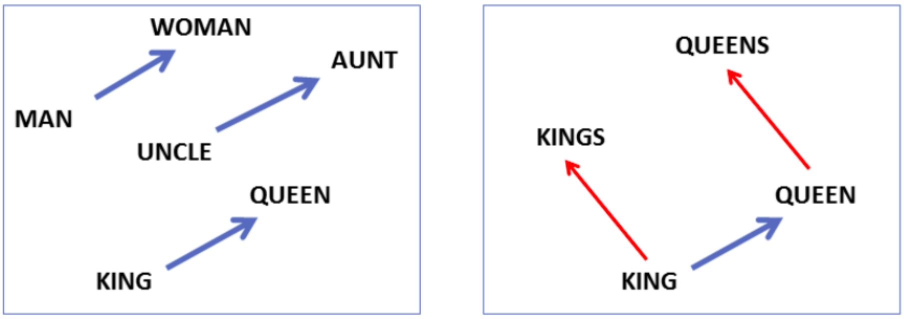


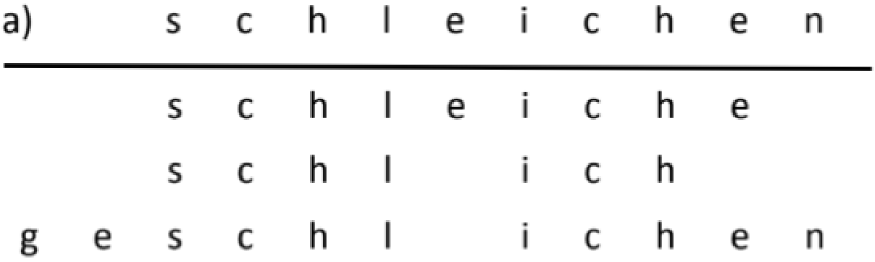
Fig. 3.1: Regular spatial transformations encode semantic or grammatical content (Mikolov et al. 2013)

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). Fig. 3.1 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.

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 Fig. 3.2), 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 Fig. 3.3).



a)		s	c	h	l	e	i	c	h	e	n
		s	c	h	l	e	i	c	h	e	
		s	c	h	l		i	c	h		
g	e	s	c	h	l		i	c	h	e	n

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

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.

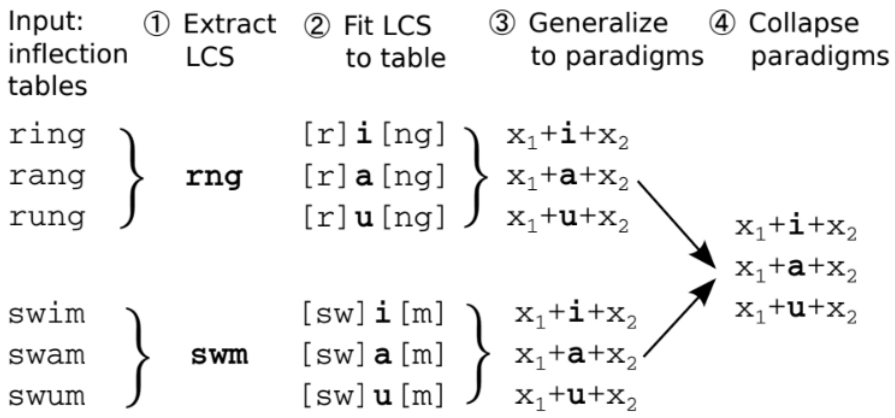


Fig. 3.3: Conceptual depiction of a typical example of a string transduction method of morphology learning (Hulden et al. 2014).

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.

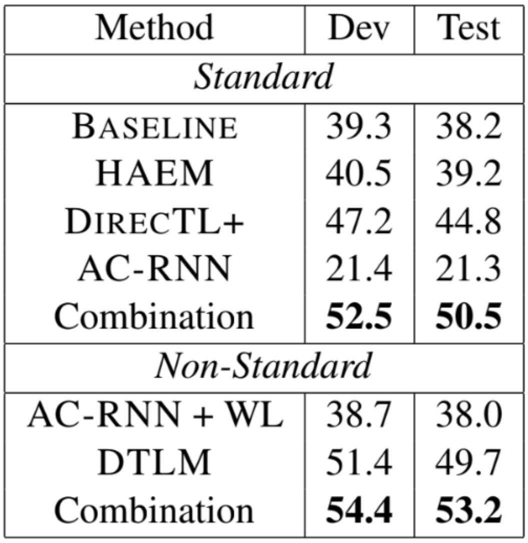
The poorer performance of string transduction relative to neural models has led to a move of the field away from string transduction since about 2016 (Cotterell, Kirov, Sylak-Glassman, Walther, et al. 2018).

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).

3.4 Ensembling approaches



Method	Dev	Test
Standard		
BASELINE	39.3	38.2
HAEM	40.5	39.2
DIRECTL+	47.2	44.8
AC-RNN	21.4	21.3
Combination	**52.5**	**50.5**
Non-Standard		
AC-RNN + WL	38.7	38.0
DTLM	51.4	49.7
Combination	**54.4**	**53.2**

Fig. 3.4: The low-resource accuracy scores of the various University of Alberta models submitted to SIGMORPHON 2018 (Najafi et al. 2018).

Ensembling was used in SIGMORPHON 2018 by various groups between neural and non-neural methods, specifically to ameliorate the poor performance of neural models in low-data settings (Cotterell, Kirov, Sylak-Glassman, Walther, et al. 2018). For example, the University of Alberta submission `UA-05` uses weighted voting to combine the 2019 baseline, a hard-attention LSTM model, a soft-attention LSTM model, and a string transduction model, favoring the output of generally more accurate models unless other models can agree on a different output. The `UA-08` system submitted by the same team combines the output of a string transduction and a soft-attention LSTM in a totally different way, by linear combination of their self-determined confidence scores (Najafi et al. 2018).

Fig. 3.4 shows the accuracy scores of the UA models averaged over all 103 languages of SIGMORPHON 2018 in the low-resource setting. DIRECTL+ and DTLM are string transduction models, BASELINE, HAEM, AC-RNN, and AC-RNN + WL are LSTM models, and the two combinations are `UA-05` and `UA-08`, respectively; notice that the string transduction models score higher accuracy than neural models, but both ensembling methods have the highest accuracy of all.

