Exploring Cross-Lingual Transfer of Morphological Knowledge In Sequence-to-Sequence Models

Huiming Jin
Beihang University, China
huiming.jin.buaa@gmail.com

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

Abstract

Multi-task training is an effective method to mitigate the data sparsity problem. It has recently been applied for cross-lingual transfer learning for paradigm completion—the task of producing inflected forms of lemmata—with sequence-to-sequence networks. However, it is still vague how the model transfers knowledge across languages, as well as if and which information is shared. To investigate this, we propose a set of data-dependent experiments using an existing encoder-decoder recurrent neural network for the task. Our results show that indeed the performance gains surpass a pure regularization effect and that knowledge about language and morphology can be transferred.

1 Introduction

Neural sequence-to-sequence models define the state of the art for paradigm completion (Cotterell et al., 2016, 2017; Kann and Schütze, 2016), the task of generating inflected forms of a lemma's paradigm, e.g., filling the empty fields in Table 1 using one of the non-empty fields.

However, those models are in general very data-hungry, and do not reach good performances in low-resource settings. Therefore, Kann et al. (2017) propose to leverage morphological knowledge from a high-resource language (*source language*) to improve paradigm completion in a closely related language with insufficient resources (*target language*). This is achieved by a form of multi-task learning – they train an encoder-decoder model simultaneously on training examples for both languages. While closer related languages seem to help more than distant ones, the mechanisms *how* this transfer works still remain largely obscure. Several possibilities exist: (i) learning of target tag specific word transformations from the high-resource language (**trans**); (ii) training of the character language model of the decoder (**LM**); (iii) learning a bias to copy a large part of the input (**copy**), since members of the same paradigm mostly share the same stem; (iv) a general regularization effect obtained by multi-task training (**reg**).

	Present		Past	
	Singular	Plural	Singular	Plural
1	*sueño*	*soñamos*	*soñé*	*soñamos*
2	*sueñas*	*???*	*soñaste*	*soñasteis*
3	*sueña*	*sueñan*	*soñó*	*???*

Table 1: Partial inflection table for indicative forms of the Spanish verb *soñar*.

In this work, we intend to shed light on the way cross-lingual transfer learning for paradigm completion with an encoder-decoder model works, and will especially focus on the role of the character and tag embeddings. In particular we aim at answering the following questions: (i) What does the neural model learn from the tags of a high-resource language for the tags of a low-resource language? (ii) Is sharing an alphabet important for the transfer? (iii) How much of the transfer learning can be reduced to a regularization effect achieved by multi-task learning?

For our analysis, we present a set of detailed experiments for the target language Spanish [ES]. Source languages are either members of the Romance language family (Catalan [CA], French [FR], Italian [IT], Portuguese [PT]) of different levels of similarity to Spanish, cf. Table 2, or an unrelated language (Arabic [AR]). We show which parts of the information are learned from the characters or tags and discuss where sequences of letters or tags from a second language contribute to or restrain performance on the paradigm comple-

	PT	CA	IT	FR
similarity to ES	89%	85%	82%	75%

Table 2: Lexical similarities of Spanish and the Romance languages used for our experiments (Lewis, 2009).

tion task in the low-resource language.

2 Transfer Learning for Paradigm Completion

In this section, we describe cross-lingual transfer learning for morphology and the model used for it.

Cross-lingual transfer. Transfer learning for paradigm completion is much more language-specific than most semantic natural language processing tasks, like entity typing or machine translation. An extreme example is the infeasible task of transferring morphological knowledge from Chinese to Portuguese as Chinese does not make use of inflection at all. Even between two morphologically rich languages transfer is difficult if they are unrelated, since inflections often mark dissimilar subcategories and word forms do not share similarities.

However, Kann et al. (2017) show that transferring morphological knowledge from Spanish to Portuguese, two languages with similar morphology and 89% lexical similarity, works well and, more surprisingly, even supposedly very different languages like Arabic and Spanish can benefit from each other. They make this possible by training an encoder-decoder model and appending a special tag (i.e., embedding) for each language to the input of the system, similar to (Johnson et al., 2016). It is currently unclear, though, what the nature of this transfer is, motivating our work which explores this in more detail.

