Minimal Supervision for Morphological Inflection

Omer Goldman
Bar Ilan University
omer.goldman@gmail.com

Reut Tsarfaty
Bar Ilan University
reut.tsarfaty@biu.ac.il

Abstract

Neural models for the various flavours of morphological reinflection tasks have proven to be extremely accurate given ample labeled data, yet labeled data may be slow and costly to obtain. In this work we aim to overcome this annotation bottleneck by bootstrapping labeled data from a seed as small as *five* labeled inflection tables, accompanied by a large bulk of unlabeled text. Our bootstrapping method exploits the orthographic and semantic regularities in morphological systems in a two-phased setup, where word tagging based on *analogies* is followed by word pairing based on *distances*. Our experiments with the Paradigm Cell Filling Problem over eight typologically different languages show that in languages with relatively simple morphology, orthographic regularities on their own allow inflection models to achieve respectable accuracy. Combined orthographic and semantic regularities alleviate difficulties with particularly complex morpho-phonological systems. We further show that our bootstrapping methods substantially outperform hallucination-based methods commonly used for overcoming the annotation bottleneck in morphological reinflection tasks.

1 Introduction

The introduction of neural models into natural language processing in the last decade has led to huge improvements in all *supervised* generation tasks, including morphological inflection.[1] In particular, previous works (Cotterell et al., 2017; Silfverberg and Hulden, 2018) have achieved near-perfect performance over the Paradigm Cell Filling Problem (PCFP) (Ackerman et al., 2009), wherein models are required to provide any form in an inflection table, given a few forms of the same lexeme.[2]

Two lines of recent work made progress towards less supervision, in different fashions. The first simply provided scenarios with smaller training sets — for example, in SIGMORPHON's shared tasks (Cotterell et al., 2017, 2018). The second research avenue aims to discover the paradigmatic structure of an unknown language given a large bulk of unlabeled data, either alone (Soricut and Och, 2015; Elsner et al., 2019), accompanied by a list of all relevant forms in the vocabulary (Erdmann et al., 2020), or by a list of lemmas (Jin et al., 2020).

The problem with the first kind of attempts is that given the neural nature of the most successful models, their performance on limited supervision is capped, and data augmentation is likely to help only if the initial data is diverse enough. As for the second scenario of no supervision at all, it is somewhat pessimistic and unrealistic. Even if much labeled data for a language does not exist for a low-resourced language, typically there exists *knowledge* about its paradigm structure that can be employed. UniMorph (Kirov et al., 2018), for example, includes small amounts of labeled inflection tables for many languages, from obscure ones like Ingrian to national languages with widespread usage that lack global attention like Georgian.

In this work we propose a new, *low-resourced morphological inflection* scenario, which is more optimistic and realistic for those widely-spoken sparsely-annotated languages. We assume a *minimal* supervision set and a large bulk of unlabeled text, thus balancing both trends of lowering supervision resources. We bootstrap a tiny amount of as little as *five* inflection tables, that could be eas-

[1] In recent years the term *reinflection* has surfaced as a reference to morphological inflection done not necessarily from the lemma. In this paper we will refer to both inflection and reinflection as "inflection", and specify whenever we refer to inflection done exclusively from the lemma.

[2] Throughout the paper we conform to the linguistic terminology, where 'lexeme' stands for an abstract lexical entry, e.g., the English RUN, and its 'forms' are the words that convey this lexical meaning with some inflectional relations between them, e.g. *run*, *running* but not *runner*. 'Paradigm' will stand for a group of lexemes sharing a POS tag, in the same manner as the English lexeme RUN is part of the verbal paradigm.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2078–2088
November 7–11, 2021. ©2021 Association for Computational Linguistics

ily written by hand, into a full-blown training set, albeit noisy, for an inflection model trained in a supervised manner. Our approach makes use of the regularities abundant in inflectional morphology, both orthographic regularities and semantic ones.

Based on this method we train morphological inflection models for eight languages and show that, for the five Indo-European of them, orthographic regularity is enough to train a morphological inflector that achieves reasonable success. We further show that for languages with complicated morpho-phonological systems, such as Finnish, Turkish and Hungarian, a method combining both orthographic and semantic regularities is needed in order to reach the same level of performance. An error analysis reveals that the closer an inflection is to the verge of disappearance, the poorer our system performs on it, as less examples exist in the data-derived vocabulary. Our models outperform Makarov and Clematide (2018)'s model designed for low-resourced setting, even when equipped with additional hallucinated data (Anastasopoulos and Neubig, 2019). We also outperform the best model of Jin et al. (2020), that didn't use any inflection tables, and their skyline for most languages.

