Neural Morphological Tagging of Lemma Sequences for Machine Translation

Costanza Conforti cc918@cam.ac.uk
Matthias Huck mhuck@cis.lmu.de
Alexander Fraser fraser@cis.lmu.de
Center for Information and Language Processing, LMU Munich, Munich, Germany

Abstract

Translation to morphologically rich languages is a difficult task because of sparsity caused by morphological richness. In this work we perform a pilot study on predicting the morphologically rich POS tags of sequences of lemmas. Similar studies have been conducted in the context of phrase-based statistical machine translation. We implement a state-of-the-art tagger taking lemmas as input and show that we can successfully predict the morphologically rich POS tags, with accuracies of up to 91%.

1 Introduction

Modeling sequences of tokens in morphologically rich languages (MRLs) is a difficult task of great importance in many applications of NLP. For instance, translation from a morphologically poor language (such as English) to an MRL (such as German or Czech) is known to be difficult. An effective approach for modeling MRLs is to break the sequence into a factorized representation, such as lemmas paired with their morphologically rich POS representations (e.g., for a German noun, the rich representation includes the noun POS tag, and the three grammatical features gender, number, and case).

In this paper, we assume that we have a good system for generating lemmas and study whether we can automatically recover the morphologically rich POS representation. This is more difficult than morphologically rich POS tagging, which takes a sequence of surface forms and recovers the most likely morphologically rich POS representation, because lemma input is underspecified. This task was previously studied by Minkov et al. (2007). We differ in two ways: (1.) we implement a state-of-the-art neural tagger, rather than a Maximum Entropy Markov model, and (2.) we predict rich morphological POS, rather than surface forms.

Studying the prediction of morphologically rich POS given lemmas is an interesting problem in its own right. It has implications for NLP applications involving the generation of MRL sentences including machine translation. A concrete application is to apply it in an end-to-end MT system. Similar morphological prediction systems have been applied by Toutanova et al. (2008), Bojar and Kos (2010) and Fraser et al. (2012) in phrase-based SMT. A pipeline of such a system is depicted in Figure 1.

Given the promising results in this initial study, we plan to combine our tagger with a standard neural machine translation model, resulting in a multi-task system which produces pairs of lemmas and morphologically rich POS tags. An important benefit of such a system over previous approaches which produce such pairs directly using a standard NMT model (e.g., Tamchyna et al. (2017)) is that we will be able to train it in a multi-task fashion, where some

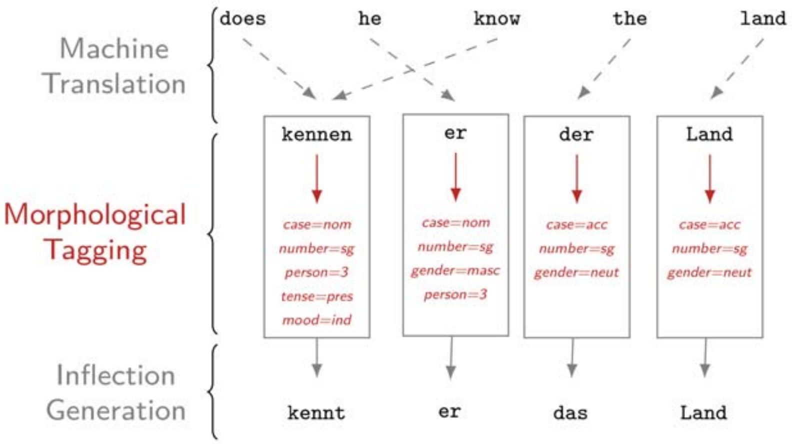

Figure 1: Example of a machine translation system using our morphology tagger (in red) as an intermediate step in the translation process into a target MRL.

training examples contain source language text (from parallel data), while others do not (from monolingual data).

In this paper, we present our language-independent neural tagging system implemented and trained for this task. German, an MRL belonging to the Indo-European family, has been chosen for a case study. This choice is motivated by the fact that, within MRLs, German has been widely studied in recent years, and many resources are publicly available.

2 Motivation

2.1 Rich Morphology in Machine Translation: Still Challenging

MRLs are difficult to deal with in natural language processing applications. Naive technological approaches without any proper analysis and modeling of morphological phenomena tend to result in underperforming systems for MRLs. Computational morphology research is therefore a long-standing subdiscipline of NLP, with an impact on almost any use case that involves an MRL. Information retrieval systems have traditionally benefited from stemmers or lemmatizers, which reduce inflected surface forms to a word stem or a canonical form. In MT, translating *from* source-side English *into* an MRL is notoriously more difficult than the other way around (Bojar et al., 2017). The English→MRL direction requires the system to decide amongst many possible inflected forms on the output side. The source-side lexical counterpart is morphologically underspecified, which complicates both statistical modeling and search. Data-driven approaches over MRLs furthermore suffer eminently from data sparsity issues under medium- to low-resource conditions. Many inflected forms are observed rarely.

