Neural Multi-Source Morphological Reinflection

Katharina Kann
CIS
LMU Munich, Germany
kann@cis.lmu.de

Ryan Cotterell
Department of Computer Science
Johns Hopkins University, USA
ryan.cotterell@jhu.edu

Hinrich Schütze
CIS
LMU Munich, Germany
inquiries@cislmu.org

Abstract

We explore the task of multi-source morphological reinflection, which generalizes the standard, single-source version. The input consists of (i) a target tag and (ii) multiple pairs of source form and source tag for a lemma. The motivation is that it is beneficial to have access to more than one source form since different source forms can provide complementary information, e.g., different stems. We further present a novel extension to the encoder-decoder recurrent neural architecture, consisting of multiple encoders, to better solve the task. We show that our new architecture outperforms single-source reinflection models and publish our dataset for multi-source morphological reinflection to facilitate future research.

1 Introduction

Morphologically rich languages still constitute a challenge for natural language processing (NLP). The increased data sparsity caused by highly inflected word forms in certain languages causes otherwise state-of-the-art systems to perform worse in standard tasks, e.g., parsing (Ballesteros et al., 2015) and machine translation (Bojar et al., 2016). To create systems whose performance is not deterred by complex morphology, the development of NLP tools for the generation and analysis of morphological forms is crucial. Indeed, these considerations have motivated a great deal of recent work on the topic (Ahlberg et al., 2015; Dreyer, 2011; Nicolai et al., 2015).

In the area of generation, the most natural task is morphological inflection—finding an inflected form for a given target tag and lemma. An example for English is as follows: (`trg:3rdSgPres`, *bring*) $\mapsto$ *brings*. In this case, the 3rd person singular present tense of *bring* is generated. One generalization of inflection is morphological *re*inflection (MRI) (Cotterell et al., 2016), where we must produce an inflected form from a triple of target tag, source form and source tag. The inflection task is the special case where the source form is the lemma. As an example, we may again consider generating the English past tense form from the 3rd person singular present: (`trg:3rdSgPres`, *brought*, `src:Past`) $\mapsto$ *brings* (where `trg` = "target tag" and `src` = "source tag"). As the starting point varies, MRI is more difficult than morphological inflection and exhibits more data sparsity. However, it is also more widely applicable since lexical resources are not always complete and, thus, the lemma is not always available. A more complex German example is given in Table 1.

	Present Ind		Past Ind		Past Sbj	
	Sg	Pl	Sg	Pl	Sg	Pl
1	treffe	treffen	traf	trafen	träfe	träfen
2	triffst	trefft	trafst	traft	träfest	träfet
3	trifft	treffen	traf	trafen	träfe	träfen

Table 1: The paradigm of the strong German verb TREFFEN, which exhibits an irregular ablaut pattern. Different parts of the paradigm make use of one of four bolded theme vowels: e, **i**, **a** or **ä**. In a sense, the verbal paradigm is partitioned into subparadigms. To see why multi-source models could help in this case, starting only from the infinitive treffen makes it difficult to predict subjunctive form träfest, but the additional information of the fellow subjunctive form träfe makes the task easier.

In this work, we generalize the MRI task to a multi-source setup. Instead of using a single source form-tag pair, we use *multiple* source form-tag pairs. Our motivation is that (i) it is often beneficial to have access to more than one source form since different source forms can provide complementary information, e.g., different stems; and (ii)

in many application scenarios, we will have encountered more than one form of a paradigm at the point when we want to generate a new form.

We will make the intuition that multiple source forms provide complementary information precise in the next section, but first return to the English verb *bring*. Generating the form *brings* from *brought* may be tricky—there is an irregular vowel shift. However, if we had a second form with the same theme vowel, e.g., *bringing*, the task would be much easier, i.e., (`trg`:3rdSgPres, `form1`:*brought*, `src1`:Past, `form2`:*bringing*, `src2`:Gerund). A multi-source approach clearly is advantageous for this case since mapping *bringing* to *brings* is regular even though the verb itself is irregular.

The contributions of the paper are as follows. (i) We define the task of multi-source MRI, a generalization of single-source MRI. (ii) We show that a multi-source MRI system, implemented as a novel encoder-decoder, outperforms the top-performing system in the SIGMORPHON 2016 Shared Task on Morphological Reinflection on seven out of eight languages, when given an additional source form. (iii) We release our data to support the development of new systems for MRI.

2 The Task: Multi-Source Reinflection

Previous work on morphological reinflection has assumed a single source form, i.e., an input consisting of exactly one inflected source form (potentially the lemma) and the corresponding morphological tag. The output is generated from this input. In contrast, multi-source reinflection (MRI), the task we introduce, is a generalization in which the model receives multiple form-tag pairs. In effect, this gives the model a partially annotated paradigm from which it predicts the rest.