Model description. The model Kann et al. (2017) use and we explore in more detail here is an encoder-decoder recurrent neural network (RNN) with attention (Bahdanau et al., 2015). It is trained on maximizing the following log-likelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{(k,w_{\ell_t})\in\mathcal{D}_t} \log p_{\boldsymbol{\theta}}\left(f_k[w_{\ell_t}]|\ell_t, w_{\ell_t}, t_k\right) + \sum_{(k,w_{\ell_s})\in\mathcal{D}_s} \log p_{\boldsymbol{\theta}}\left(f_k[w_{\ell_s}]|\ell_s, w_{\ell_s}, t_k\right) \quad (1)$$

We denote the source training examples as $\mathcal{D}_s$ and the target training examples as $\mathcal{D}_s$. w_{ℓ_s} represents a lemma in a high-resource source language ℓ_s and w_{ℓ_t} represents a lemma in a low-resource target language ℓ_t. k represents a given slot in the paradigm and $f_k[w_\ell]$ is the inflected form of w_ℓ corresponding to the morphological tag t_k. The parameters θ of the model are tied for both the high-resource language and the low-resource language to enable transfer learning.

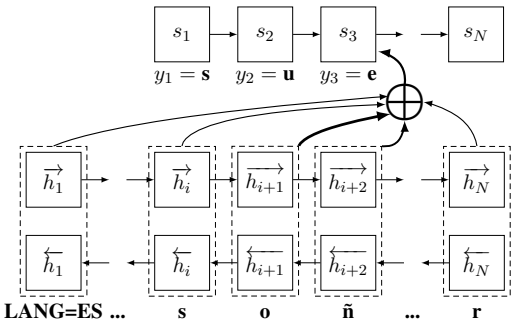

Figure 1: Overview of an encoder-decoder RNN, mapping the Spanish lemma *soñar* to the target form *sueña*. The thickness of the arrows towards the circled plus symbol corresponds to each *attention weight*. All tags in the input are omitted.

In detail, a bidirectional gated RNN is used to *encode* the input sequence, which consists of a language tag, morphological tags and characters of the input language. The decoder generates the output sequence from the characters of the same language, and consists of a unidirectional RNN with an attention mechanism over the encoder hidden states. Notably, the elements of the input and the output are represented by embeddings living in separate spaces.

Hyperparameters. Encoder and decoder RNNs have 100 hidden units and we use 300-dimensional embeddings. We train using ADADELTA (Zeiler, 2012) with minibatch size 20. All models for all experiments are trained for a maximum of 150 epochs. The best model is applied at test time.

3 Exploration of Transfer Learning

In order to answer the questions raised in the introduction, we conduct the following experiments.

3.1 Data

We use the Romance and Arabic language data from Kann et al. (2017). In particular, each training file contains 12,000 high-resource examples mixed with 50 or 200 fixed Spanish instances. We

	trans	LM	copy	reg
l-ciph	X	X		
t-ciph	X			
l-emb	X	X	X	
t-emb	X			

Table 3: Expected effect of different modifications of the high-resource training data. Learning of the marked fields is likely to be influenced, descriptions in the text, cf. §1.

use the same development and test files for all experiments. Arabic is transcribed into Latin characters.

3.2 Experiments

Letter cipher (l-ciph). Let $\mathcal{C} = \mathcal{C}_{low} \cup \mathcal{C}_{high}$ be the union of the sets of all characters in the alphabets of the low-resource language and the high-resource language, respectively.[1] We define a bijective cipher function $f_{ciph} : \mathcal{C} \mapsto \mathcal{C}$, mapping each character to a different character, chosen at random. Then, we apply this function to the elements of the input and output words in the high-resource language and train the model on this modified data. The low-resource samples in train, dev and test remain unchanged.

We expect this to have the following effects: (i) languages do not share affixes anymore; (ii) as we use the same embeddings for the changed and unchanged characters, the model might learn *wrong* affixes for tags; (iii) an incorrect character language model could be learned; and (iv) a general bias to copy should remain unchanged.

Tag cipher (t-ciph). We further consider the union of the sets of all morphological tags existing in the low- and high-resource languages: $\mathcal{T} = \mathcal{T}_{low} \cup \mathcal{T}_{high}$. We define a bijective cipher function $f_{ciph} : \mathcal{T} \mapsto \mathcal{T}$. We then apply this function to all tags in the high-resource language input and train a new model. The low-resource examples in train, dev and test are not changed.

We expect this to: (i) disturb the learning of correspondences between target tags and output characters; (ii) not influence anything else.

Language-dependent letter embeddings (l-emb). We now use different embeddings for the characters of the two languages. This corresponds to a setting where the source and target languages do not share the same vocabulary.

[1]Note that for the languages considered in our experiments we have $\mathcal{C}_{low} \approx \mathcal{C}_{high}$.