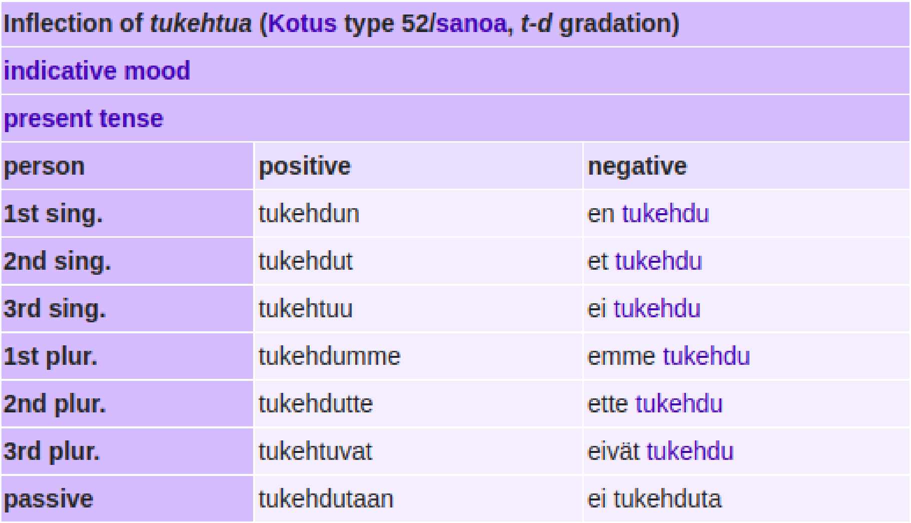
4 Data

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, target grammatical category values, and correctly inflected form - for 103 languages, partitioned into training, development, and test sets. An example of some Finnish triples is shown in Fig. 4.2. 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-resourced 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.

4.1.2 Sourcing and sampling

The English Wiktionary, a collaborative online dictionary, has become something of a standard source of supervised morphological data (Cotterell, Kirov, Sylak-Glassman, Yarowsky, et al. 2016, Cotterell, Kirov, Sylak-Glassman, et al. 2017, Cotterell, Kirov, Sylak-Glassman, Walther, et al. 2018). 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 Fig. 4.1. 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. 4.2, 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 similarity are discussed in sections 5.2 and 5.3 respectively.

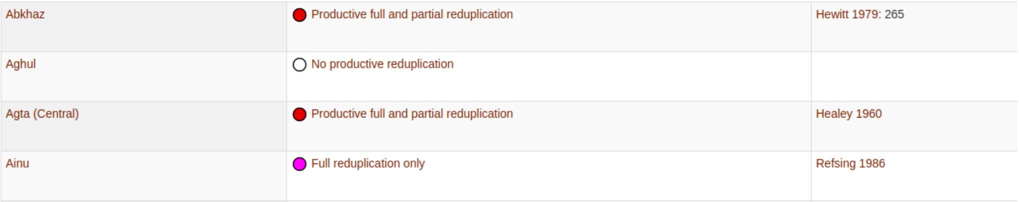
4.2 Language typology and genealogy data

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.

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.



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, a database of basic typological and socioliguistic information about all recognized languages, 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.

4.3 A note on publishing

I have elected not to include my generated data in an appendix to this paper. Please find it online at `https://github.com/cstuartroe/thesis/tree/master/csv`.

5 Methods

This section is essentially a prose description of the computational methods employed to generate data and results for this study. For a more precise look at methodology, the code for this project can be viewed at `https://github.com/cstuartroe/thesis`.

5.1 Assessing the impact of transfer learning

In order to assess transfer learning, I conducted an original data generation phase, with a model based on the SIGMORPHON 2019 baseline model.

5.1.1 Pair selection

242 language pairs were selected, from among 21 languages: Turkish, Bashkir, Crimean Tatar, Uzbek, Spanish, Portuguese, Italian, Romanian, Georgian, Navajo, Arabic, Hebrew, Danish, Swedish, Czech, Slovak, Quechua, Zulu, and Basque. These languages were grouped into 11 groups: Turkic, Romance, Semitic, Germanic, Slavic, and six single-language groups. Each language was the target language in pairs with each language from their group as well as one from every other group as source languages; for instance, Arabic was a target language with Georgian, Estonian, Bashkir, Hebrew, Zulu, Czech, Navajo, Swedish, Quechua, Spanish, and Basque as source languages; pairs were chosen such that each language from a group was a source language roughly the same number of times.

5.1.2 Neural model changes and comparisons

The model used was based on the SIGMORPHON 2019 baseline model, but with two major changes. Firstly, the means of utilizing source language material was switched from concurrent training to pretraining. In the SIGMORPHON 2019 baseline, 10000 examples from the source language and 100 example from the target language were homogenously sampled, such that only about 1% of training examples were from the target language. In this study, a model was pretrained on the source language before being fine-tuned on the target language. This allowed relatively less training time to be spent on the source language, and a single pretrained model of the source language to be adapted to multiple target languages. Secondly, the data setting for the target language was altered from the low-volume to medium-volume SIGMORPHON training sets; that is, from 100 examples of target language morphology to 1000.

Pretraining and training were conducted identically in epochs, with each training example being attempted once per epoch in random order. Hyperparameters were modified in a validation step in between each epoch. 10 epochs were conducted in pretraining, and 20 in training.

The SIGMORPHON 2019 baseline code actually had four models, differentiated by LSTM attention mechanism: soft attention, hard attention with dynamic programming, 0th-order hard attention, and 1st-order hard attention. This study conducted identical data generation with both soft attention and dynamic programming hard attention models, to assess the relationship between model architecture and transfer learning efficacy.

5.1.3 Non-transfer baseline

To assess the impact of transfer learning, outcomes of models with pretraining were compared to models with no transfer learning. These were identical in structure to the transfer learned models, but simply skipped a pretraining step. Target language training of non-transfer learning models was conducted in the same fashion as the transfer learning models: 20 epochs of training over data sets of 1000 examples. The primary dependent variable in this study was the difference in performance between a transfer learning model for a particular language pair, and a non-pretrained model with the same target language.

5.2 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.

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

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

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.

5.3 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).

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

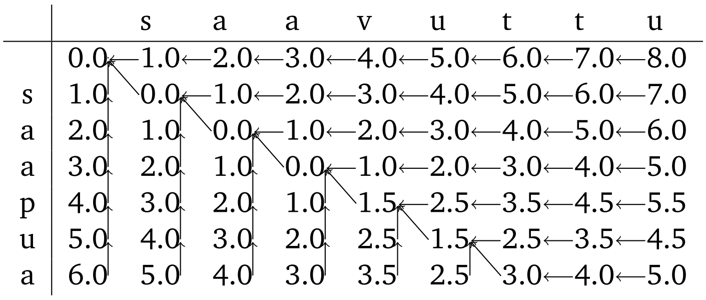
$$POSDS(language\ A, language\ B) = \sum_{POS} \left| \frac{f_{POS,language\ A}}{f_{language\ A}} - \frac{f_{POS,language\ B}}{f_{language\ B}} \right|$$

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

5.4 Inflection shape

Inflection shape - that is, occurrence of prefixing, suffixing, or infixing morphology - may be important in determining, for instance, how attention mechanisms 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. For instance, if the source language indicates plurality with a prefix, then a soft attention model initially trained on that language may weight its attention toward the beginning of a word more highly when attempting to mark a word for plurality. Such focusing of attention may be a desired behavior if that model is adapted to another language which also marks plurality with a prefix.