MRI is a more natural problem than single-source morphological reinflection since we often have access to more than just one form.[1] For example, corpora such as the universal dependency corpus (McDonald et al., 2013) that are annotated on the token level with inflectional features often contain several different inflected forms of a lemma. Such corpora would provide an ideal source of data for the multi-source MRI task.

Formally, we can think of a morphological paradigm as follows. Let Σ be a discrete alphabet for a given language and $\mathcal{T}$ be the set of morphological tags in the language. The inflectional table or morphological paradigm π of a lemma w can be formalized as a set of pairs:

$$\pi(w) = \{(f_1, t_1), (f_2, t_2), \ldots, (f_N, t_N)\}, \quad (1)$$

where $f_i \in \Sigma^+$ is an inflected form of w, and $t_i \in \mathcal{T}$ is the morphological tag of the form f_i. The integer N is the number of slots in the paradigm that have the syntactic category (POS) of w.

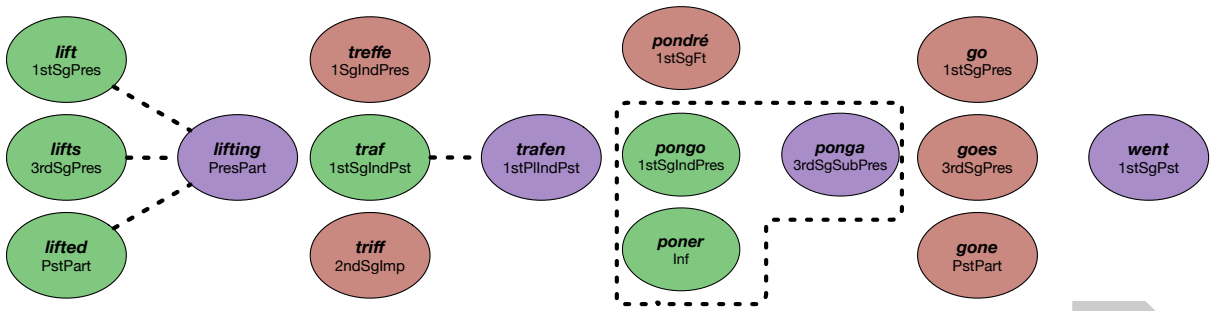
Using this notation, single-source morphological reinflection (MRI) can be described as follows. Given a target tag and a pair of source form and source tag $(t_{\text{trg}}, (f_{\text{src}}, t_{\text{src}}))$ as input, predict the target form f_{trg}. There has been a substantial amount of prior work on this task, including systems that participated in Task 2 of the SIGMORPHON 2016 shared task (Cotterell et al., 2016). Thus, we may define the task of *multi-source morphological reinflection* as follows: Given a target tag and a set of k form-tag source pairs $(t_{\text{trg}}, \{(f^1_{\text{src}}, t^1_{\text{src}}), \ldots, (f^k_{\text{src}}, t^k_{\text{src}})\})$ as input, predict the target form f_{trg}. Note that single-source MRI is a special case of multi-source MRI for $k = 1$.

2.1 Motivating Examples

Figure 1 gives examples for four different configurations that can occur in multi-source MRI.[2] We have colored the source forms green and drawn a dotted line to the target if they contain sufficient information for correct generation. If two source forms together are needed, the dotted line encloses both of them. Source forms that provide no information in the configuration are colored red (no arrow); note these forms could provide (and in most cases will provide) useful information for other combinations of source and target forms.

[1] Scenarios where a single form is available and that form is the lemma are perhaps not infrequent. In high-resource languages, an electronic dictionary may have near-complete coverage of the lemmata of the language. However, paradigm completion is especially crucial for neologisms and low-resource languages.

[2] Figure 1 is not intended as a complete taxonomy of possible MRI configurations, e.g., there are hybrids of ANYFORM and NOFORM (some forms are informative, others are suppletive) and fuzzy variants (a single form gives pretty good evidence for how to generate the target form, but another single form gives better evidence). All of our examples make additional assumptions, e.g., that we have not seen other similar forms in training either of the same lemma (*siente*) or of a similar lemma (*consientes*). Hopefully, the examples are illustrative of the main conceptual distinction: several single forms each are sufficient by themselves (ANYFORM), a single, but a carefully selected form is sufficient (SINGLEFORM), multiple forms are needed to generate the target (MULTIFORM) and the target form cannot be predicted (irregular) from the source forms (NOFORM).