Source-side MRL. Rich morphology on the source side can to some extent be tackled via preprocessing. Syntactic and morphological analyzers can be employed, based on which a source sentence representation can be constructed which is more appropriate as the input to a translation system (Popović and Ney, 2004; Popović et al., 2005; Goldwater and McClosky, 2005). E.g., certain morphological features of the source words may be dismissed beforehand. Non-reversible modifications are uncritical on the source side and can potentially alleviate the modeling problem and counteract data sparsity. Arabic is a prominent example of a language that typically undergoes heavy morphosyntactic analysis and preprocessing on the source side (Lee, 2004; Habash and Sadat, 2006; Hasan et al., 2011).

Target-side MRL. The more challenging question of how to tackle rich morphology on the target side has undergone quite some research. Factored phrase-based models (Koehn and Hoang, 2007) can be used to produce inflected output via a separate decoding path and a generation step (Bojar, 2007). The blow-up of the search space can make such models impractical. Backoff techniques (Koehn and Haddow, 2012) or flat factored models with supplementary features over lemmas and linguistic annotation (Stymne et al., 2008; Huck and Birch, 2015) are more tractable alternatives.

Phrase-based translation models, accompanied with n-gram language models, have a relatively limited local view. Some morphological phenomena go beyond local context and require agreement across long distances. In syntax-based systems with a chart-based decoding procedure, engineering adequate agreement constraints is more viable (Williams and Koehn, 2011, 2014). Pursuing a different idea, Avramidis and Koehn (2008) and Daiber and Sima'an (2015) have attempted to annotate input sentences beforehand with morphological features that are exhibited by the target-side MRL, thus taking the burden of the inflection selection away from the phrase-based decoder. Chahuneau et al. (2013) and Huck et al. (2017c), on the other hand, have specifically looked into how to produce unseen morphological variants without resorting to a factored generation step. To that end, they augment their phrase tables with synthetic entries.

Other researchers have proposed two-step approaches to MT into MRLs, where the output of the first step (the actual translation) lacks certain morphological features of the MRL, which in turn have to be predicted by a separate module in a second step in order to end up with inflected target language sentences (Toutanova et al., 2008; Fraser et al., 2012). Slightly less supervision might be required by another technique for tackling rich morphology on the target side: word segmentation, and subsequent modeling on a subword level. Inflected target words in the training data can e.g. be segmented into stems and morphological affixes (Fishel and Kirik, 2010; Clifton and Sarkar, 2011; Passban et al., 2017). The segmentation of output hypotheses of the MT system needs to be reverted in postprocessing.

In modern neural machine translation engines, word segmentation by means of a Byte Pair Encoding (BPE) style algorithm is a common trick to shrink the vocabulary size (Sennrich et al., 2016). Recent research has shown that NMT of MRLs benefits from word segmentation techniques that are linguistically more informed than plain BPE (Ataman et al., 2017; Pinnis et al., 2017; Huck et al., 2017b,a). Not all prior research on MRLs in traditional phrase-based MT can be readily transferred to NMT. One of the most promising attemps to date is following the theme of two-step MT. A second-step module generates inflections from lemmas and morphological tags. The first-step NMT module outputs interleaved sequences of such lemmas and their respective tags (Burlot et al., 2016, 2017; Tamchyna et al., 2017). Research on how to best model morphology with neural networks is ongoing, in MT and in other areas of NLP (Botha and Blunsom, 2014; Ebert et al., 2016; Vania and Lopez, 2017; Belinkov et al., 2017; García-Martínez et al., 2016; García-Martínez et al., 2017; Burlot and Yvon, 2017).

2.2 Morphological Tagging of Lemmas: Utility and Limitations

Morphological tagging is the task of marking up each token in an input sequence with the corresponding morphological features which describe its inflectional properties. In the case of morphologically poor languages, a word can usually be described sufficiently using information about its POS tag (Mueller et al., 2013). MRLs require a more detailed analysis. The term MRL refers to a language where word shapes encode a consistent number of syntactic and semantic features. This behavior is particularly productive in fusional and agglutinating languages, where it can involve both verbal conjugation and nominal declination. In these situations, a more accurate morphological label is usually attached to the POS tag. For this reason, morphological tagging has been also defined as fine-grained POS tagging (Labeau et al., 2015).

A divide-and-conquer approach is often adopted to deal with data sparsity in MRLs. First, the MRL inflectional structure is simplified, thus reducing the number of word types and consequently data sparsity. Then, the NLP task is carried out over the simplified MRL data. Finally, the complete MRL's inflection is reconstructed as a separate post-processing step. In this way, a complex problem is decoupled into approachable subtasks. Previous authors have pursued this approach in MT, e.g. from English or Chinese into simplified representations of MRLs like Czech (Bojar, 2007), German (Fraser et al., 2012), Spanish (Costa-Jussà and Escolano, 2016). A disadvantage of these previous attempts is that they require a careful, linguistically-motivated analysis in order to optimize the choice of morphological features which can be removed from the MRL, thus decreasing data sparsity, without increasing the difficulty of the final morphology generation step too much (as in Costa-Jussà and Escolano (2016)). In this work, by contrast, we investigate the possibility of generating morphological annotations from *completely underspecified* input sequences. We implement and train a neural morphological tagger which deals with lemma sequences. The lemmas are uninflected canonical base forms with no morphological features at all. Given a morphologically fully underspecified lemma sequence, the task considered in this work consists of annotating each input symbol with the complete set of morphological features needed to generate the inflected word forms.