To generate inflection shape metrics, lemma-inflected form pairs were first aligned by a Levenshtein distance algorithm. As shown in the graphic below, Levenshtein distance was calculated using the classic dynamic programming algorithm with a substitution penalty of 1.5, and traced backwards to identify a string alignment corresponding to the set of insertions, deletions, and substitutions that minimizes edit distance.



		s	a	a	v	u	t	t	u
	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0
s	1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0
a	2.0	1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
a	3.0	2.0	1.0	0.0	1.0	2.0	3.0	4.0	5.0
p	4.0	3.0	2.0	1.0	1.5	2.5	3.5	4.5	5.5
u	5.0	4.0	3.0	2.0	2.5	1.5	2.5	3.5	4.5
a	6.0	5.0	4.0	3.0	3.5	2.5	3.0	4.0	5.0

Fig. 5.1: Computing string alignment using the Levenshtein algorithm.

In the above example, the string pair is aligned as:

```
saapua--
saavuttu
```

with the substitutions `p` → `v` and `a` → `t`, and the addition of a suffix `-tu`. A boolean value is then calculated for the presence or absence of each inflection shape (prefixing, infixing, alternation, and suffixing) based on the aligned pair of strings: if there are unequal characters at the beginning or end of the strings, these are considered to be part of a prefix or suffix (even if the Levenshtein algorithm regarded them as a substitution, so that for instance in this example the `a` → `t` substitution is considered to be part of a suffix, not an alternation. Any non-spacer character aligned with a gap in the word interior is considered to be part of an infix, and two unequal non-spacer characters aligned in the word interior are considered to be part of an alternation, so that for instance in this example the $\mathtt{p} \rightarrow \mathtt{v}$ substitution is regarded as an alternation.

For each language in the dataset, a proportion was calculated for each inflection shape of how many lemma-inflected form pairs in the training set exhibited that inflection shape. For each language pair, an inflection shape similarity coefficient was calculated equal to $1 - \frac{\Sigma_{shape}|p_{s,shape}+p_{t,shape}}{\Sigma_{shape}|p_{s,shape}-p_{t,shape}}$, where $p_{s,shape}$ is the prevalence of a particular inflection shape in the source language, and $p_{t,shape}$ the same for the target language.

5.5 Fusion

To measure fusion between a pair of grammatical categories, average Levenshtein distance was 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 Levenshtein distance from each is 1.5. *Hablo* and *habló* differ in both tense and subject person but also have a Levenshtein distance of 1.5; 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 Levenshtein distance of 1 from both other forms while they have an Levenshtein distance of 1.5 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*.

For each language in the dataset, a set of form pairs which differ only by one UniMorph category (e.g., number, case, tense, or gender) was generated for each category present in that language, and then a set of form pairs which differ by two categories was generated, and Levenshtein distance calculated for each pair. Average distance for each single category and category pair was calculated, and then a coefficient generated for every category pair $c_{a,b} = 2 - \frac{2d_{a,b}}{d_a+d_b}$ where a and b are the two categories, d_a and d_b the average Levenshtein distance between pairs which differ along only one particular category, and $d_{a,b}$ the average Levenshtein distance between pairs which differ by both categories and no other categories. This coefficient typically ranges between 0, indicating no fusion whatsoever, and 1, indicating total fusion between marking of the two categories, although there are not hard boundaries on its range. For instance, Spanish verbs which differ by tense only have an average Levenshtein distance of 2.7, those which differ by person only have an average Levenshtein distance of 3.2, and those which differ by both tense and person but are the same in all other grammatical categories have an average Levenshtein distance of 4.6, so the fusion coefficient for tense and person in Spanish is $2 - \frac{2 \cdot 4.6}{3.2+2.7} \approx .44$, indicating that tense and person marking are somewhat fused in Spanish. By contrast, tense and aspect have a fusion coefficient of 1.07 in Spanish, while person and aspect have a fusion coefficient of .19 (in the SIGMORPHON dataset, the Spanish preterite is considered to be perfective while the imperfect is considered to be imperfective, with all other Spanish verb forms unmarked for aspect).

The similarity of fusion patterns between two languages was a coefficient between 0 and 1, calculated as follows: where $c_{x,y,languageA}$ is the fusion coefficient for tag categories x and y in a particular language A, and $c_{x,y,languageB}$ the same for language B, their fusion similarity is

$$\frac{\sum_x \sum_{y \neq x} |c_{x,y,languageA} - c_{x,y,languageB}|}{\sum_x \sum_{y \neq x} |c_{x,y,languageA}| + |c_{x,y,languageB}|}$$

6 Results and Discussion

In order to demonstrate results, throughout this section I use a sampling of procedurally generated scatter plots demonstrating correlations between performance metrics and other measurements. All scatter plots, including many not shown in this section, can be accessed at

`https://github.com/cstuartroe/thesis/tree/master/images/generated`.

I reference combinations of architectures and performance metrics. There are two of each - hard and soft attention mechanism architectures, and accuracy and Levenshtein distance model performance metrics - and so there are four combinations.

I frequently distinguish between results for closely related language pairs and distantly related and unrelated language pairs. Among the 242 language pairs, 22 were closely related, 26 were distantly related, and the other 194 were unrelated.

6.1 Overall model performance and architectural comparison

I conducted learning trials with two architectures from the SIGMORPHON 2019 baseline - hard monotonic and soft attention. Mean and standard deviation for baseline (non-transfer) model accuracy and Levenshtein distance from correct solution across all 21 languages, and the same information for transfer learning models across all 242 language pairs, is given below:

		μ	σ
Baseline Accuracy	hard	30.6%	23.3%
	soft	25.8%	17.2%
Baseline Levenshtein	hard	2.66	1.41
	soft	2.8	1.01
Transfer Accuracy	hard	48.7%	27.9%
	soft	34.9%	21.8%
Transfer Levenshtein	hard	1.95	1.56
	soft	2.7	1.73

Overall, the hard attention model performed better on average across all metrics. However, variability from language to language was quite substantial and on some languages, the soft attention model performed more strongly. On the whole, performance of the hard and soft attention models on a given language or pair were correlated, but not tightly so.