(a) ANYFORM (b) SINGLEFORM (c) MULTIFORM (d) NOFORM

Figure 1: Four possible input configurations in multi-source morphological reinflection (MRI). In each subfigure, the target form on the right is **purple**. The source forms are on the left and are **green** if they can be used to predict the target form (also connected with a dotted line) and **red** if they cannot. There are four possible configurations: (i) ANYFORM is the case where one can predict the target form from any of the source forms. (ii) SINGLEFORM is the case where only one form can be used to regularly predict the target form. (iii) MULTIFORM is the case where multiple forms are *necessary* to predict the target form. (iv) NOFORM is the case where the target form cannot be regularly derived from any of the source forms. Multi-source MRI is expected to perform better than single-source MRI for the configurations SINGLEFORM and MULTIFORM, but not for the configurations ANYFORM and NOFORM.

The first type of configuration is ANYFORM: each of the available source forms in the subset of the English paradigm (*lift*, *lifts*, *lifted*) contains enough information for a correct generation of the target form *lifting*. The second configuration is SINGLEFORM: there is a single form that contains enough information for correct generation, but it has to be carefully selected. Inflected forms of the German verb *treffen* 'to meet' have different stem vowels (see Table 1). In single-source reinflection, producing a target form with one stem vowel (*a* in *trafe* in the figure) from a source form with another stem vowel (e.g., *e* in *treffe*) is difficult.[3]

In contrast, the learning problem for the SINGLEFORM configuration is much easier in multi-source MRI. The multi-source model does not have to learn the possible vowel changes of this irregular verb; instead, it just needs to pick the correct vowel change from the alternatives offered in the input. This is a relatively easy task since the theme vowel is identical. So we only need to learn one general fact about German morphology (which suffix to add) and will then be able to produce the correct form with high accuracy. This type of regularity is typical of complex morphology: there are groups of forms in a paradigm that are similar and it is highly predictable which of these groups a particular target form for a new word will be a member of. As long as one representative of each group is part of the multi-source input, we can select it to generate the correct form.

[3]It is not impossible to learn, but *treffen* is an irregular verb, so we cannot easily leverage the morphology we have learned about other verbs.

In the MULTISOURCE configuration, we are able to use information from multiple forms if no single form is sufficient by itself. For example, to generate *ponga*, 3rdSgSubPres of *poner* 'to put' in Spanish, we need to know what the stem is (*ponga*, not *pona*) and which conjugation class (*-ir*, *-er* or *-ar*) it is part of (*ponga*, not *pongue*). The single-source input *pongo*, 1stSgIndPres, does not reveal the conjugation class: it is compatible with both *ponga* and *pongue*. The single-source input *poner*, Inf, does not reveal the stem for the subjunctive: it is compatible with both *ponga* and *pona*—we need both source forms to generate the correct form *ponga*.

Again, such configurations are frequent cross-linguistically, either in this "discrete" variant or in more fuzzy variants where taking several forms together increases our chances of producing the correct target form. Finally, we call configurations NOFORM if the target form is completely irregular and not related to any of the source forms. The suppletive form *went* is our example for this case.

2.2 Principle Parts

The intuition behind the MRI task draws inspiration from the theoretical linguistic notion of **principle parts** (Finkel and Stump, 2007; Stump and Finkel, 2013). The notion is that a paradigm has a subset that allows for maximum predictability. In terms of language pedagogy, the principle parts would be a minimial set of forms a student has to learn in order to be able to generate any form in the paradigm. For instance for the partial German paradigm in Table 1, the forms *treffen*, *trifft*,

trafen, and *träfen* could form *one* potential set of principle parts.

From a computational learning point of view, maximizing predictability is always a boon—we want to make it as easy as possible for the system to learn the morphological regularities and subregularities of the language. Giving the system the principle parts as input is one way to achieve this.

3 Model Description

Our model is a multi-source extension of MED, Kann and Schütze (2016b)'s encoder-decoder network for MRI. In MED, a single bidirectional recurrent neural network (RNN) encodes the input. In contrast, we use multiple encoders to be able to handle multiple source form-tag pairs. In MED, a decoder RNN produces the output from the hidden representation. We do not change this part of the architecture, so there is still a single decoder.

3.1 Input and Output Format

For k source forms, our model takes k different inputs of parallel structure. Each of the $1 \leq i \leq k$ inputs consists of the target tag and the source form f_i and its corresponding source tag t_i. The output is the target form. Each source form is represented as a sequence of characters; each character is represented as an embedding. Each tag—both the target tag and the source tags—is represented as a sequence of subtags; each subtag is represented as an embedding.