We conclude that bootstrapping datasets for inflectional morphology in low-resourced languages, is a viable strategy to be adopted and explored.[3]

2 The Minimally-Supervised Setup

Problem Statement Let $\mathcal{L}$ be a set of lexemes, each corresponding to an inflection table $W^l = \{w^l_{t_1}...w^l_{t_m}\}$ where $w^l_{t_i}$ is a form of lexeme l bearing the feature bundle t_i. Our goal is to train an inflection model that maps words bearing one set of features to words bearing another set within the same lexeme.

$$(\langle t_j, w^{l_i}_{t_j}\rangle, t_k) \mapsto w^{l_i}_{t_k}$$

For example, in the French verbal paradigm:
$\big(\langle$ INDPRS1SG, *finis* $\rangle$, SBJVPST2SG$\big)$
$\mapsto$ *finisses*

In order to induce this function, we propose a *minimally-supervised* scenario where we are only given a small set of n examples of complete inflectional tables $\mathcal{L} = \{W^{l_1}...W^{l_n}\}$ (each of which is of size m), and a large bulk of naturally occurring (that is, in-context) unlabeled data in that language, that is, $w_1.....w_c$, such that $c >> nm$.

3 The Algorithmic Framework

This work suggests utilizing the patterns exhibited in the small supervision seed, and finding words that exhibit *similar* (or, *analogous*) patterns.

The algorithm proposed here works in two phases. First we tag words with morphological features if they are found similar (analogous) to examples in the minimal supervision. Then, we pair the tagged words such that each pair will include two forms of the same lexeme. This algorithmic division of labor allows the pairing module to provide a sanity check, and reduce the noise potentially lingering from the word-tagging module.

The above sketch of the algorithm is quite generic, that is, we can get different instantiations of this framework by plugging in different ways to calculate *similarities* or *analogies* between words. In our various algorithmic implementations, we will use the regularities that prevail in inflectional morphology and are detectable even from a tiny amount of supervision. These regularities are manifested both orthographically, as edits between forms tend to repeat themselves across lexemes, and semantically, as forms that share an inflectional function tend to be used similarly.

In the rest of this section we will describe both modules, each with its different variants depending on the different notion of *similarity* used.

3.1 Morphological Features Assignment

In order to assess whether a pair of unseen words w_1, w_2 belong to the same lexeme, we first need to characterize the relationship between those two words. It is then imperative to compare the concrete relation between w_1, w_2 to some representation of the abstract relation R_{t_j,t_k} between 2 morphological categories t_j and t_k in the same paradigm. If these concrete and abstract relations are sufficiently similar, w_1 and w_2 will be tagged as bearing features t_j and t_k, respectively. The idea, in a nutshell, is to obtain the representation R_{t_j,t_k} by aggregating differences between the forms of the t_j, t_k entries in all n inflection tables in the minimal seed. These differences can be stated in terms of either semantics, orthography or a combination thereof.

3.1.1 Orthography-Based Tagging

In the orthographic case we define the difference between a pair of words as the edits needed to get from one word to the other. The edits our system expects are a list of sub-strings that were deleted

[3] Our code is available on https://github.com/OnlpLab/morphodetection.

and a list of those added, sorted from left to right.[4]

For every pair of morphological categories (t_j, t_k), their orthographic relation $R^O_{t_j,t_k}$ is defined by the set of edits observed in the n supervision inflection tables, one from each lemma l_i

$$\hat{R}^O_{t_j,t_k} := \left\{ \text{edits}(w^{l_i}_{t_j}, w^{l_i}_{t_k}) \right\}^n_{i=1}$$

In order to check whether a pair of new words (w_a, w_b) exhibits a relation between two t_j, t_k morphological categories, we check whether the edits between them belong to the relation representation:

$$\text{edits}(w_a, w_b) \overset{?}{\in} \hat{R}^O_{t_j,t_k}$$

Consider the example from the French verbal paradigm where the examples for the relation (IndPrs1Sg, IndPrs3Sg) are {*(finis,finit), (bois, but), (parle,parla)*}. In this case the representation of this relation is:

$$\hat{R}^O_{\text{IndPrs1Sg,IndPrs3Sg}} = \\ \{(\text{'}s\text{'}, \text{'}t\text{'}), (\text{'}ois\text{'}, \text{'}ut\text{'}), (\text{'}e\text{'}, \text{'}a\text{'})\}$$

Since the edits between (*aime, aima*) are identical to those between (*parle,parla*), then *aime* would be correctly considered for tagging as IndPrs1Sg and *aima* – as IndPstPrf3Sg.

This procedure is however highly prone to coincidence. For example, the edits between *le* and *la* are the same as between *parle* and *parla*, although the former are actually determiners, and not part of the verbal paradigm. Given the multitude of relations available, we can expect many edits between *incidental* pairs of words to match an edit seen in the gold data. To overcome this, we propose to take the complete paradigm structure into account, rather than considering word pairs in isolation.