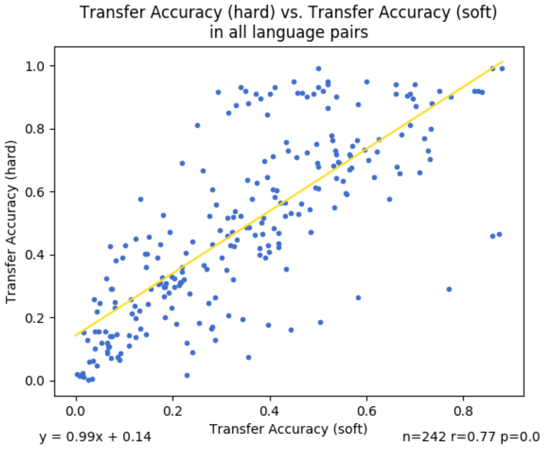


Fig. 6.1: Relationship between the transfer accuracy of hard and soft models, across all 242 language pairs. With r=.77, there is clear correlation of moderately high strength. Comparisons of baseline performance and Levenshtein distances between the two models yielded similar relationships.

Comparison of transfer learning with baseline outcomes demonstrates that transfer learning conferred substantial benefits on models: an average of 18.1 percentage points improved accuracy for the hard attention models, and 9.1 percentage points for the soft attention model.

As will be seen in the following results sections, a number of statistical effects only appear for the hard attention models, not for soft attention models. The best explanation I have for this fact is that, given the overall weaker performance boosts that transfer learning seemed to confer to soft attention models, statistical effects may simply have been more difficult to detect for soft attention models, as random performance differences between target languages would be comparatively larger in relation to transfer learning effects.

6.2 Relationship between source language model performance and transfer learning efficacy

To ascertain whether there is any relationship between the accuracy of a pretrained model and the eventual accuracy of fine-tuned models based on it, correlations were sought between the baseline performance for each source language and the average performance of transfer learned models using it as a source language. The tables and graphics given here use accuracy as the metric for assessing both pretrained and trained models, but results based on Levenshtein distance were essentially the same.

There were no statistically significant differences in average performance improvements between source languages, nor any statistically significant correlations between pretrained model accuracy and average fine-tuned model accuracy. This suggests a lack of evidence that any one language is a universally good source language, or that a pretrained model must be highly accurate on the source language to be useful as a basis for transfer learning.

Source Language	SL Baseline Accuracy (hard)	TL Accuracy Improvement (hard)	
		μ	σ
Basque	0.037	-0.148	0.246
Georgian	0.435	0.005	0.114
Slovak	0.44	0.045	0.132
Uzbek	0.04	0.075	0.141
Hebrew	0.23	0.077	0.096
Danish	0.598	0.111	0.105
Navajo	0.12	0.168	0.169
Portuguese	0.437	0.177	0.147
Turkish	0.145	0.187	0.297
Bashkir	0.745	0.198	0.243
Zulu	0.041	0.199	0.203
Finnish	0.096	0.207	0.132
Quechua	0.248	0.259	0.193
Crimean Tatar	0.93	0.264	0.262
Romanian	0.239	0.275	0.099
Swedish	0.489	0.292	0.232
Italian	0.275	0.314	0.18
Estonian	0.16	0.338	0.198
Spanish	0.359	0.349	0.128
Czech	0.214	0.37	0.22
Arabic	0.149	0.395	0.213

Source Language	SL Baseline Accuracy (soft)	TL Accuracy Improvement (soft)	
		μ	σ
Bashkir	0.348	-0.083	0.191
Slovak	0.233	-0.059	0.196
Georgian	0.184	-0.049	0.134
Danish	0.454	-0.013	0.161
Hebrew	0.203	-0.007	0.123
Uzbek	0.41	0.011	0.123
Basque	0.062	0.028	0.178
Crimean Tatar	0.77	0.029	0.075
Zulu	0.043	0.034	0.19
Turkish	0.176	0.069	0.187
Navajo	0.103	0.096	0.162
Swedish	0.36	0.119	0.143
Portuguese	0.457	0.126	0.243
Romanian	0.185	0.129	0.172
Finnish	0.046	0.139	0.183
Czech	0.164	0.173	0.134
Italian	0.253	0.176	0.181
Arabic	0.105	0.201	0.15
Quechua	0.423	0.225	0.175
Spanish	0.201	0.243	0.142
Estonian	0.233	0.27	0.157

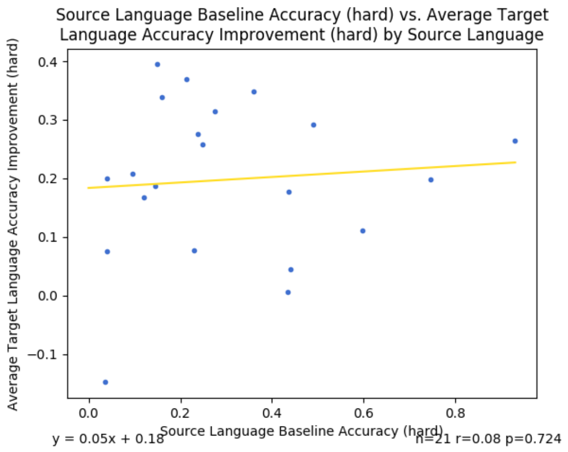


Fig. 6.2

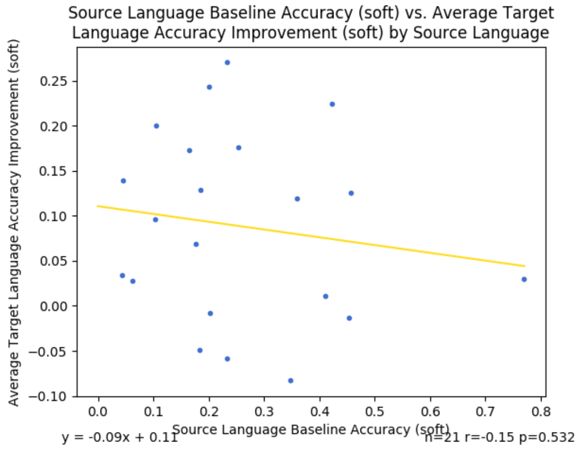


Fig. 6.3

6.3 Part of speech distribution similarity

Using both accuracy rate and Levenshtein distance as performance metrics, greater part of speech distribution similarity seems to predict a larger performance boost from transfer learning in hard attention models, among all language pairs and distantly related and unrelated language pairs. For soft attention models and among closely related languages, there is no clear suggestion in the data of any contradictory effect of part of speech distribution similarity. It seems most likely that (as discussed in 6.1) transfer learning effects were similar but too small to statistically detect for soft attention models, and that the sample size of closely related languages was too small to rise to the level of statistical significance.