More formally, we define the alphabet Σ_{lang} as the set of characters in the language and Σ_{subtag} as the set of subtags that occur as part of the set of morphological tags $\mathcal{T}$ of the language; e.g., if 1st-SgPres $\in \mathcal{T}$, then 1st, Sg and Pres $\in \Sigma_{\text{subtag}}$. Each of the k inputs to our system is of the following format: $S_{\text{start}}\Sigma^+_{\text{subtag}}\Sigma^+_{\text{lang}}\Sigma^+_{\text{subtag}}S_{\text{end}}$ where the first subtag sequence is the source tag t_i and the second subtag sequence is the target tag. The output format is: $S_{\text{start}}\Sigma^+_{\text{lang}}S_{\text{end}}$, where the symbols S_{start} and S_{end} are predefined start and end symbols.

3.2 Multi-Source Encoder-Decoder

The encoder-decoder is based on the machine translation model of Bahdanau et al. (2014) and all specifics of our model are identical to the original presentation unless stated otherwise. Whereas in model of Bahdanau et al. (2014), there is only one encoder, our model consists of $k \geq 1$ encoders and processes k sources simultaneously. The k sources have the form $X_m = (t_{\text{trg}}, f^m_{\text{src}}, t^m_{\text{src}})$, represented as $S_{\text{start}}\Sigma^+_{\text{subtag}}\Sigma^+_{\text{lang}}\Sigma^+_{\text{subtag}}S_{\text{end}}$ as described above. Characters and subtags are embedded.

The input to encoder m is X_m. Each encoder consists of a bidirectional RNN that computes a hidden state h_{mi} for each position, the concatenation of forward and backward hidden states. Decoding proceeds as follows:

$$p(y|X_1,\ldots,X_k) = \prod_{t=1}^{|Y|} p(y_t|\{y_1,...,y_{t-1}\}, c_t) = \prod_{t=1}^{|Y|} g(y_{t-1}, s_t, c_t), \quad (2)$$

where $y = (y_1,...,y_{|Y|})$ is the output sequence (a sequence of $|Y|$ characters), g is a nonlinear function, s_t is the hidden state of the decoder and c_t is the sum of the encoder states h_{mi}, weighted by attention weights $\alpha_{mi}(s_{t-1})$ that depend on the decoder state:

$$c_t = \sum_{m=1}^{k}\sum_{i=1}^{|X_m|} \alpha_{mi}(s_{t-1})h_{mi}. \quad (3)$$

A visual depiction of this model may be found in Figure 2. A more complex hierarchical attention structure would be an alternative, but this simple model in which all hidden states contribute on the same level in a single attention layer (i.e., $\sum_{m=1}^{k}\sum_{i=1}^{|X_m|}\alpha_{mi} = 1$) works well as our experimental results. The k encoders share their weights.

4 Multi-Source Reinflection Experiment

We evaluate the performance of our model in an experiment based on Task 2 of the SIGMORPHON Shared Task on Morphological Reinflection (Cotterell et al., 2016). This is a single-source MRI task as outlined in Section 1.

4.1 Experimental Settings

Datasets. Our datasets are based on the data from the SIGMORPHON 2016 Shared Task on Morphological Reinflection (Cotterell et al., 2016). Our experiments cover eight languages: Arabic, Finnish, Georgian, German, Hungarian, Russian, Spanish and Turkish. The languages were chosen to represent different types of morphology. Finnish, German, Hungarian, Russian, Turkish and Spanish are all suffixing. In addition to being suffixing, three of these languages employ vocalic (German, Spanish) and consonantal