However, in our approach, not all morphological features can in each and every case accurately be recovered from target-side lemma sequences only, without ever taking the source sentence into account. Consider the English sentence *"the flowers are blossoming"* and its German translation *"die Blumen blühen"*. The German lemma sequence is *"der Blume blühen"*. Rather than annotating the article lemma *"der"*, the noun lemma *"Blume"*, and the verb lemma *"blühen"* as plural each, a morphological tagger could just as well predict singular here, resulting in the grammatically correct sentence *"die Blume blüht"* (English: "*the flower is blossoming"*). Under many circumstances, however, our morphological tagger is able to correctly disambiguate most features. E.g., for the English sentences (1.) *"a flower is blossoming"* and (2.) *"many flowers are blossoming"* with their respective German lemma correspondences (1.) *"ein Blume blühen"* and (2.) *"viel Blume blühen"*, the number feature can be unambiguously established from the first lemma in each of the two German lemma sequences, the indefinite article *"ein"* and the adverb *"viel"*. Given their context, noun and verb should also be annotated correctly as singular in the first example and plural in the second. In ambiguous cases, the morphological tagger will benefit from such contextual clues, as well as from the semantics of the lemmas, their syntactic order, and co-occurrence frequencies in the training data. Our neural morphological tagger performs very well on the very difficult task of predicting rich POS from this very underspecified representation.

3 A Neural Architecture for Morphological Tagging of Lemmas

We consider a tagging task in which, given a lemmatized input sequence $\mathbf{x} = \{x_1, x_2, ..., x_N\}$, each input symbol is assigned one out of a set $\mathcal{T}$ of predefined fine-grained tags, resulting in the output sequence $\mathbf{y} = \{y_1, y_2, ..., y_N\}$, where $\mathbf{x}$ and $\mathbf{y}$ have the same length N. Our architecture leverages information coming from multiple input channels. For each token, three features are considered, none of which requires any language-specific tool to be extracted:

- **Lemma.** The input token itself, which is a canonical word form.
- **Capitalization.** Following Collobert et al. (2011) and Santos and Zadrozny (2014), a capitalization feature has been implemented, which indicates whether a given lemma is completely uppercased, completely lowercased, capitalized, contains at least one uppercase character in a position other than the first, or none of these cases. Encoding information about the capitalization of a token can be useful, in particular for unknown input symbols.

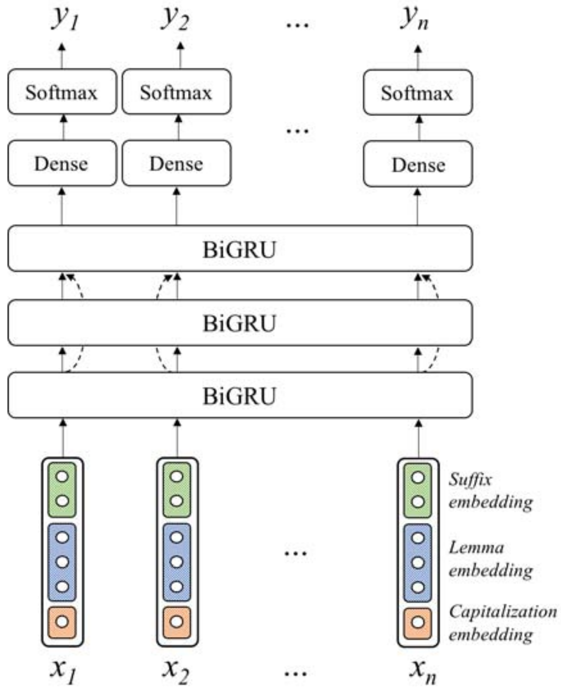

Figure 2: The architecture of our neural network model for morphological tagging.

Tagset	Tagset size	Avg. # labels/stem	Max. # labels/stem
morph	255	3.96	96
POS+morph	678	4.00	128

Table 1: Tagset statistics.

Feature	Values
Gender	`Masc, Fem, Neut, *`
Case	`Nom, Gen, Dat, Acc, *`
Number	`Sg, Pl, *`
Person	`1, 2, 3`
Degree	`Pos, Comp, Sup`
Tense	`Pres, Past`
Mood	`Ind, Subj, Imp`

Table 2: Morphological features.

Parameter	Value
batch size	32
optimizer	adadelta
dropout	0.3
lemma embedding size	128
suffix embedding size	40
capitalization embedding size	10
recurrent layer size	256

Table 3: Hyperparameters of the network.

- **Suffix.** For our purposes, a suffix is defined as the last n characters of a token. Suffixes can be indicative of a lemma's syntactic category, or POS. As the inflection pattern largely depends on the POS, suffixes can also help predict the morphological features of lemmas.

3.1 Neural Network Components

The proposed architecture is composed of a series of modules, or layers (Figure 2). Given an input sequence of length N, in the first layer each input symbol is mapped to three one-hot vector representations, corresponding to the three features described above. The one-hot vector representations are then projected into real-valued dense vector representations (embeddings). The lemma, the capitalization, and the suffix embeddings are then concatenated in order to obtain a single vector representation for each input symbol.