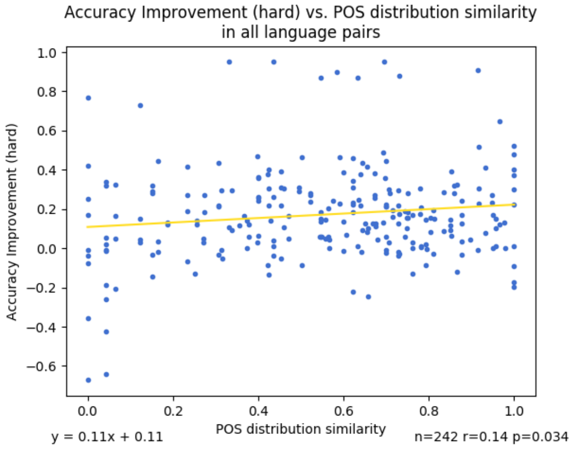


Fig. 6.4

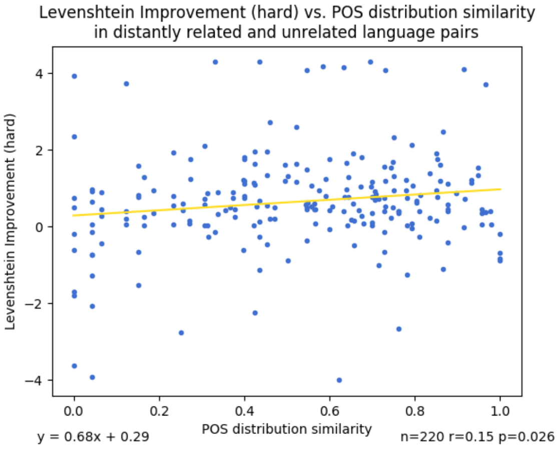


Fig. 6.5

6.4 Relationships between language similarity metrics and model performance

By and large, the takeaway is that all measurements of language similarity predict more effective transfer learning, *except* similarity of adjectival categories and similarity of inflection shape, although what effects do appear rise to the level of statistical significance less often with soft attention models.

6.4.1 Part of speech category overlap

In general, part of speech category overlap was often predictive of effective transfer learning, though there were more nuanced patterns apparent.

Using both accuracy rate and average Levenshtein distance as metrics of performance, nominal category overlap was predictive of performance at a statistically significant level only for hard attention models and not among closely related language pairs.

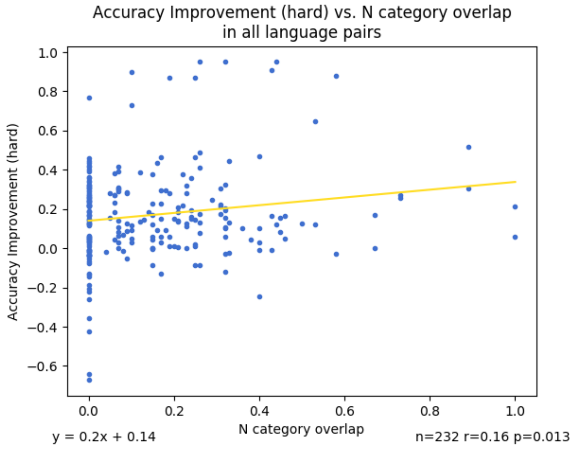


Fig. 6.6: Nominal category overlap has a positive correlation with hard attention model performance.

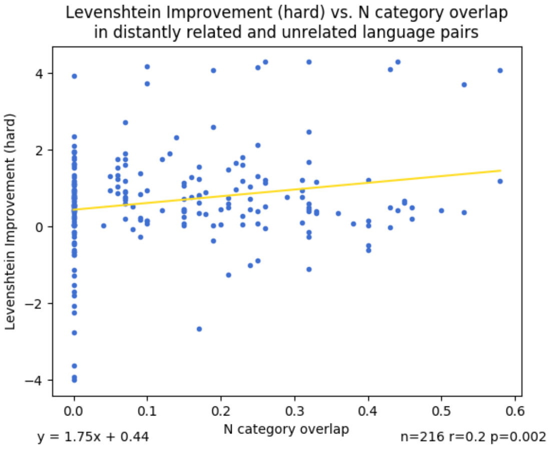


Fig. 6.7: Nominal category overlap has a positive correlation with hard attention model performance.

In fact, for soft attention models, the data was not at all suggestive of a relationship with model performance.

Even given the conditions on the relationship between nominal category overlap and transfer learning efficacy, it appears impossible to dismiss what statistically significant relationships do appear. Figure 6.7 demonstrates that, among all 220 distantly related and unrelated pairs, nominal category overlap is predictive of transfer learning efficacy, and there are no obvious clustering or outlier effects at work.

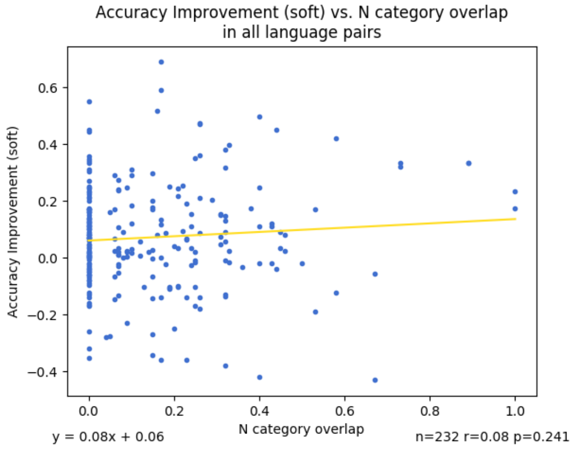


Fig. 6.8: Nominal category overlap has no substantial correlation with soft attention model performance.

Verbal category overlap was broadly predictive of effective transfer learning across both architectures and metrics of performance, but unlike with nominal category overlap, the effect appeared to be driven entirely by differences among closely related language pairs, and differences between closely related language pairs and other language pairs as groups. The data was not at all suggestive of a relationship between verbal category overlap and model performance when limited to distantly related and unrelated pairs.

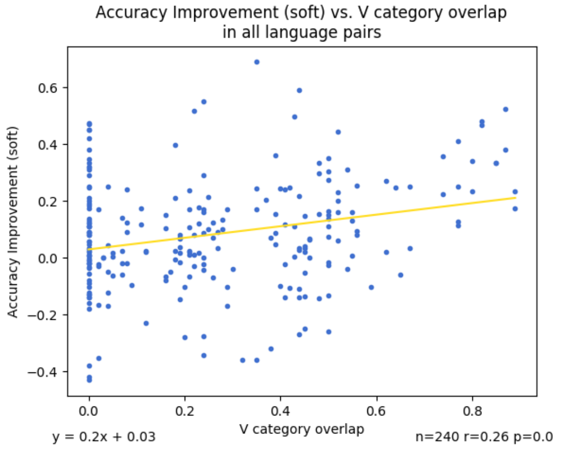


Fig. 6.9: Verbal category overlap is predictive of model performance among all language pairs, but comparison with the next two figures makes it clear that the relationship is characterized by the difference of the rightmost cluster, entirely composed of closely related language pairs, from the rest of the pairs.