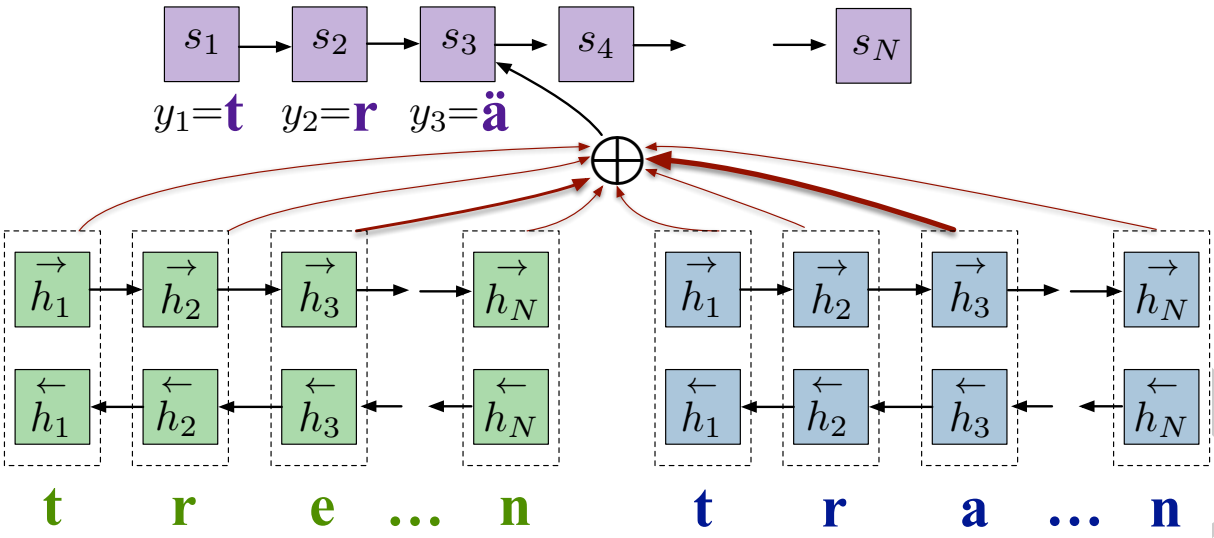


Figure 2: Visual depiction of our multi-source encoder-decoder RNN. We sketch a two encoder model, where the left encoder reads in the present form treffen and the right encoder reads in the past tense form trafen. They work together to predict the subjunctive form träfen. The shadowed red arcs indicate the strength of the attention weights—we see the network is focusing more on **a** because it helps the decoder better predict **ä** than **e**. We omit the source and target tags as input for conciseness.

(Russian) stem changes for many inflections. The members of the remaining sub-group are agglutinative. Georgian makes use of prefixation as well as suffixation. Arabic morphology contains both concatenative and templatic elements. We build multi-source versions of the dataset for Task 2 of the SIGMORPHON shared task in the following way. We use data from the UNIMORPH project,[4] containing complete paradigms for all languages of the shared task. The shared task data was sampled from the same set of paradigms; our new dataset is a superset of the SIGMORPHON data.

We create our new dataset by uniformly sampling three additional word forms from the paradigm of each source form in the original data. In combination with the source and target forms of the original dataset, this means that our dataset is a set of 5-tuples consisting each of four source forms and one target form.[5] Ideally, we would like to keep the experimental variable k, the number of sources we use in multi-source MRI, constant for a particular experiment or vary it systematically across other experimental conditions. Table 2 gives an overview of the number of different source forms per language in our dataset.

[4] http://unimorph.org

[5] One thing to note is that the original shared task data was sampled depending on word frequency in unlabeled corpora. We do not impose a similar condition, so the frequency distributions of our data and the shared task data are different. Also, we excluded Maltese and Navajo due to a lack of data to create the additional multi-source datasets.

	1	2	3	≥ 4
ar	0	0	0	12,800
fi	0	0	0	12,800
ka	1015	84	2	11,699
de	0	0	0	12,800
hu	0	0	0	19,200
ru	0	0	5	12,794
es	1575	25	877	10,323
tu	0	0	0	12,800

Table 2: Number of target forms in the training set for which 1, 2, 3 or ≥ 4 source forms (in the training set) are available for prediction. The tables for the development and test splits show the same pattern and are omitted.

Our dataset is available for download at http://cistern.cis.lmu.de.

Hyperparameters. We use embeddings of size 300. Our encoder and decoder GRUs have 100 hidden units each. Following Le et al. (2015), we initialize all encoder and decoder weights as well as the embeddings with an identity matrix. All biases are initialized with zero. We use stochastic gradient descent, Adadelta (Zeiler, 2012) and a minibatch size of 20 for training. Training is done for a maximum number of 90 epochs. If no improvement occurs for 20 epochs, we stop training early. The final model we run on test is the model that performs best on the development data.

Baselines. For the single-source case, we apply MED, the top-scoring system in the SIGMOR-

	source form(s) used					
	1	2	3	4	1–2	1–4
ar	.871	.813	.796	.830	.905	**.944**
fi	.956	.929	.941	.934	.965	**.978**
ka	.967	.943	.942	.934	.969	**.979**
de	.954	.922	.931	.912	.959	**.980**
hu	**.992**	.962	.963	.963	.988	.989
ru	.876	.795	.824	.817	.888	**.911**
es	.975	.961	.963	.968	.977	**.984**
tu	.967	.928	.947	.944	.970	**.983**

Table 3: Accuracy on MRI for single-source (1, 2, 3, 4) and multi-source (1–2, 1–4) models. Best result in bold.