Concretely, we propose to tag only words that have been found to answer this criterion for *multiple relations in the same paradigm*, covering at least half of the size of the paradigm. So *aime* would be tagged as IndPrs1Sg since edits(*aime, aima*) = edits(*parle,parla*) and edits(*aime, aimons*) = edits(*parle, parlons*) and so on, but *le* won't be tagged as IndPrs1Sg since the French vocabulary does not contain *lons*.

With this tightened criterion, the orthographic algorithm might be sufficiently precise, but may not be sufficiently diverse, so as to obtain high recall.

3.1.2 Semantics-Based Tagging

The orthographic criterion only considers exact match with the observed edits so it is expected to miss irregulars and classes unattested in the n supervision inflection tables.[5] This can pose a significant problem to paradigms that have more than n classes or display significant morpho-phonological processes not present in the labeled examples.

To overcome the generalization problem of orthographic edits we propose to consider semantic regularities, since the differences in meaning and usage rarely correlate with orthography. Semantic regularity arises from agreement, a phenomenon in which words in a sentence must have the same morphological features as some other words in the sentence, effectively creating equivalence classes. Modern algorithms for word embeddings that extract semantics from co-occurrences, following Firth (1957), are naturally suitable to exploit this kind of regularity.

In this setting the difference between words is defined as the difference between their embedded vectors. And for every pair of morphological feature bundles (t_j, t_k) the representation of their semantic relation is estimated by the average over those relevant examples

$$\hat{R}^S_{t_j,t_k} = \frac{1}{n}\sum_{i=1}^{n} v(w^{l_i}_{t_j}) - v(w^{l_i}_{t_k})$$

A new word pair will be tagged t_j, t_k if their difference is close enough, in cosine-distance terms, to $\hat{R}^S_{t_j,t_k}$

$$D_C\left(\hat{R}^S_{t_j,t_k}, v(w_a) - v(w_b)\right) \overset{?}{\leq} \hat{C}^S_{t_j,t_k}$$

where D_C is the cosine distance function and $\hat{C}^S_{t_j,t_k}$ is an estimation of a relation-specific cut-off score set by the average scatter of the relevant supervision examples around their average:

$$\hat{C}^S_{t_j,t_k} = \frac{1}{n}\sum_{i=1}^{n} D_C\left(\hat{R}^s_{t_j,t_k}, v(w^{l_i}_{t_j}) - v(w^{l_i}_{t_k})\right)$$

Although lacking the orthographic disadvantages, here $\hat{R}^S_{t_j,t_k}$ might be a biased representation that misses many examples, or mistakenly tags incorrect words. For this reason we suggest the third algorithm combining both types or regularities.

[4] Though without position indices, in order to deal with words of various lengths.

[5] the term 'class' refers to a group of lexemes in a paradigm that display similar inflection patterns. Traditionally known as 'conjugation' and 'declension' in the description of verbal and nominal paradigms, respectively. E.g., the Spanish verbal paradigm is said to include 3 classes: *-er*, *-ar* and *-ir* verbs.

3.1.3 Combined Tagging

The idea behind this variant is to consider word pairs that answer both the orthographic and semantic criteria as *semi-gold* examples that will be added to better estimate the relation $\hat{R}^S_{t_j,t_k}$. Thus we harness the accuracy of the orthographic criterion to combat the bias in the semantic representation.

Specifically, a pair of words (w_1, w_2) is considered *semi-gold* if their edit script is in $\hat{R}^O_{t_j,t_k}$, **and** the distance of their difference vector from the relation vector $\hat{R}^S_{t_j,t_k}$ is smaller than the furthest corresponding difference vector of gold examples.

Note that we relaxed both criteria comparing to those in the previous sections. The orthographic criterion is relaxed by dropping the requirement for the complete paradigm, and the semantic criterion is relaxed by replacing $\hat{C}^S_{t_j,t_k}$ with a more inclusive cut-off relying on max rather than average distance

$$\hat{C}^{Comb}_{t_j,t_k} = \max_i D_C\left(\hat{R}^s_{t_j,t_k}, v(w^{l_i}_{t_j}) - v(w^{l_i}_{t_k})\right)$$

The relaxations allow inclusion of more *semi-gold* examples, and they are sensible as both criteria are mutually constraining.

Pairs of words that have satisfied both criteria are very likely correct examples of the morphological relation (t_j, t_k). Thus they are added to create a new semantic representation $\hat{R}^S_{t_j,t_k}$. The new representation may be used again to find more semi-gold examples, executing this stage iteratively. We set the stopping criterion when either no examples are added to $R^S_{t_j,t_k}$ after an iteration, or when so many examples where added so that the run time per iteration exceeded 48 hours in our implementation.

Once $\hat{R}^S_{t_j,t_k}$ is settled, and in order to include tagging of words with different edits, we finally tag words according to the semantic criterion alone with the corresponding $\hat{C}^S_{t_j,t_k}$.

3.2 Pairing Tagged Words

Given the output of the first module, a list of words tagged with some morphological sets of features, the second module pairs tagged words from the first module, such that both tagged words are forms of the same lexeme.