This should result in: (i) making it impossible for the model to learn which affixes have to be produced for which tag, maybe resulting in benefits for more distant and worse performance for extremely close languages; and (ii) transfer of the decoder's character language model getting impossible.

Language-dependent tag embedding (t-emb). Additionally, we also experiment with different embeddings for the morphological tags in different languages.

We expect the following to happen: (i) the model can learn a character language model in the output, which might be good for related and bad for more distant languages; (ii) it should not be possible for the model to learn a correspondence between tags and characters in the output sequence; and (iii) the model cannot get information about tags in the low-resource language from the high-resource language's examples.

We additionally perform two last experiments:

Language-dependent letter embeddings with separation symbol (l-emb-sep). This is the same as l-emb, but we introduce a new separation symbol `SEP` between the tags and the characters, solving the problem that it is not clear where the tag ends and the word starts. We expect equal or better performance than for l-emb.

Language-dependent tag embedding with separation symbol (t-emb-sep). This is equivalent to t-emb, but we again insert a new separation symbol `SEP` between the tags and the input word's characters. We expect equal or better performance than for t-emb.

3.3 Intuition

In Table 3 we display an overview of which of the working mechanisms of cross-lingual transfer learning we expect to be effected by which changes to the high-resource training data. Depending on the relationship between the source and the target language, e.g., whether they use the same affixes to express the same morphosyntactic properties, we anticipate stronger or weaker effects. The regularization effect should not be influenced by our changes to the data.

3.4 Results and Analysis

For the low-resource training set of size 50, the models with the original setup and without transfer perform best and worst, respectively. However,

	50						200					
	ES (+0)	AR	FR	IT	CA	PT	ES (+0)	AR	FR	IT	CA	PT
original	**.0075(.00)**	.1496(.01)	.4277(.02)	**.5161(.01)**	**.6216(.02)**	**.4755(.01)**	**.5012(.03)**	.6596(.01)	.7080(.01)	.7713(.01)	.8142(.01)	.6885(.01)
l-ciph	-	.1209(.01)	.1837(.03)	.3207(.02)	.2937(.02)	.1005(.06)	-	.6626(.01)	.6491(.02)	.7032(.02)	.7151(.00)	.6155(.03)
t-ciph	-	.1208(.02)	.3491(.01)	.4823(.01)	.4963(.02)	.3623(.02)	-	.6405(.01)	.7058(.01)	.7768(.01)	.8040(.01)	.6317(.01)
l-emb	-	.1353(.06)	.2905(.01)	.2842(.09)	.4327(.03)	.2723(.06)	-	**.7109(.02)**	.7048(.01)	.7412(.01)	.7655(.02)	**.7323(.01)**
t-emb	-	.1363(.03)	.3941(.02)	.5012(.02)	.5610(.02)	.4300(.02)	-	.6464(.00)	.7364(.01)	.7760(.01)	.8142(.01)	.6690(.01)
l-emb-sep	-	.1312(.03)	.3240(.03)	.3554(.04)	.4282(.03)	.2883(.06)	-	.6464(.00)	.7180(.02)	.7522(.01)	.7757(.01)	.7250(.02)
t-emb-sep	-	**.1672(.02)**	**.4516(.01)**	.5138(.01)	.5944(.02)	.4608(.02)	-	.6668(.01)	**.7434(.00)**	**.7946(.01)**	**.8305(.01)**	.6824(.01)

Table 4: Results for all experiments and all high-resource source languages. ES denotes experiments without transfer. 50 and 200 are the numbers of low-resource training examples. All results are averaged over 5 training runs, standard deviation in parenthesis.

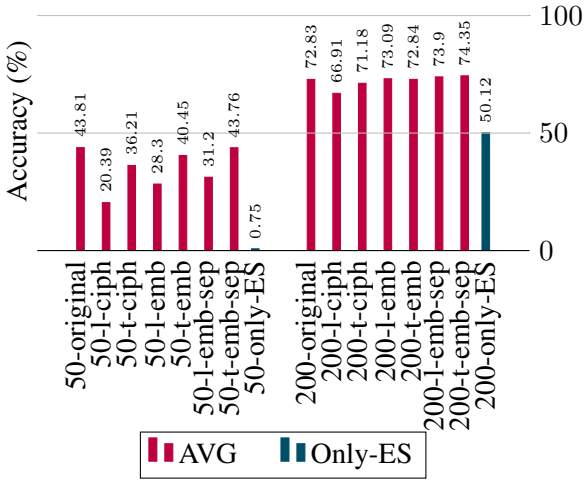

Figure 2: Results for all experiments, averaged over all languages. *only*-ES denotes a model trained exclusively on 50 or 200 Spanish examples.

for low-resource training size 200, t-emb-sep performs best in most case, and without transfer still performs worst. The order of the accuracies averaged over all languages can be seen in Figure 2: original > t-emb-sep > t-emb > t-ciph > l-emb-sep > l-emb > l-ciph for 50 and t-emb-sep > l-emb-sep > l-emb > t-emb > original > t-ciph > l-ciph for 200 low-resource examples. The detailed results of each language can be found in Table 4.