PHON 2016 Shared Task on Morphological Reinflection (Cotterell et al., 2016; Kann and Schütze, 2016b). At the time of writing, MED constitutes the state of the art on the dataset. For Arabic, German and Turkish, we run an additional set of experiments to test two additional architectural configurations of multi-source encoder-decoders: (i) In addition to the default configuration in which all encoders share parameters, we also test the option of each encoder learning its own set of parameters (shared par's: yes vs. no in Table 4). (ii) Another way of realizing a multi-source system is to concatenate all sources and give this to an encoder-decoder with a single encoder as one input (encoders: $k = 1$ vs. $k > 1$ in Table 4).

Evaluation Metric. We evaluate on 1-best accuracy (exact match) against the gold form. We deviate from the shared task, which also evaluates under mean reciprocal rank and edit distance. We omit the later two, ince all these metrics were highly correlated (Cotterell et al., 2016).

4.2 Results

Table 3 shows the results of the MRI experiment on test data. We compare using a single source, the first two sources and all four sources. The first source (in column "1") is the original source from the SIGMORPHON shared task. Recall that we used uniform sampling to identify additional forms whereas the sampling procedure of the shared task took into account frequency. We suspect that this is the reason for the worse performance of the new sources compared to the original source; e.g., in German there are rarely used subjunctive forms like *befähle* that are unlikely to help generate related forms that are more frequent.

The main result of the experiment is that multi-source MRI performs better than single-source MRI for all languages except for Hungarian and that, clearly, the more sources the better: using four sources is always better than using two sources. This result confirms our hypothesis, illustrated in Figure 1, that for most languages, different source forms provide complementary information when generating a target form and thus performance of the multi-source model is better than of the single-source model. Table 3 demonstrates that the two configurations we identified as promising for multi-source MRI, SINGLEFORM and MULTIFORM, occur frequently enough to boost the performance for seven of the eight languages, with the largest gains observed for Arabic (7.3%) and Russian (3.5%) and the smallest for Spanish (0.9%) and Georgian (2.0%) (comparing using source form 1 with using source forms 1–4).

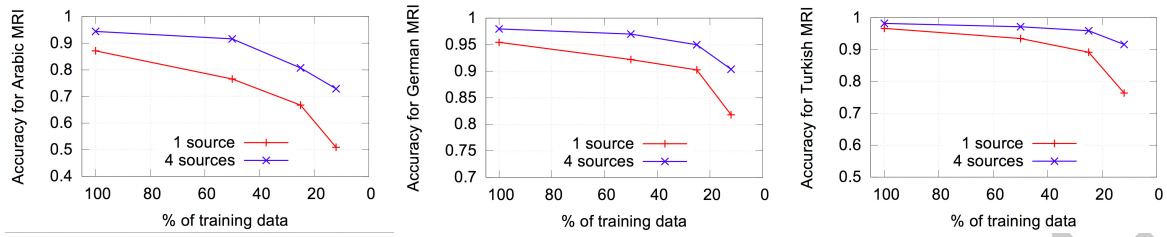
encoders:		$k = 1$	$k = 4$	
par's shared:			yes	no
language	ar	**.944**	**.944**	.920
	de	**.980**	**.980**	.975
	tu	**.985**	.983	.969

Table 4: Results of different architectures for the dataset with 4 source forms being available for prediction. The best result for each row is in bold.

Hungarian is the only language for which performance decreases, by a small amount (.03%). We attribute this to overfitting: the multi-source model has a larger number of parameters, so it is more prone to overfitting. We would expect the performance to be the same in a comparison of two models that have the same size.

Error Analysis. We compare errors of single-source and multi-source models for German on development data. Most mistakes of the multi-source model are stem-related: *versterbst* for *verstirbst*, *erwerben* for *erwürben*, *Apfelsinenbaume* for *Apfelsinenbäume*, *lungenkränkes* for *lungenkrankes* and *übernehmte* for *übernähme*. In most of these cases, the stem of the lemma was used, which is correct for some forms, but not for the form that had to be generated. In one case, the multi-source model did not use the correct inflection rule: *braucht* for *gebraucht*—the inflectional rule that the past participle is formed by *ge-* was not applied.

Errors of the single-source model that were "corrected" by the multi-source model include



(a) Arabic (b) German (c) Turkish

Figure 3: Learning curves for single-source and multi-source models for Arabic, German and Turkish. We observe that the multi-source model generalizes faster than the single soure case—this is to be expected since the multi-source model often faces an easier transduction problem.

empfahlt for *empfiehl*, *Throne* for *Thron* and *befielen* for *befallen*. These are all SINGLEFORM cases: the multi-source model will generate the correct form if it succeeds in selecting the most predictive source form. The single-source model is at a disadvantage if this most predictive source form is not part of its input.