The central body of the architecture are three stacked bidirectional recurrent layers, using GRUs as recurrent cells. The choice of bidirectional recurrent layers over traditional feedforward layer was motivated by the promising results obtained in other labeling tasks, such as named entity recognition (Lample et al., 2016). The use of GRUs was preferred over LSTMs because their simpler internal structure allows for faster training, showing comparable results in preliminary experiments. Inspired by Heigold et al. (2016), we add *skip connections* between the first and the third bidirectional recurrent layer to allow for direct propagation of information between layers at different levels of depth. At the top of the architecture, a time-distributed densely-connected layer produces one $|\mathcal{T}|$-dimensional vector per time step, where $|\mathcal{T}|$ is the tagset size. Finally, the output label at each time step is given by a softmax operation over the tagset. The weights of the network (θ) are jointly estimated using the conditional log-likelihood $F(\theta) = -\sum_{n=1}^{N} \log p_\theta(y_n|x_1, ..., x_N)$.

3.2 Training

In order to train our neural model, a high amount of lemmatized data, tagged with morphological annotations, is required. In a multi-component end-to-end NLP system that involves our tagger, one would typically strive for a good match between the training data of the tagger and the training data of the other components, so as to achieve ideal interaction between them. But corpora that are manually annotated with lemma and fine-grained POS will rarely ever be at hand for most tasks. Common practice in most practical scenarios would be to synthetically annotate the task-specific training corpus. We follow this real-world rationale and work with synthetic annotation in our study.

Data and preprocessing. We train on the Europarl v7 corpus (Koehn, 2005). The conventional Europarl test sets (*test2006*, *test2007*, *test2008*) that had been released for the WMT shared task are used for development and testing.[1] Our main tagging evaluation results will be reported on *test2007*, which we abbreviate as *test* in most tables, while *test2006* serves as our *dev* set. The corpora are tokenized and frequent-cased using scripts from the Moses toolkit.[2] They are then annotated with lemmas, POS tags, and morphological tags with the pretrained tagging model for German provided by the MARMOT toolkit.[3] MARMOT is a CRF-based tagger with a reported accuracy of 97.94 for POS tagging and of 91.65 for morphological annotations on the TIGER test set (Mueller et al., 2013). The toolkit produces lemmas using LEMMING (Müller et al., 2015), a language-independent token-based lemmatizer which is reaching state-of-the-art accuracy of 98.10 for the German language. Our training corpus contains 1.9M sentences, the dev and test set 2,000 sentences each.

Tagsets. As in (Mueller et al., 2013), two tagsets are considered in our work. The first tagset (morph) is composed of morphological annotations, while the second (POS+morph) is obtained by concatenating the POS tag and the morphological annotation of each input lemma. For example, the lemma "Parlament" in the context "im Parlament" (in the parliament) would receive the POS+morph label `NN+case=dat|number=sg|gender=neut`, where `NN` is the POS tag and the segment `case=dat|number=sg|gender=neut` corresponds to the morph label, specifying the values taken by the morphological features case, number, and gender. As shown in Table 1, considering POS+morph labels increases both tagset size and classification ambiguity. Morphological labels show a compositional property (Cotterell and Schütze, 2015). In fact, each label is represented by a concatenation operation over a set of `feature : value` pairs. Table 2 reports the morphological features used for annotation. Some features are specific to certain word classes, such as mood for finite verbs. Other features can occur in more contexts, such as gender and case (articles, pronouns, adjectives, and nouns).

Vocabularies. In order to increase training speed, we reduce the lemma vocabulary to the 40K most frequent entries. This allows for an OOV rate of 0.016 over the training and of 0.019 on the test sets. After some preliminary experiments, we commited to suffixes of size $n = 4$. No vocabulary reduction is performed on the suffix vocabulary.

Model setup. Neural models are trained to predict labels from the two tagsets, morph and POS+morph, respectively. The same hyperparameters of the network architecture are configured for both models (Table 3). During training, the input sequences are padded or cut up to the length of 70 tokens for the morph tagset and of 60 for the POS+morph tagset. At test time, the sentences are padded up to the length of the longest sequence in the dataset. Our implementation takes around one week to train 15 epochs on a Nvidia GeForce GTX 750 GPU.

[1] `http://www.matrix.statmt.org/test_sets/list/`

[2] `https://github.com/moses-smt/mosesdecoder/`

[3] `https://github.com/muelletm/cistern/tree/master/marmot`

Tagset	Max. Freq.		Freq. Lookup		CRF		Our Neural Tagger	
	dev	**test**	**dev**	**test**	**dev**	**test**	**dev**	**test**
morph	39.15	38.92	59.97	60.32	87.77	87.91	**91.22**	**91.34**
POS+morph	7.65	7.47	56.52	56.78	86.96	86.78	**90.61**	**90.92**

Table 4: Accuracy obtained on morphological tagging of lemmas considering morph and POS+morph tagsets.

4 Empirical Evaluation

Three non-neural baselines have been built to provide lower bounds, two dummy classifiers and a CRF-based model:

- **Maximum frequency.** The first baseline (Max. Freq.) always predicts the most frequent label in the tagset.
- **Frequency lookup.** The second baseline (Freq. Lookup) uses a lookup table to return, for each lemma, the label it is most frequently annotated with in the training corpus.
- **CRF.** A CRF model was trained on the lemmatized Europarl corpus, using the MARMOT toolkit with its default parameters.