It appears from the scatter plots of verbal category overlap and model improvements among closely related languages that the positive correlation is driven primarily by six outlying pairs with substantially lower verbal category overlap than the others;

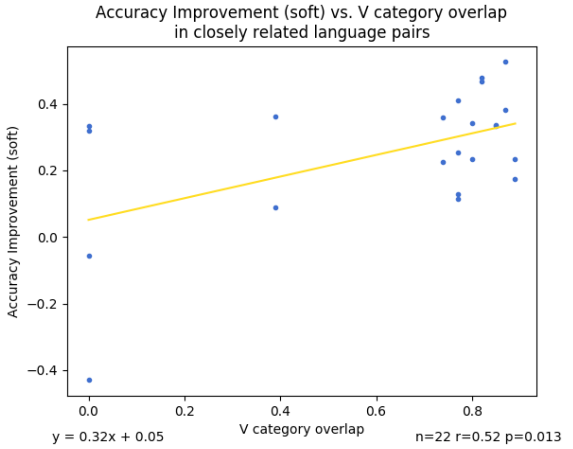


Fig. 6.10: The main cluster of language pairs visible here can be seen as well in Figure 6.9.

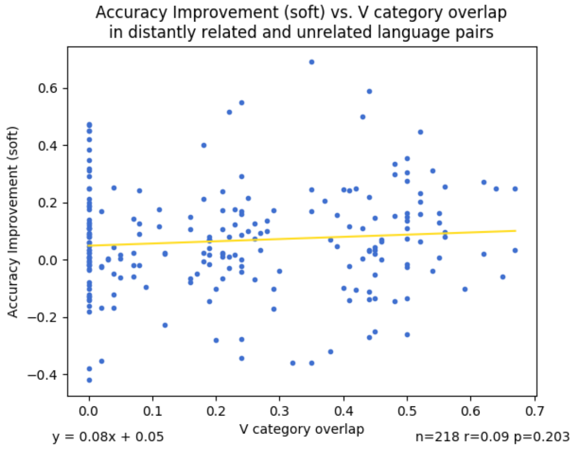


Fig. 6.11: With the cluster of closely related languages removed, verbal category overlap is no longer predictive of performance.

these pairs are Bashkir ↔ Crimean Tatar, Czech ↔ Slovak, and Arabic ↔ Hebrew. Given that this relationship boils down to three pairs of closely related language which happen to have substantially different verbal systems (at least as represented by the input data), it is probably best to avoid drawing too strong a conclusion.

The only robust correlation between verbal category overlap and model performance, then, is the separate clustering of closely related languages and others, and seems likely to simply be reflective of the facts that closely related languages are much more likely to have similar verbal systems and that genealogical similarity is predictive of effective transfer learning. That is, despite many statistically significant relationships between verbal category overlap and transfer learning efficacy, this appears on close examination to be a case of mutual correlation with genealogical similarity.

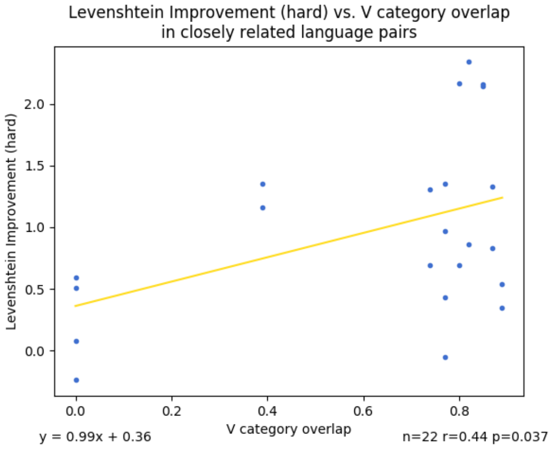


Fig. 6.12: Verbal category overlap is predictive of effective transfer learning for closely related language pairs, regardless of model architecture and performance metric.

A similar picture appears for adjectival category overlap, though even more distinctly so. Adjectival category overlap predicts effective transfer learning as measured by average Levenshtein distance among closely related languages, but this seems to be entirely driven by the outlying pairs Bashkir ↔ Crimean Tatar and Czech ↔ Slovak. No broader conclusions should be drawn about the relationship between adjectival category overlap and transfer learning.

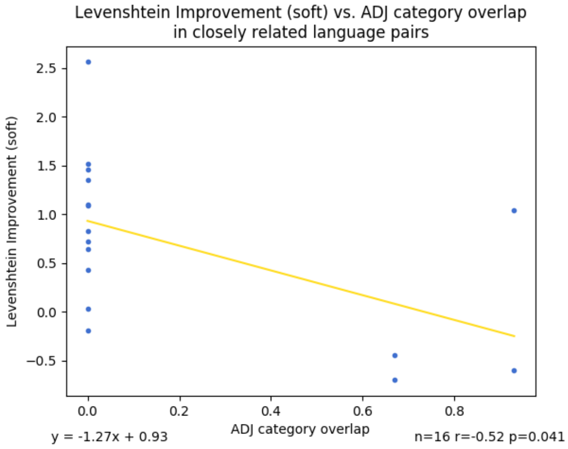


Fig. 6.13: The four rightmost data points are Bashkir ↔ Crimean Tatar and Czech ↔ Slovak. The adjectival category overlap of 0 among most pairs is usually due to one or both languages not having any adjectives in their data set; adjectives were less common by far than nouns and verbs across all the input data sets.

6.4.2 Inflection shape

Across all metrics of performance and language pair subsets tested, any correlation between inflection shape similarity and transfer learning efficacy never rose to a

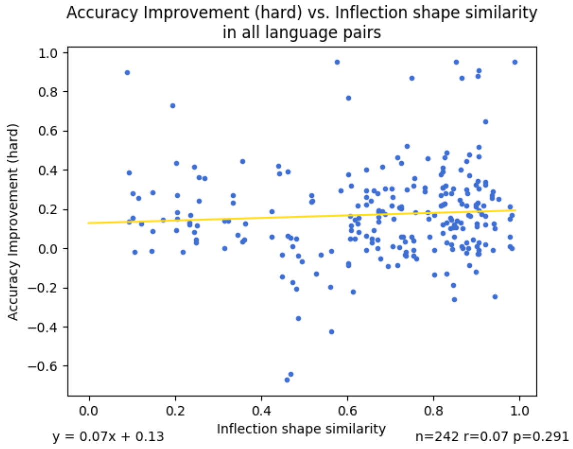


Fig. 6.14

statistically significant level. Of all metrics of language similarity assessed in this study, inflection shape is certainly the most unambiguously lacking in correlation with effective transfer learning. From this, it seems straightforward to conclude that there is no evidence of similarity in inflection shapes exhibited by two languages being predictive of effectiveness as a transfer learning pair, at least with the transfer learning strategies tested here.