We start by grouping the tagged words from the first module into m bins $\{B_{t_i}\}^m_{i=1}$, each containing all words tagged with t_i. We then need to find for each word $w^{l_j}_{t_i} \in B_{t_i}$ corresponding words from other bins that are forms of the same lexeme l_j.

Ideally, if each bin contained exactly one correct form of every lexeme, pairing the most similar words across bins should suffice in order to collect all the forms in the paradigm. In reality, due to noise, bins may include several forms of the same lexeme or none at all. Assuming enough words are tagged, it seems better to simply drop the cases where several forms of the same lexeme are occupying the same bin, rather then trying to locate the one that was tagged correctly. Therefore, we would like to pair words such that they are *distinct* nearest neighbors across bins, i.e., where the second nearest neighbor is much more distant.

To achieve this, we scale the distance using the Cross-domain Similarity Local Scaling (CSLS) score, suggested by Lample et al. (2018) in the context of bi-lingual embeddings. The CSLS score scales the cosine distance between vectors across groups $x \in X, y \in Y$ with the average distance from their top-k nearest neighbors. We set $k = 2$ when applying this method here, since we aim to discard cases where a form $w^{l_q}_{t_i} \in B_{t_i}$ has *at least* two corresponding forms in another bin $w^{l_q}_{t_j}, w^{l_q}_{t_k} \in B_{t_j}$, with one presumably misplaced. The definition of CSLS for $k = 2$ can be written as

$$CSLS(x, y) = D(x, y) - \frac{1}{2}D(x, NN^x_2(Y)) - \frac{1}{2}D(NN^y_2(X), y)$$

where $NN^x_2(Y)$ is x's second nearest neighbor in group Y.

Although the original distance measure $D(\cdot, \cdot)$ used by Lample et al. (2018) is the cosine distance function, any distance measure will do. In the semantic algorithm (Sec. 3.1.2), we apply cosine distance that reflects semantic similarity. In the orthgraphic algorithm (Sec. 3.1.1), we plug in the Levenshtein edit distance, to utilize the orthography rather than the semantics.

After scoring all possible word pairs, the best scored pairs are taken as a training set for a supervised neural morphological inflector.

4 Experimental Evaluation

4.1 Experimental Setup

We set out to empirically examine whether it is possible to train a morphological inflection model while starting with a minimal number of labeled inflection tables, using regularities of different types. Our experiments include the verbal paradigm of eight languages: English, Russian, Latvian, Spanish, French, Finnish, Hungarian and Turkish, although our methods are suitable for any paradigm.

To make evaluation of the inflection model possible, we had to choose languages with a sufficient amount of gold (test) data, and *simulate* a low-resource scenario for training. This limited us to western languages, and we aimed to include as many non-Indo-European languages as possible.

The Unlabeled Data The problem setting we specified in Section 2 designates the use of a bulk of unlabeled text. In actuality, the proposed algorithm makes use of the text for (i) collecting a vocabulary to seek tagging candidates, and (ii) training embeddings to be used for calculating the semantic relations between candidate pairs. In our experiments, we simply employed language-specific pre-trained word embeddings for both purposes.

We used the pre-trained FastText vectors provided by Grave et al. (2018), following their reported success over morphological analogies in the Bigger Analogy Test Set (Li et al., 2018). We clipped the 2 million long FastText vocabulary to include only real words, i.e., we keep only lower-cased tokens that do not include non-alphabetic characters, and include at least one vowel.[6] This procedure downsized the vocabulary size to between about $200k - 500k$ words per language.[7] Additionally, for run time reasons, we capped the size of the vocabulary for the orthographic variant to $200k$ words per language.

Minimal Supervision Source For every language we extracted inflection tables for 5 lexemes from UniMorph (Kirov et al., 2018). We aimed for lexemes that are both frequently used, to have robust embedded representations, and from diverse classes, to capture as much as possible of the linguistic behaviors of the language. To this end we targeted lexemes with the largest amount of forms appearing in the embeddings' vocabulary and manually selected from them such that they belong to as diverse classes as possible.[8]

The Inflection Model As our supervised inflection model, to be trained on the bootstrapped data, we used an out-of-the-box sequence-to-sequence LSTM model with attention over the input's characters and the features of the source and target forms. We used the publicly available model of Silfverberg and Hulden (2018). We trained the model for 50 epochs, on either up to $10,000$ data training examples outputted by our system in the minimally supervised scenario, or on $6,000$ training examples in the supervised scenario. The latter provides an upper-bound for the performance of our minimally supervised system. Silfverberg and Hulden (2018) provided train and test data for 4 of the languages examined here, and we hand-crafted similar data sets for English, Russian, Hungarian and Turkish.

The model's evaluation metric is exact match of the outputted string to the gold output.