4.1 Intrinsic Evaluation: Tagging

Table 4 reports on the results of tagging on the development and test set. Our neural tagger clearly beats all the baselines taken as lower bound, considering both tagsets.

Quantitative analysis. In order to understand the performance of our models at predicting each single feature, the morphological labels were split into their components and performance is measured according to the following metrics:

- **F1-score A.** Performance is measured only across word classes which present the given feature in the gold. For example, degree is measured only in adjectives.
- **F1-score B.** Performance is measured across all word classes. If a given feature is not predicted for a label, or it is not present in the gold annotation, its value is set to an artificial NNN class. In this way, features which are correctly *not* predicted by the system, such as gender for verbs, count as true positives.

Results of this evaluation are reported in Table 5. The overall feature performance is satisfactory in line with both evaluation criteria. The performance scores for all features are slightly improved using the model predicting POS+morph labels. In fact, as POS tags are indicators of word classes, jointly predicting them with the morphological labels could help the system learn which features should be predicted and which should not be produced.

Considering evaluation of type A, the best results are obtained for the gender feature. Contrary to what happens for the other features, gender constitutes a lexical attribute of nouns, and an inflectional feature for other nominal constituents. This could have had the effect of simplifying the classification problem for nouns, thus also strengthening performance on dependent tokens, such as determiners and adjectives. Moving to evaluation of type B, an overall enhancement in performance can be observed for all features, suggesting that the systems are successfully able to learn when a certain morphological feature should be predicted or not.

In general, the performance of tagging with respect to single morphological features seems to highly depend on the distributional characteristics of the corpus, as well as on the relative balance within a single feature's values. Highest performance is obtained on the morphological features which present the highest support in the training set. Figure 3 reports the confusion

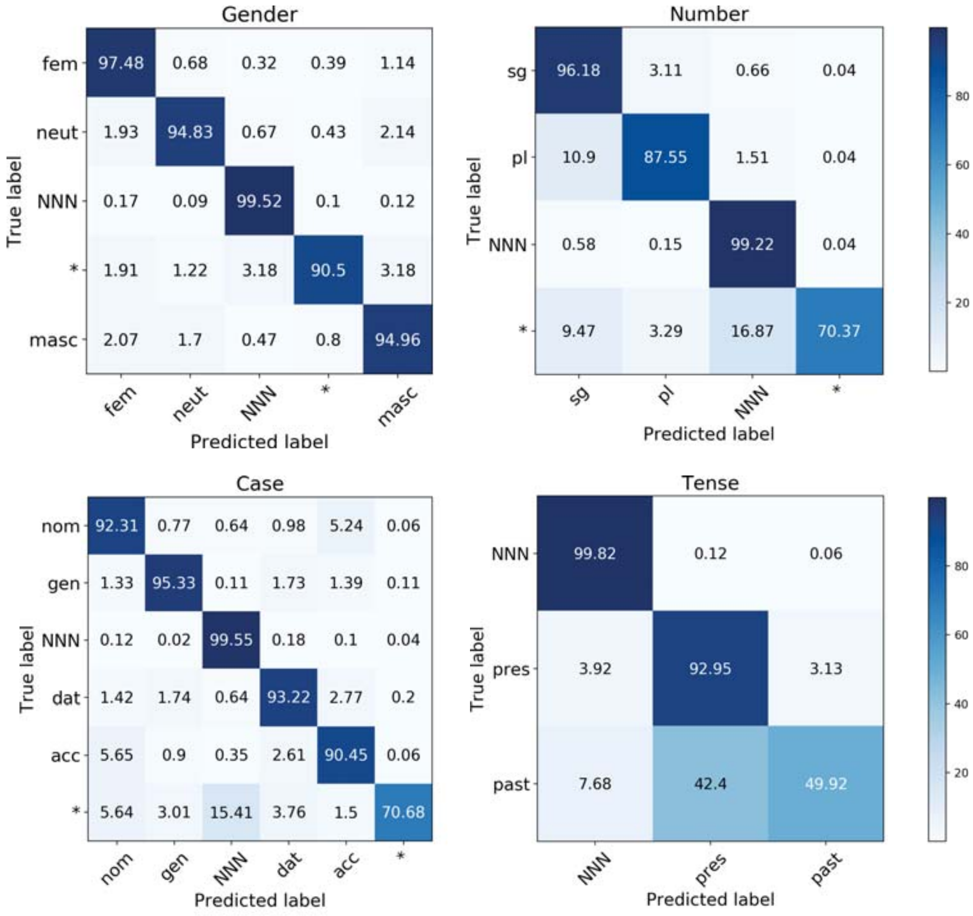

Figure 3: Normalized confusion matrices considering the morphological features gender, number, case, and tense. The value NNN refers to morphological labels where the given feature is *not* present in gold or predicted by the system, while $\star$ indicates that the given feature is present or predicted, but undefined.