6.4.3 Fusion

Viewed across the dataset of all language pairs tested, similarity in morphological fusion had a statistically significant positive correlation with transfer learning improvement for all combinations of architecture and performance metric. It was also significant across the dataset of only closely-related language pairs in three out of four architecture × performance metric combinations.

However, for the data set of distantly related and unrelated languages, the correlation of fusion similarity and transfer learning efficacy was only significant for one combination: soft attention architecture and accuracy, with $p = .004$, while for all other combinations the p-value was greater than .15. Unlike with part of speech category overlaps, this is not apparently due to separate clustering of closely related languages. The data sets with fusion similarity which did not rise to any level of statistical significance all still show a similar positive correlation, and so it may be that the effect exists and simply went undetected for those data sets.

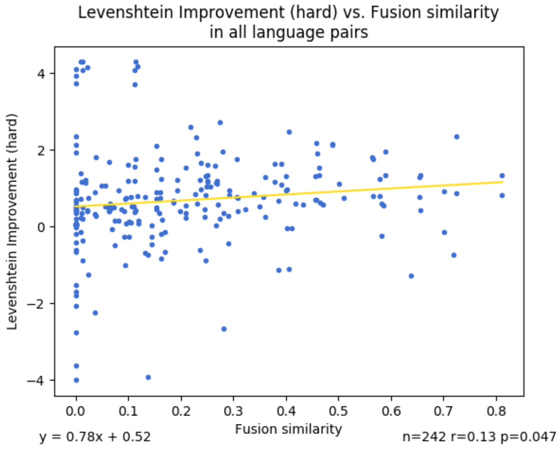


Fig. 6.15: This was the combination of architecture and performance metric for which the correlation with performance improvement, among all language pairs, was weakest. For all three other combinations, correlation was significant at a $p < .01$ level.

It is still unclear why the effect was most pronounced for the combination of soft attention and accuracy as a performance metric.

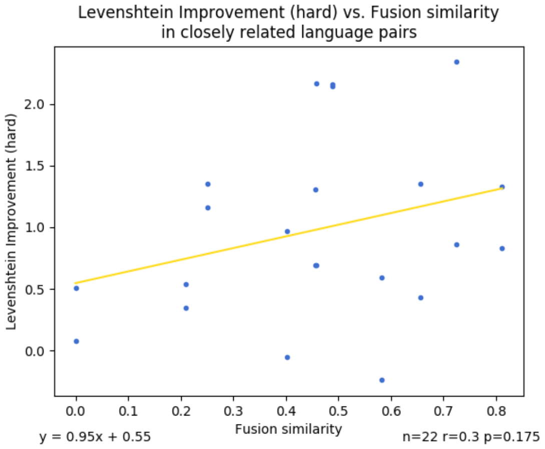


Fig. 6.16: Although Levenshtein improvement × hard attention × fusion similarity shows a statistically significant correlation among all languages, it does not among the subset of closely related languages. Given that there is still a similar trendline, it may be that the sample size here is simply too small to detect the effect.

6.4.4 Genealogical distance

For all combinations of architecture and performance metric, genealogical similarity was correlated with more effective transfer learning pairs. That is, a source language more closely related to a particular target language was most likely to result in a higher-performing transfer learned model by all metrics.

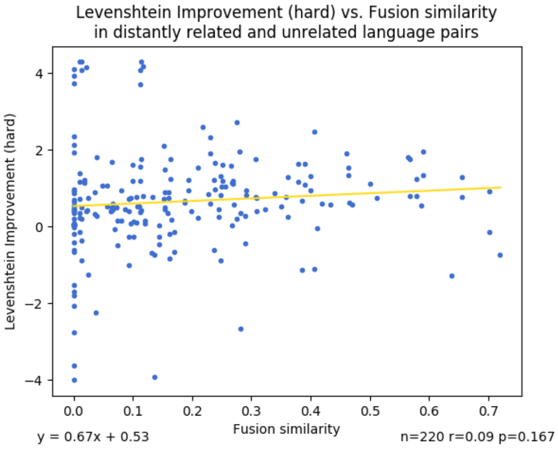


Fig. 6.17: Although Levenshtein improvement × hard attention × fusion similarity shows a statistically significant correlation among all languages, it does not among the subset of distantly related and unrelated languages.

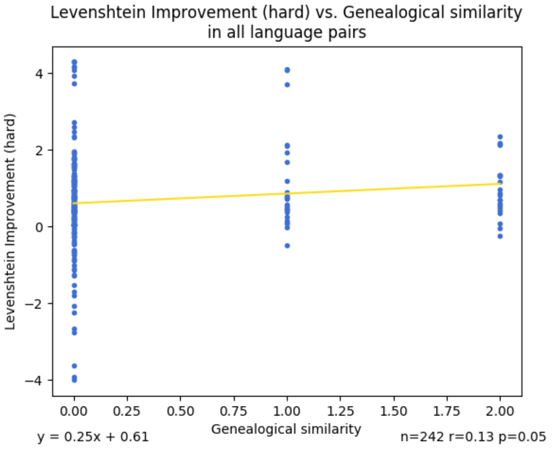


Fig. 6.18: This was the combination of architecture and performance metric for which the correlation between genealogical similarity performance improvement was weakest. For all three other combinations, correlation was significant at a $p < .005$ level.

Separate data sets for closely related languages and distantly related and unrelated languages were not considered for this explanatory variable, for obvious reasons.

7 Conclusion and Future Work

The goal of this study was to disentangle some of the factors influencing whether a particular high-resources language is a good candidate for a source language to conduct transfer learning with a particular low-resource target language. This research question has value to computational linguists attempting to develop new morphologically savvy models of low-resource languages. As suggested by McCarthy et al. 2019, this study produced evidence that a source language closely related to the target language is moderately predictive of effective transfer learning. It also elucidated other types of similarities relevant to transfer learning.