Models, Baseline and Skyline We report results for the three system variants we tested:

- ORTH: The orthography-based system (§3.1.1);
- SEM: The embedings-based system (Sec. 3.1.2);
- COMB: The combined system (Sec. 3.1.3).

In addition, in order to assess the added value in our systems we include accuracy for two baseline models trained in the low-resourced setting without finding new exmaples:

- OVERFIT: Our sequence-to-sequence inflection model, based on Silfverberg and Hulden (2018).
- CA: The neural transition-based model model of Makarov and Clematide (2018).[9]

We also include the performance of a model trained in a fully-supervised fashion (SUP) as an upper-bound.

4.2 Results

Table 1 summarizes the inflection accuracies for all models. Although the results vary widely across languages, COMB consistently achieves best or near-best results. While COMB outperforms the other two variants in Finnish, Hungarian and Turkish, its performance is roughly on par with ORTH for the five Indo-European (IE) languages in our selection, pointing to the marginal role semantics played in those languages. Both COMB and ORTH outperform CA, our stronger baseline.

Comparing our best system to the SUP skyline, it seems that the room for improvement is bigger for

[6] This last criterion happens to be suitable for all 8 languages examined. Admittedly there are languages, like Czech, that allow less sonorant syllable nucleus.

[7] To reduce run time the Finnish and Hungarian vocabularies were further reduced to include only words that appeared at least twice in the Finnish Parole and Hungarian Gigaword corpora (Bartis, 1998; Oravecz et al., 2014).

[8] Cases of syncretism were trivially solved by merging 2 categories into one category with 2 tags if all supervision lexemes had the same forms in those 2 categories. This strategy might pose a problem in cases of *partial* syncretism, where categories differ in forms only in classes that happen to be absent from the minimal supervision. The manually enforced class diversity in the selection process was devised to solve this problem as well.

[9] This model was designed for low-resourced settings and achieved state of the art results in SIGMORPHON's 2018 shared-task on inflection from lemma (Cotterell et al., 2018). We adapted it to allow for inflection from an arbitrary form.

Inflection	Average	Eng	Rus	Lav	Fra	Fin	Spa	Tur	Hun
Overfit	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.00
CA	0.18	0.38	0.30	0.15	0.06	0.06	0.02	0.22	0.23
Orth	0.40	0.86	**0.75**	**0.49**	0.31	0.13	0.50	0.32	0.01
Sem	0.09	0.01	0.33	0.01	0.07	0.00	0.24	0.03	0.00
Comb	**0.54**	**0.90**	0.73	0.48	**0.32**	**0.48**	**0.53**	**0.49**	**0.48**
Supervised	0.93	0.95	0.93	0.86	0.84	0.95	0.95	0.94	0.98
Paradigm size		5 (8)	15 (15)	27 (34)	32 (48)	39 (39)	56 (65)	30 (30)	48 (49)

Table 1: Morphological inflection accuracies for all languages and systems. Paradigm sizes in number of forms (functions) is also included for reference. The best minimally-supervised results are in **bold**.

	Avg	Eng	Rus	Fin	Spa	Tur
CoNLL17	0.49	0.71	0.37	0.25	**0.76**	0.35
PCS	0.31	0.74	0.31	0.06	0.31	0.12
Comb	**0.62**	**0.92**	**0.60**	**0.51**	0.57	**0.49**

Table 2: Comparison between our best system (Comb), Jin et al. (2020)'s best system per-language (PCS) and their skyline (CoNLL17) on the inflection-from-lemma task.

languages with bigger paradigms (paradigm sizes are indicated in the table, both in number of forms and number of functions). Impressively, the results for English fall short only by 5 percentage points from the fully-supervised upper-bound.

In terms of tagging accuracy, outputted data sets are quite invariably precise across most languages and models (details in the supplementary material). The success of the inflection model seem to be correlated with the amount of words tagged by the first module, rather than on it's precision. The Pearson correlation between the inflection accuracy and the log of the averaged tagged amount per paradigm cell is 0.87.

To exemplify the added value of our minimally-supervised scenario we provide in Table 2 a comparison with the completely unsupervised model of Jin et al. (2020). We compare their best model to our best model (Comb) applied on inflection from lemma. We provided our model with the same 100 lemmas in their test set, and tested our model's capability to complete the respective inflection tables. For most languages, our model's inflection accuracy surpasses even the skyline named by Jin et al. (2020), an edits-based minimally-supervised algorithm that uses 10 inflection tables, while we are using only 5.

4.3 Analysis

The results on our selection of languages, suggest a clear division between Indo-European (IE) languages and non-IE ones. In the former, adding semantic knowledge yields minute improvements at best, while in the latter, Comb clearly outperforms over Orth. We conjuncture that this is because the IE languages in our selection exhibit a fairly simple morpho-phonological system, with relatively few classes and almost no phonological stem-suffix interaction. In contrast, all non-IE languages selected exhibit vowel harmony that multiply the edits related to a single morphological relation, in addition to consonant gradation in the case of Finnish.