Morph	**Support**	**Number**	**morph labels**		**POS+morph labels**	
Feature	**(% in train set)**	**of values**	**F1-score A**	**F1-score B**	**F1-score A**	**F1-score B**
POS	100.00	48	-	-	98.45	98.45
Case	50.41	5	93.69	96.41	93.87	96.52
Gender	49.86	4	96.57	97.93	96.61	97.94
Number	58.05	3	95.12	96.63	94.99	96.57
Person	12.00	3	94.69	99.52	94.70	99.56
Degree	8.42	3	82.36	99.63	82.21	99.63
Mood	7.65	2	90.14	99.27	90.01	99.29
Tense	7.64	2	85.06	98.86	85.44	98.90

Table 5: Performance of tagging considering single morphological features. Support is defined as the number of labels where a given feature is defined, divided by the total number of labels in the training set.

matrices of case and gender, two highly frequent morphological features which distinguish between the highest number of possible values. Considering case, the tagger was able to learn to discriminate relatively well between the four cases, due to their distributionally different characteristics. The highest number of misclassifications occurs between `Nom` and `Acc`, which present some similar distributional patterns in the German language. Excluding the `*` value, which occurs with a support lower than 1%, the lowest performance is obtained on `Gen`, which is also the less frequent value for this morphological feature. Moving to gender, the best performance is achieved with `Fem`. This is not only the most frequent value, but also the gender which contains most of the substantives obtained through derivational morphology (Matthews, 1991), thus presenting a pattern which can be easily spotted by the suffix feature. In contrast, misclassifications are more common in case of morphological features which present a high class imbalance, especially when the classes tend to appear in similar context. This is exemplified by the features number and tense, whose confusion matrices are reported in Figure 3. In the case of number, the morphological feature with the highest support, our tagger tends to misclassify `Plur` occurrences in favour of the most frequent value `Sing`, which occurs roughly twice as often. The same pattern can be observed also in the confusion matrix of tense, a feature with a considerably lower support in the training corpus (as it occurs only for verbs). Here, the extreme feature imbalance induces the system to wrongly label almost half of the `Past` occurrences as `Pres`, which accounts alone for almost 85% of the training samples.

Qualitative analysis. These observations are supported by a qualitative analysis of the systems' output on the test set[4]. In particular, our tagger is able to learn to produce verbs in the correct tense when an explicit temporal mention is present in the sentence, as in the example below, which contains both a past and a present adverbs (in italics):

> Wie ich bereits *gestern Abend* <u>sagte</u>, und ich <u>tue</u> dies *heute* erneut, [...]
> *As I said last night - I will say it again today, [...]*

Number can be disambiguated when a clue such as *viel* (En.: *many*) is found:

> [...], enthält der diesjährige Haushaltsentwurf <u>viele politische Botschaften</u> [...]
> *[...], there are many political announcements coming out of this year's Budget [...]*

Moreover, our system can also learn complex distributional patterns, being also able to cope with long-distance dependencies. In the following example, the combination of the two modal verbs *hätten ... müssen* (separated by 16 tokens) can only accept the `Subj` mood. The tagger correctly predicts:

> Dann *<u>hätten</u>* sich die französischen Behörden [...] an uns *wenden müssen*.
> *The French authorities ought to have consulted us [...]*

However, when no explicit clue is present in the sentence, and the distributional characteristics of a morphological feature's values are similar, the tagger chooses the value which was most frequently associated with the given lemma in the training corpus, such as the `Plur` Number considering the substantive in the Nominative case *Nachbarland* (English: *neighbouring country*):

> <u>Unsere</u> <u>Nachbarländer</u> <u>haben</u> (*correct*: <u>Unser</u> <u>Nachbarland</u> <u>hat</u>) sich [...] während der letzten zehn Jahre [...] bemüht.
> *<u>Our</u> <u>neighbouring</u> <u>countries</u> <u>have</u> (correct: <u>Our</u> <u>neighbouring</u> <u>country</u> <u>has</u>) struggled [...] over the last decade [...]*

[4] For the sake of clarity, in this section we report the re-inflected samples obtained using the predicted morphological labels, instead of showing the labels themselves (which could be difficult to interpret).

Our Tagger					**Lamb**	
Suffix Embeddings			**Lemma Embeddings**			
-ISCH	-RUNG	-EREN	SOLIDARISCH	KOOPERATION	SOLIDARISCH	KOOPERATION
-lich	-lung	-rken	generell	Kommunikation	Solidarität	Zusammenarbeit
-ativ	-hung	-ösen	uneingeschränkt	Kohäsion	Elitenförderung	Kooperationsprojekt
-rell	-tung	-üben	systematisch	Steuerung	Selbstverantwortung	Gemeinschaftsprojekt
-iell	-gung	-üfen	schrittweise	Produktion	Lebensrecht	Kooperationsvertrag

Table 6: 4-nearest neighbors of the embeddings encoding of three suffixes (columns 1-3) and two lemmas (columns 4-5) from the training set, computed using cosine similarity on the vectors jointly learned by our POS+morph neural tagger after 15 epochs. The 4-nearest neighbors of the corresponding purely semantic Lamb lemma embeddings (Ebert et al., 2016) are reported for comparison purposes (columns 6-7).

In fact, where the sentence offers no clue to disambiguate a particular morphological feature, it is in principle impossible to recover the correct feature's value from the lemmatized sequence. The system can rely only on the statistical characteristics of the corpus to infer it. It should be observed, however, that even when the system predicts a wrong morphological feature, the complete sequence of labels is nevertheless usually grammatical and coherent, as shown in the example above (where the singular subject *unser Nachbarland* agrees in number with the principal verb *hat*).