Broadly, it appears that the largest hurdles for the types of LSTM models tested here center around interpreting grammatical categories, rather than understanding spelling and word structure of languages. Anecdotally, when viewing many of the incorrect predictions that my test models produced, they often corresponded to other correct forms in the target language but didn't match the intended set of grammatical categories; at the least, they typically appeared to be consistent with the phonology and spelling rules of the language, and fairly close to the lemma and/or the correct form. I would interpret many of the statistical results of this study as being consistent with that conjecture: similarity in patterns of grammatical fusion was moderately predictive of transfer learning efficacy, and the most predictive explanatory variable was similarity between source and target language in the overall set of grammatical categories exhibited on nouns. In contrast, similarity in inflection shape has no perceivable bearing on model performance. Lastly, the statistical relationship between part of speech distribution similarity and transfer learning efficacy suggests that grammatical tag similarities between the *data sets* for source and target language, separate from their actual structure, bears on the usefulness of transfer learning. On the basis of these facts, my distillation of the suggestions of the results of this study are:

> When selecting a source language for transfer learning of a particular target language, morphophonological and inflection shape similarities appear to be less relevant than similarities in the grammatical structure of the two languages and similarities in the set of grammatical tags in the source and target data sets.

That said, there are questions raised by this study that I consider to be not fully answered, in particular questions of the relationship between attention mechanism and linguistic characteristics in determining model performance. Overall, hard attention models better leveraged transfer learning and showed stronger performance overall, and showed a relatively stronger sensitivity to nominal category overlap, while soft attention models showed a greater sensitivity to fusion pattern similarities. These questions are indirectly addressed by Cotterell, Kirov, Sylak-Glassman, Walther, et al. 2018 and McCarthy et al. 2019 as they compare the performance of models which vary in design, but those studies do not isolate attention as a variable or explicitly take linguistic properties into account.

I believe that research on that front could be pursued by various further analyses of the data generated here, such as comparing differences in performance between hard and soft attention models with actual linguistic properties of languages, rather than simply similarity of those properties as done here. Study could also be done using new model architectures that explicitly takes into account linguistic typological information such as that used in this study.

On that note, a ripe area for further research involves making model design choices not considered here. For instance, while McCarthy et al. 2019 used a low-resource setting of 100 training examples for target languages, this study used a resource setting of 1000 training examples for each target language, in large part because typical accuracy rates at the lower data setting were too low to reliably detect statistical trends. This may be related to intrinsic learning curve weaknesses of LSTMs (Cotterell, Kirov, Sylak-Glassman, et al. 2017, Cotterell, Kirov, Sylak-Glassman, Walther, et al. 2018), and so a study similar to this one that makes use of ensembling to ameliorate this issue may be of value, since models of very low-resource languages are likely to make use of some type of ensembling.

Another design choice that has not been previously investigated for this family of computational morphology problems is the use of multiple source languages. It seems perfectly probable that pretraining with several source languages, perhaps languages similar to the target language in different ways, could further push the potential of models.

Bibliography

Ahlberg, Malin, Markus Forsberg, and Mans Hulden (Jan. 2015). “Paradigm classification in supervised learning of morphology”. In: pp. 1024–1029. DOI: 10.3115/v1/N15-1107.

Alexandrescu, Andrei and Katrin Kirchhoff (Jan. 2006). “Factored Neural Language Models.” In: DOI: 10.3115/1614049.1614050.

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.

Cotterell, Ryan and Georg Heigold (2017). “Cross-lingual, Character-Level Neural Morphological Tagging”. In: *CoRR* abs/1708.09157. arXiv: 1708.09157. URL: http://arxiv.org/abs/1708.09157.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden (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. DOI: 10.18653/v1/K18-3001.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden (2017). “CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages”. In: *CoRR* abs/1706.09031. arXiv: 1706.09031. URL: http://arxiv.org/abs/1706.09031.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden (Aug. 2016). “The SIGMORPHON 2016 Shared Task—Morphological Reinflection”. In:

Cotterell, Ryan and Hinrich Schütze (2019). “Morphological Word Embeddings”. In: *CoRR* abs/1907.02423. arXiv: 1907.02423. URL: http://arxiv.org/abs/1907.02423.

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. URL: http://dl.acm.org/citation.cfm?id=3044805.3045095.

Dreyer, Markus, Jason Smith, and Jason Eisner (Jan. 2008). “Latent-Variable Modeling of String Transductions with Finite-State Methods.” In: pp. 1080–1089. DOI: 10.3115/1613715.1613856.

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*. URL: http://aclweb.org/anthology//N/N13/N13-1138.pdf.

English Wiktionary (n.d.). https://en.wiktionary.org. Accessed: 2019-10-25.

Ethnologue (n.d.). https://www.ethnologue.com/. Accessed: 2019-11-17.

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. URL: http://arxiv.org/abs/1512.06110.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). “Long Short-Term Memory”. In: *Neural Comput.* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hogan, P.C. (2010). *The Cambridge Encyclopedia of the Language Sciences*. Cambridge University Press. ISBN: 9780521866897. URL: https://books.google.com/books?id=t7T-AAAACAAJ.

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. DOI: 10.3115/v1/E14-1060. URL: https://www.aclweb.org/anthology/E14-1060.

Krogh, Anders and Jesper Vedelsby (1995). “Neural network ensembles, cross validation, and active learning”. In: *Advances in neural information processing systems*, pp. 231–238.

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. DOI: 10.18653/v1/D15-1166. URL: https://www.aclweb.org/anthology/D15-1166.

McCarthy, Arya D., Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sebastian J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden (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. DOI: 10.18653/v1/W19-4226.

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.

Najafi, Saeed, Bradley Hauer, Rashed Rubby Riyadh, Leyuan Yu, and Grzegorz Kondrak (Oct. 2018). “Combining Neural and Non-Neural Methods for Low-Resource Morphological Reinflection”. In: *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*. Brussels: Association for Computational Linguistics, pp. 116–120. DOI: 10.18653/v1/K18-3015. URL: https://www.aclweb.org/anthology/K18-3015.

Nicolai, Garrett, Colin Cherry, and Grzegorz Kondrak (Jan. 2015). “Inflection Generation as Discriminative String Transduction”. In: pp. 922–931. DOI: 10.3115/v1/N15-1093.

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. DOI: 10.1109/TKDE.2009.191.

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.

Soricut, Radu and Franz Och (Jan. 2015). “Unsupervised Morphology Induction Using Word Embeddings”. In: pp. 1627–1637. DOI: 10.3115/v1/N15-1186.

Sylak-Glassman, John (2016). “The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema)”. In:

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. ISBN: 978-3-319-23980-4.

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. DOI: 10.3115/v1/P15-2111.

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

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. DOI: 10.18653/v1/P19-1148. URL: https://www.aclweb.org/anthology/P19-1148.

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. DOI: 10.18653/v1/D16-1163. URL: https://www.aclweb.org/anthology/D16-1163.