First, this shows clearly that the character embeddings are more important for the task than the tag embeddings. Second, l-emb (resp. t-emb) and l-ciph (resp. t-ciph) correspond to a setting with *no* additional information vs. a setting with *potentially wrong* information. Generally higher accuracies for separate embedding spaces indicate that the model can learn incorrect information via transfer. Thus, the choice of the source language seems to be very important. The differences in performance between original and l-emb represent the influence of shared vs. separate embedding spaces, i.e., vocabularies in the case of the letters. Sharing a vocabulary seems to influence the final accuracy a lot, and more positively for 50 low-resource examples. We can explain this with the model learning to copy – it has no intrinsic way of knowing which input character equals which output character in the vocabulary unless it has seen it at least once. However, for 200 Spanish examples, we can expect all characters to appear in the Spanish training data, such that the character language model and tag-output correspondence get more important. This explains the unexpected result that l-emb performs best for Arabic (200) and Portuguese (200): both source languages potentially confuse the language model; in Portuguese we contribute this to a big overlap of lemmata in the two languages with Portuguese often inflecting in a different way (Kann et al., 2017). Further, the differences in performance between original and t-emb show that the model indeed learns information from the tags, supposedly which output sequence is more likely to appear with which tag.

The l-emb-sep and t-emb-sep results show that a separation symbol clearly improves the model's performance.

4 Related Work

Transfer learning with encoder-decoder networks. Encoder-decoder RNNs were introduced by Cho et al. (2014) and Sutskever et al. (2014) and extended by an attention mechanism by Bahdanau et al. (2015). Lately, much work was done on *multi-task learning* and *transfer learning* with encoder-decoder RNNs. Luong et al. (2015) investigated multi-task setups for sequence-to-sequence learning, combining multiple encoders and decoders. In contrast, in our experiments, we use only one encoder and one decoder. There exists much work on multi-task learning with encoder-decoder RNNs for machine translation (Johnson et al., 2016; Dong et al., 2015; Firat et al., 2016;

Ha et al., 2016). Alonso and Plank (2016) explored multi-task learning empirically, analyzing *when* it improves performance. Here, we focus on *how* transfer via multi-task learning works.

Paradigm completion. SIGMORPHON hosted two shared tasks on paradigm completion (Cotterell et al., 2016, 2017), in order to encourage the development of systems for the task. One approach is to treat it as a string transduction problem by applying an alignment model with a semi-Markov model (Durrett and DeNero, 2013; Nicolai et al., 2015). Recently, neural sequence-to-sequence models are also widely used (Faruqui et al., 2016; Kann and Schütze, 2016; Aharoni and Goldberg, 2017; Zhou and Neubig, 2017). All the above mentioned work were designed for one single language.

5 Conclusion

We conducted a set of experiments to explore the mechanisms behind cross-lingual transfer learning for morphological reinflection. Our findings indicate that knowledge about a language's typical character sequences and outputs for certain morphological tags can be transferred. In particular, this means that the effect cannot be reduced to sole regularization.

Acknowledgments

We would like to thank Hinrich Schütze and the anonymous reviewers for their helpful comments.

References

Roee Aharoni and Yoav Goldberg. 2017. Sequence to sequence transduction with hard monotonic attention. In *ACL*.

Héctor Martínez Alonso and Barbara Plank. 2016. Multitask learning for semantic sequence prediction under varying data conditions. *arXiv preprint arXiv:1612.02251* .

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *SSST*.

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. The CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *CoNLL-SIGMORPHON*.

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

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *ACL*.

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-HLT*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073* .

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798* .

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Vigas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558* .

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. One-shot neural cross-lingual transfer for paradigm completion. In *ACL*.

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

M Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 16 edition.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114* .

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

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

Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR* abs/1212.5701.

Chunting Zhou and Graham Neubig. 2017. Multi-space variational encoder-decoders for semi-supervised labeled sequence transduction. In *ACL*.