4.3 Comparison of Different Architectures

Table 4 compares different architectural configurations. All experiments use 4 sources. We see that sharing parameters is superior as expected. Using a single encoder on 4 sources performs as well as 4 encoders (and very slightly better on Turkish). Apparently, it has no difficulty learning to understand an unstructured (or rather lightly structured) concatenation of form-tag pairs; on the other hand, this parsing task, i.e., learning to parse the sequence of form-tag pairs, is easy, so this is not a surprising result.

4.4 Learning Curves

Figure 3 shows learning curves for Arabic, German and Turkish. We iteratively halve the training set and train models for each subset. In this analysis, we train all models for 90 epochs, but use the numbers from the main experiment for the full training set. For the single-source model, we use the SIGMORPHON source. The figure shows that the single-source model needs more individual paradigms in the training data to achieve the same performance as the multi-source model. The largest difference between single-source and multi-source is $> 20\%$ for Arabic when only 1/8 of the training set is used. This suggests that multi-source MRI is an attractive option for low-resource languages since it exploits available data better than single-source.

4.5 Attention Visualization

Figure 4 shows for one example, the generation of the German form *wögen*, 3rdPlSubPst, the attention weights of the multi-source model at each time step of the decoder, i.e., for each character as it is being produced by the decoder. For characters that simply need to be copied, the main attention lies on the corresponding characters of the input sources. For example, the character *g* is produced when attention is on the characters *g* in *wögest*, *wöge* and *wogen*. This aspect of the multi-source model is not different from the single-source model, offering no advantage.

However, even for *g*, the source form that is least relevant for generating *wögen* receives almost no weight: *wägst* is an indicative singular form that does not provide helpful information for generating a plural form in the subjunctive; the model seems to have learned that this is the case. In contrast, *wogen* does receive some weight; this makes sense as it is a past indicative form and the past subjunctive is systematically related to the past indicative for many German verbs. These observations suggest that the network has learned to correctly predict (at least in this case) which forms provide potentially useful information. For the last two time steps (i.e., characters to be generated), attention is mainly focused on the tags. Again, this indicates that the model has learned the regularity in generating this part of the word form: the suffix, consisting of *en*, is predictable from the tag.

5 Related Work

Recently, variants of the RNN encoder-decoder have seen widespread adoption in many areas of NLP due to their strong performance. Encoder-decoders with and without attention have been applied to tasks such as machine translation (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et

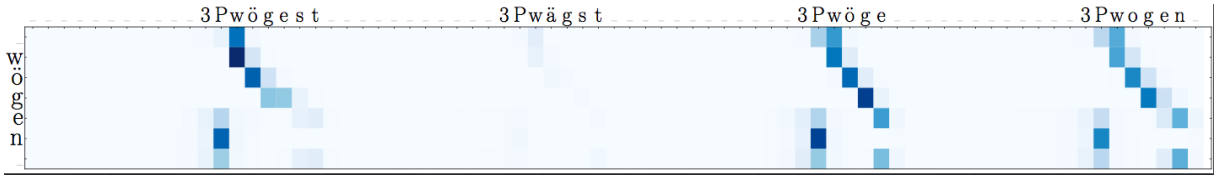


Figure 4: Attention heatmap for the multi-source model. The example is for the German verb *wiegen* 'to weigh'. The model learns to focus most of its attention on forms that share the irregular subjunctive stem *wög* in addition to the target subtags *3* and *P* that encode that the target form is 3rd person plural. We omit the tags from the diagram to which the model hardly attends.

al., 2014), parsing (Vinyals et al., 2015) and automatic speech recognition (Graves and Schmidhuber, 2005; Graves et al., 2013).

The first work on multi-source models was presented for machine translation. Zoph and Knight (2016) made simultaneous use of source sentences in multiple languages in order to find the best match possible in the target language. Unlike our model, they apply transformations to the hidden states of the encoders that are input to the decoder. Firat et al. (2016)'s neural architecture for MT translates from any of N source languages to any of M target languages, using language specific encoders and decoders, but sharing one single attention-mechanism. In contrast to our work, they obtain a single output for each input.

Much ink has been spilled on morphological reinflection over recent years. Dreyer et al. (2008) develop a high-performing weighted finite-state transducer for the task, which was later hybridized with an LSTM (Rastogi et al., 2016). Durrett and DeNero (2013) apply a semi-CRF to heuristically extracted rules to generate inflected forms from lemmata using data scraped from Wikitionary. Improved systems for the Wiktionary data were subsequently developed by Hulden et al. (2014), who used a semi-supervised approach, and Faruqui et al. (2016), who used a character-level LSTM. All of the above worked has focused on the single input case. Two important exceptions, however, have considered the multi-input case. Both Dreyer and Eisner (2009) and Cotterell et al. (2015) define a string-valued graphical model over the paradigm and apply the missing values.