This difficulty is magnified with a large amount of classes, as in the Finnish verbal paradigm that includes 27 classes[10] of which about a dozen seem to include more than a few lexemes according to the statistics over Wiktionary entries.[11]

Another imbalance in the results is the inverse correlation between the size of the paradigm and the performance of the inflector trained over the outputted data. We speculate that the effect arises from the fact that languages with bigger paradigms include inflections for functions that are in *exceedingly rare* use, either because they are considered archaic or literary, or because they are used for functions that are far less common. It means that the *data-driven* vocabulary is likely to include less forms with those features, or miss them completely. It also means that these forms will have noisier vector embeddings.

To probe this conjecture empirically, we plotted the Spanish inflection by morphological category against the number of forms found in FastText's vocabulary (Fig. 1). The figure includes results for both best systems in terms of inflection accuracy, namely Orth and Comb. It shows that on common forms the inflection accuracy is on par with the performance on small-paradigm languages, while the rarer forms are mostly the ones driving the total

[10] according to the Research Institute for the Languages of Finland.

[11] https://en.wiktionary.org/wiki/Appendix:Finnish_conjugation

Inflection +hallucination	Average	ENG	RUS	LAV	FRA	FIN	SPA	TUR	HUN
OVERFIT	0.10 (+9)	0.08 (+8)	0.07 (+7)	0.21 (+20)	0.04 (+3)	0.04 (+4)	0.10 (+10)	0.19 (+18)	0.10 (+10)
CA	0.34 (+16)	0.46 (+8)	0.40 (+10)	0.41 (+26)	0.22 (+16)	0.24 (+18)	0.21 (+19)	0.38 (+16)	0.40 (+17)
COMB	0.64 (+9)	0.87 (-3)	0.71 (-2)	0.57 (+9)	0.38 (+6)	0.54 (+6)	0.73 (+20)	0.62 (+13)	0.70 (+22)

Table 3: Results of the baselines and of our best model COMB when 5000 hallucinated examples are added. In parenthesis are the differences comparing to the relevant result in Table 1.

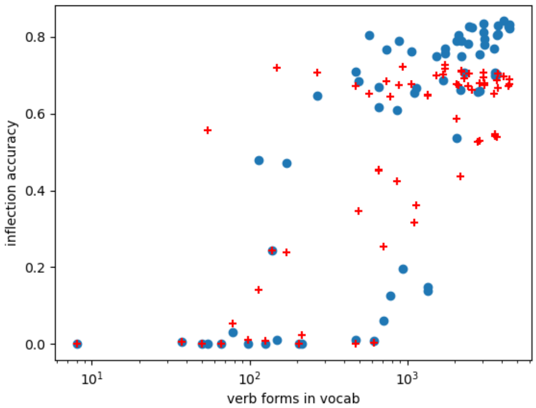

Figure 1: Spanish inflection accuracy per morphological category as a function of the category's abundance in the vocabulary. Plotted results for ORTH (+) and COMB (●).

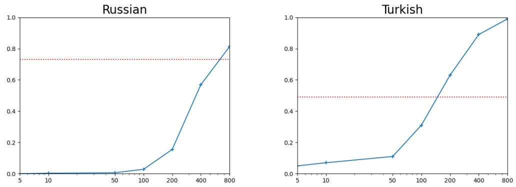

Figure 2: Learning curves of the supervised model in 4 of the languages: Russian and Turkish, with our best performance marked by the dashed line. Plots for all languages can be found in the supplementary material

accuracy down.[12]

In Fig. 2 we display the amount of annotation labor saved by using our model for a subset of the languages. We compare the performance of COMB to models trained on increasing amounts of inflection tables, and we show that our model harnessed 5 annotated inflection tables to output performance equivalent to over 400 lexemes needed in a supervised scenario for most languages. That's a reduction of two orders of magnitude in the labor needed for annotation. This may lead to more sophisticated annotation procedures that could capitalize on the bootstrapping approach proposed in this paper, to save time and efforts in creating morphological resources needed for many languages.

Data Hallucination The methods introduced here to combat the annotation bottleneck are somewhat orthogonal to the method of data hallucination introduced in Anastasopoulos and Neubig (2019), as we aim for identifying new *real* examples from a vocabulary, rather than hallucinate nonce ones. Nonetheless we added 5000 hallucinated examples to both baselines and to our best system (COMB) to assess the combined power of both methods (Table 3). Unsurprisingly, we found that hallucinated data helped improving the results of the baselines, since the hallucinated data is the sole source of new examples. However, this source of example is also quite noisy, and we see that for COMB adding hallucinated examples marginally harms the results for English and Russian, those languages with best COMB performance on our bootstrapped data. Furthermore, note that even with the hallucinated examples, the baselines still underperform compared to COMB models *without* hallucination.