An analysis of the embedding matrices jointly learned during training shows that our model is able to learn complex relations of *morphological similarity*, leveraging information coming from both lemmas and suffixes. As reported in Table 6, our learned suffix embeddings tend to cluster together with suffixes denoting the same word class (adjectives for *-isch*, feminine nouns for *-rung*, and verbs for *-eren*). This holds true also for lemma embeddings, which cluster with input symbols belonging to the same morphological class, and with which they share almost no semantic content. This is particularly evident when comparing our learned lemma embeddings with the purely semantic Lamb embeddings (Ebert et al., 2016). The nearest neighbor of the lemmatized adjective *solidarisch* (En.: *solidary*) in our model is the adjective *generell* (En.: *general*), while the corresponding Lamb nearest neighbor is the noun *Solidarität* (En.: *solidarity*), as reported in Table 6. For *Kooperation* (feminine noun, En.: *cooperation*), the nearest neighbors in our space are all feminine nouns, while the nearest Lamb vectors are semantically related nouns with different genders.

4.2 Extrinsic Evaluation: Inflection Generation

Inflection generation, also called *morphology generation*, is the NLP task of generating an inflected word from its lemma paired with its morphological tag. This task offers a nice opportunity for an extrinsic evaluation of our tagger's predictions. We implemented an inexpensive lookup-based inflection generation system. At each position i, the inflected word w_i corresponding to the lemma l_i is produced according to the following chain of backoff operations:

1. POS+morph bigram$_{+1}$: $\max_{w_i}(count(\{(l_i, t_i, t_{i+1}) \rightarrow w_i\}))$
2. POS+morph unigram: $\max_{w_i}(count(\{(l_i, t_i) \rightarrow w_i\}))$
3. Lemma bigram$_{+1}$: $\max_{w_i}(count(\{(l_i, l_{i+1}) \rightarrow w_i\}))$
4. Lemma unigram: $\max_{w_i}(count(\{(l_i) \rightarrow w_i\}))$
5. Unseen lemma: $l_i \rightarrow l_i$

where t_i corresponds to the POS+morph label predicted by our tagger. Lookup tables are calculated over the training corpus.

MT System	**Lemma-BLEU**	
	test2007	**test2008**
Baseline	28.9	28.7
Lemma NMT	29.3	29.2

Table 7: Quality of lemma translation. Baseline hypothesis translations and references have been lemmatized with LEMMING.

MT System	**BLEU**	
	test2007	**test2008**
Baseline	25.8	25.7
Pipelined	24.8	24.5

Table 8: Machine translation quality. We report case-sensitive BLEU of fully inflected, postprocessed translations.

Reinflection accuracy over the test set reaches **99.22** using gold labels, and **95.54** using our predicted labels. We believe that adopting state-of-the-art neural inflection generation systems such as the one by Kann and Schütze (2016), or using language-specific tools, as done in previous works on German (Fraser et al., 2012) and Spanish (Costa-Jussà and Escolano, 2016), could further enhance this performance. The purpose of this extrinsic evaluation, however, was not to propose a new competitive Inflection Generation system, but rather to prove that our labels constitute a good input to such a system and that, even with the limitations discussed in the previous section, it is possible to obtain satisfactory results in reconstructing the inflection of lemmatized input.

4.3 Machine Translation Evaluation

We build neural machine translation engines to evaluate the pipelined MT approach as illustrated in Figure 1. For comparison, a baseline NMT system translates directly from English to fully inflected German word surface forms. The pipelined architecture from Figure 1 is evaluated against this baseline. For the pipelined architecture, we train an NMT engine on a parallel corpus with lemmatized German target side. At test time, the latter engine performs the first step (MT from English words to German lemmas) in the pipeline. The second step is conducted by our tagger, which annotates the lemma hypothesis translation with morphological tags. Finally, the lookup-based inflection generator from Section 4.2 is employed to map the paired lemmas and predicted morphological tags to inflected German words.

We use the Nematus toolkit's implementation of encoder-decoder NMT with attention and GRUs (Sennrich et al., 2017). We train and test on the English–German Europarl data. In the NMT systems' training corpus, words are tokenized and frequent-cased, then segmented via byte-pair-encoding (BPE) (Sennrich et al., 2016) with 50K merge operations; likewise for lemmas, but with BPE operations extracted from the lemmatized data. We configure dimensions of 500 for the embeddings and 1024 for the hidden layer. We train with the Adam optimizer, a learning rate of 0.0001, batch size of 50, and dropout with probability 0.2 applied to the hidden layer. Translation quality is measured case-sensitive with BLEU (Papineni et al., 2002).