The SIGMORPHON 2016 Shared Task on Morphological Reinflection (Cotterell et al., 2016), based on the UNIMORPH (Sylak-Glassman et al., 2015) data, resulted in the development of numerous different methods. RNN encoder-decoder models (Aharoni et al., 2016; Kann and Schütze, 2016a; Östling, 2016) obtained the strongest performance and are the current state of the art on the task. The best-performing model made use of an attention mechanism (Kann and Schütze, 2016a), first popularized in machine translation (Bahdanau et al., 2014). We generalize this architecture to the multi-source case in this paper for the reinflection task.

6 Conclusion and Future Work

Generation of unknown inflections in morphologically rich languages is an important task that remains unsolved. We provide a new angle on the problem by considering systems that are allowed to have multiple inflected forms as input. To this end, we define the task of multi-source morphological reinflection as a generalization of single-source MRI (Cotterell et al., 2016) and presented a model that solves the task. We extended an attention-based RNN encoder-decoder architecture from the single-source case to the multi-source case. Our new model consists of multiple encoders, each receiving one of the inputs. We showed that our model improves over the state of the art for seven out of eight languages, demonstrating the promise of multi-source MRI. Additionally, we publically released our implementation.[6]

We created and publically released a dataset for multi-source morphological reinflection that is a superset of the dataset of the SIGMORPHON 2016 Shared Task on Morphological Reinflection to facilitate research on morphological generation. One focus of future work should be the construction of more complex datasets, e.g., datasets that have better coverage of irregular words and datasets in which there is no overlap in lemmata between training and test sets. Further, for difficult inflections, it might be interesting to find an effective way to include unsupervised data into the setup. For example, we could define one of our k inputs to be a form mined from a corpus that is not

[6] http://cistern.cis.lmu.de

guaranteed to have been correctly tagged morphologically, but likely to be helpful.

Acknowledgements

The second author was supported by a DAAD Long-Term Research Grant and an NDSEG fellowship.

References

Roee Aharoni, Yoav Goldberg, and Yonatan Belinkov. 2016. Improving sequence to sequence learning for morphological inflection generation: The BIU-MIT systems for the SIGMORPHON 2016 shared task for morphological reinflection. In *2016 Meeting of SIGMORPHON*.

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2015. Paradigm classification in supervised learning of morphology. In *NAACL*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2015. Improved transition-based parsing by modeling characters instead of words with LSTMs. In *EMNLP*.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *WMT*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Ryan Cotterell, Nanyun Peng, and Jason Eisner. 2015. Modeling word forms using latent underlying morphs and phonology. In *TACL*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *2016 Meeting of SIGMORPHON*.

Markus Dreyer and Jason Eisner. 2009. Graphical models over multiple strings. In *EMNLP*.

Markus Dreyer, Jason R Smith, and Jason Eisner. 2008. Latent-variable modeling of string transductions with finite-state methods. In *EMNLP*.

Markus Dreyer. 2011. *A non-parametric model for the discovery of inflectional paradigms from plain text using graphical models over strings*. Ph.D. thesis, Johns Hopkins University, Baltimore, MD.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *HLT-NAACL*.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *NAACL*.

Raphael Finkel and Gregory Stump. 2007. Principal parts and morphological typology. *Morphology*, 17(1):39–75.

Orhan Firat, KyungHyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *CoRR*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610.

Alan Graves, Abdel-Rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *ICASSP*.

Mans Hulden, Markus Forsberg, and Malin Ahlberg. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *EACL*.

Katharina Kann and Hinrich Schütze. 2016a. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *2016 Meeting of SIGMORPHON*.

Katharina Kann and Hinrich Schütze. 2016b. Single-model encoder-decoder with explicit morphological representation for reinflection. In *ACL*.

Quoc V Le, Navdeep Jaitly, and Geoffrey E. Hinton. 2015. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*.

Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *ACL*.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Inflection generation as discriminative string transduction. In *NAACL*.

Robert Östling. 2016. Morphological reinflection with convolutional neural networks. In *2016 Meeting of SIGMORPHON*.

Pushpendre Rastogi, Ryan Cotterell, and Jason Eisner. 2016. Weighting finite-state transductions with neural context. In *NAACL*.

Gregory Stump and Raphael A Finkel. 2013. *Morphological typology: From word to paradigm*, volume 138. Cambridge University Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *ACL-IJCNLP*.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *NIPS*.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *arXiv preprint arXiv:1601.00710*.