Error Analysis To understand how it might be possible to improve the method further, we sampled 100 incorrect examples from the data set created using the COMB system for Spanish. We found that 64% of the mistakes were of words tagged with only 1 incorrect feature in the bundle. This suggests that a fine-grained algorithmic approach, that will tackle relations between individual features rather than complete sets, might do better in this regard.

We examined the morpho-phonological patterns that appear in the outputs of both COMB and ORTH for Finnish to better assess the reason for the gap in performance between them. We found that the data provided by COMB contains more examples for alternation unattested in the seed data comparing to ORTH. For example, the data from COMB contained 19 examples for the nt∼nn alternation and

[12] This analysis was possible only for Spanish as all other highly-inflectional languages have partial and skewed data on UniMorph, so automatically counting forms is impossible.

Inflection	Average	ENG	RUS	LAV	FRA	FIN	SPA	TUR	HUN
ORTH	0.25 (-15)	0.34 (-52)	0.52 (-23)	0.49 (+0)	0.07 (-26)	0.13 (+0)	0.10 (-40)	0.25 (-7)	0.12 (+11)
SEM	0.05 (-4)	0.00 (-1)	0.20 (-13)	0.01 (+0)	0.00 (-7)	0.00 (+0)	0.05 (-19)	0.13 (+10)	0.00 (+0)
COMB	0.49 (-6)	0.87 (-3)	0.59 (-16)	0.48 (+0)	0.15 (-17)	0.48 (+0)	0.51 (-2)	0.49 (+0)	0.35 (-13)

Table 4: Results of our models when only frequency is considered in seed selection. In parenthesis are the differences comparing to the relevant result in Table 1.

2 examples for rt∼rr, while the data from ORTH contained no examples for both. For comparison, both datasets contained over 100 examples for the attested t∼tt alternation. We conclude that while both methods find examples for attested morphophonological processes, only COMB can close the gap on unattested processes and provide a more diverse dataset for better generalization.

Seed Selection As any bootstrapping method, our algorithms results' may be vulnerable to the selection of the minimal supervision set. To that end we purposefully aimed for frequent and diverse selection (Sec. 4.1). To examine the importance of the selection strategy we altered it to disregard class membership, and we automatically selected lexemes to maximize only frequency. The results of this experiment are in Table 4.

The results show that the selection procedure is indeed important as all different algorithms suffered a loss in performance. It is also evident that diversity is particularly crucial for the orthographic algorithm that is based on the different edit scripts and has less evidence when seed examples cover less classes.

5 Related Work

In recent years, neural sequence-to-sequence models have taken the lead in all forms of morphological tagging and morphological inflection tasks.

Morphological tagging is an *analysis* task where the input is a complete sentence, i.e., a sequence of word forms, and the model aims to assign each word form in-context a morphological signature that consists of its lemma, part-of-speech (POS) tag, and a set of inflectional features (Hajic, 2000; Mueller et al., 2013; Bohnet et al., 2018).

Morphological inflection works in the opposite direction, and may be viewed as a *generation* task. Here, forms of a lexeme are generated from one another given sets of inflectional features of both the input and output. In many implementations the input form is the lemma, in which case the inflectional features of the input are not given (Faruqui et al., 2016; Aharoni and Goldberg, 2017), and the lemma can be either spelled-out, or inputted as an index in a dictionary (Malouf, 2017).[13]

While most pioneering models for *supervised* morphological inflection used statistical models based on finite-state-machines (Kaplan and Kay, 1994; Eisner, 2002), nowadays neural models for morphological inflections are a lot more pervasive (Cotterell et al., 2016, 2017, 2018) (and they go as back as Rumelhart and McClelland, 1986).

In the case of *unsupervised* learning of morphology, a key task is to induce complete paradigms from unlabled texts. Early works on unsupervised morphology induction focused on morpheme segmentation for *concatenative* morphology (Goldsmith, 2001; Creutz and Lagus, 2002; Narasimhan et al., 2015; Bergmanis and Goldwater, 2017). Notwithstanding, early unsupervised works that are *not* limited to concatenative morphology do exist (Yarowsky and Wicentowski, 2000).

More recent studies on unsupervised morphology include works on knowledge transfer within a genealogical linguistic family well- to low-resourced languages (Kann et al., 2017, 2020; McCarthy et al., 2019), as well as works aimed at modifying the approaches for the supervised problem, to allow better tackling of low resourced scenarios. These include Bergmanis et al. (2017) that focused on data augmentation, and Anastasopoulos and Neubig (2019) that modified the model itself with a separate features encoder and introduced the now commonly-used hallucination method for data augmentation. The state of the art model for classic low-resourced scenarios, with a diverse but small dataset, is the transition-based model of Makarov and Clematide (2018). In addition, some works deal with other low-resourced scenarios and assume no inflection tables at all, and are focused on paradigm detection/completion in addition to inflection (Dreyer and Eisner, 2011; Elsner et al., 2019; Jin et al., 2020). In contract, our scenario provides the knowledge on the paradigmatic structure

[13] When inflection is done from a form other than the lemma, it is sometimes referred to as "reinflection".

with a small and undiverse supervision set.

Another work that made use of semantics is that of Soricut and Och (2015), who employed analogies between embedded words to *filter* candidates affixation rules. We use embeddings to *discover* examples in a more general morphological scenario.

6 Conclusions

In this work we propose a realistic minimally-supervised scenario for morphological inflection, which includes only a handful of labeled inflection tables as well as a large bulk of unlabeled text. We showed that semantic and orthographic regularities allow bootstrapping the minimal supervision set to a large (noisy) labeled data set, by searching for word pairs in the vocabulary analogous to observed form pairs from the supervision. We demonstrate that training a neural morphological inflector over the bootstrapped dataset leads to some non-trivial successes, especially on paradigms of smaller size and on commonly-used inflections. This contribution is orthogonal and can be applied in tandem with hallucination approaches. When applied separately, our method outperforms both hallucination and current state of the art models for low-resourced settings. In the future we aim to improve performance over larger paradigms and rarer forms in order to make our method a viable substitute for the labor-intensive manual annotation for new languages.

Acknowledgements

We thank Jonathan Berant for the helpful advice and discussion all throughout. We also thank the audience of the BIU-NLP seminar, the 18th SIGMORPHON meeting, and the 1st UniMorph Meeting, for comments and discussion. This research is funded by an ERC-StG Grant 677352 by the European Research Council, and an ISF grant 1739/26 by the Israeli Science Foundation, for which we are grateful.

References

Farrell Ackerman, James Blevins, and Robert Malouf. 2009. *Parts and wholes: Implicative patterns in inflectional paradigms*, pages 54–82.

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Imre Bartis. 1998. The finnish parole corpus. FIN-CLARIN-konsortio, Nykykielten laitos, Helsingin yliopisto.

Toms Bergmanis and Sharon Goldwater. 2017. From segmentation to analyses: a probabilistic model for unsupervised morphology induction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 337–346, Valencia, Spain. Association for Computational Linguistics.

Toms Bergmanis, Katharina Kann, Hinrich Schütze, and Sharon Goldwater. 2017. Training data augmentation for low-resource morphological inflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 31–39, Vancouver. Association for Computational Linguistics.

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652, Melbourne, Australia. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, 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 *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 616–627, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Jason Eisner. 2002. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Micha Elsner, Andrea Sims, Alex Erdmann, Antonio Hernandez, Evan Jaffe, Lifeng Jin, Martha Johnson, Shuan Karim, David King, Luana Nunes, Byung-Doh Oh, Nathan Rasmussen, Cory Shain, Stephanie Antetomaso, Kendra Dickinson, Noah Diewald, Michelle McKenzie, and Symon Stevens-Guille. 2019. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, 7.

Alexander Erdmann, Micha Elsner, Shijie Wu, Ryan Cotterell, and Nizar Habash. 2020. The paradigm discovery problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7778–7790, Online. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California. Association for Computational Linguistics.

J. R. Firth. 1957. *A synopsis of linguistic theory*. The Philological Society, Oxford.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jan Hajic. 2000. Morphological tagging: Data vs. dictionaries. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707, Online. Association for Computational Linguistics.

Katharina Kann, Samuel R Bowman, and Kyunghyun Cho. 2020. Learning to learn morphological inflection for resource-poor languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(5), pages 8058–8065.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. One-shot neural cross-lingual transfer for paradigm completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1993–2003, Vancouver, Canada. Association for Computational Linguistics.

Ronald M Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational linguistics*, 20(3):331–378.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Bofang Li, Aleksandr Drozd, Tao Liu, and Xiaoyong Du. 2018. Subword-level composition functions for learning word embeddings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 38–48, New Orleans. Association for Computational Linguistics.

Peter Makarov and Simon Clematide. 2018. Neural transition-based string transduction for limited-resource setting in morphology. In *Proceedings of the 27th International Conference on Computational*

Linguistics, pages 83–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Robert Malouf. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 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*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. 2015. An unsupervised method for uncovering morphological chains. *Transactions of the Association for Computational Linguistics*, 3:157–167.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1719–1723, Reykjavik, Iceland. European Language Resources Association (ELRA).

David E Rumelhart and James L McClelland. 1986. On learning the past tenses of english verbs.

Miikka Silfverberg and Mans Hulden. 2018. An encoder-decoder approach to the paradigm cell filling problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2883–2889, Brussels, Belgium. Association for Computational Linguistics.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1627–1637, Denver, Colorado. Association for Computational Linguistics.

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 207–216, Hong Kong. Association for Computational Linguistics.