In Table 7, we use BLEU computed on lemmas (Lemma-BLEU), to show that we get a small gain in lexical choice (of the lemma) in the pipelined approach, where the NMT engine is trained to produce lemmas. However, the BLEU scores over fully inflected words in Table 8 suggest that a simple pipelined approach is not sufficient for end-to-end MT. We looked at the MT output and saw that it was mostly coherent, but there was confusion on features like number, tense, and mood. The slightly improved lexical choice of the lemma does not compensate for the loss that derives from the inherent limitations of completely decoupling lemma prediction and morphology prediction, as was discussed intuitively in Section 2.2 and later highlighted in detail empirically (Section 4.1, Figure 3, Table 5). The neural architecture yields surprisingly strong accuracy at morphological tagging of lemma sequences, but the pipelined approach from Figure 1 with completely underspecified lemma sequences and strict decoupling of the different components is too limiting for MT.

5 Relation to Previous Work on Morphological Tagging

Morphological tagging of inflected sequences. Fine-grained tagging of completely underspecified lemma sequences has not been studied much before, possibly because earlier non-neural models were deemed not powerful enough to tackle the task. Morphological tagging of inflected words, however, has been intensely investigated. The best performance for German (92% accuracy on TIGER) was achieved by Heigold et al. (2016) with a neural system which combines word and character embeddings. As observed by Santos and Zadrozny (2014) for POS tagging, character embeddings are particularly useful for processing MRLs, since they can help spot inflectional regularities. However, when dealing with completely uninflected input, this property is less useful. It is true that lemmas sometimes retain some kind of morphological information, like in derivation; however, this can be easily captured with suffix embeddings. In fact, a character-based model did not outperform our architecture in preliminary experiments.

Morphological tagging of (partially) underspecified sequences. Recently, Costa-Jussà and Escolano (2016) proposed a three-staged approach to Chinese–Spanish MT. First, a statistical MT system translates from Chinese into a morphologically impoverished version of Spanish which does not present two features: number and gender. Then, two neural classifiers separately annotate the simplified Spanish tokens with the missing features. As a last step, full forms are generated.In this way, the authors claim to beat the previous state-of-the-art performance for this specific language pair, reaching a classification accuracy higher than 90% on both features. No direct comparison can be drawn with our system, since their separate neural classifiers are trained on *partially uninflected* input (all other morphological features are still present). In our work, on the contrary, *all* features are considered. In particular, instead of *separately* training a different network for each feature, our single architecture makes one *joint* decision at each time step. In this way, our system, which is trained for a more complex task, can reach good results on all features. Furthermore, our choice of a joint prediction strategy allows for a *completely language-independent* approach. The carefully selected morphological simplification proposed by Costa-Jussà and Escolano (2016) would not generalize to other language pairs.

6 Conclusion

This work introduces a system for morphological tagging over lemmatized (that is, completely unspecified) input sequences. A detailed intrinsic and extrinsic evaluation showed that our language independent tagger reaches a very high performance by jointly predicting up to 8 morphological features, leading up to 678 possible combinations (considering POS+morph labels).

As a next step we will explore the implementation of a a multi-task system which produces pairs of lemmas and morphological labels. Niehues and Cho (2017) explored a multi-task NMT system producing coarse POS labels for the source language as well as words in the target language. We will instead produce lemmas in the target language, and at the same time use our tagger component to produce rich target POS. By giving our tagger access to the source sentences, we will overcome the limitations in our currently semantically underspecified representation, where, e.g., plural is not marked. Importantly, we will be able to train this system in a multi-task fashion, where some training examples contain source language text (from parallel data), while others do not. These examples will be taken from target monolingual data, allowing us to learn from large monolingual corpora how to inflect lemmas, as we did in this paper.

Acknowledgments

This work has received funding from the European Research Council (ERC) under grant agreement № 640550. We thank the anonymous reviewers of this paper for their efforts and for the constructive comments and suggestions.

References

Ataman, D., Negri, M., Turchi, M., and Federico, M. (2017). Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *PBML*, 108:331–342.

Avramidis, E. and Koehn, P. (2008). Enriching Morphologically Poor Languages for Statistical Machine Translation. In *Proc. of ACL*, pages 763–770, Columbus, OH, USA.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017). What do Neural Machine Translation Models Learn about Morphology? In *Proc. of ACL*, pages 861–872, Vancouver, Canada.

Bojar, O. (2007). English-to-Czech Factored Machine Translation. In *Proc. of WMT*, pages 232–239, Prague, Czech Republic.

Bojar, O. et al. (2017). Findings of the 2017 Conference on Machine Translation (WMT17). In *Proc. of WMT*, pages 169–214, Copenhagen, Denmark.

Bojar, O. and Kos, K. (2010). 2010 Failures in English-Czech Phrase-Based MT. In *Proc. of WMT*, pages 60–66, Uppsala, Sweden.

Botha, J. A. and Blunsom, P. (2014). Compositional Morphology for Word Representations and Language Modelling. In *Proc. of ICML*, Beijing, China.

Burlot, F., García-Martínez, M., Barrault, L., Bougares, F., and Yvon, F. (2017). Word Representations in Factored Neural Machine Translation. In *Proc. of WMT*, pages 20–31, Copenhagen, Denmark.

Burlot, F., Knyazeva, E., Lavergne, T., and Yvon, F. (2016). Two-Step MT: Predicting Target Morphology. In *Proc. of IWSLT*, Seattle, WA, USA.

Burlot, F. and Yvon, F. (2017). Evaluating the morphological competence of Machine Translation Systems. In *Proc. of WMT*, pages 43–55, Copenhagen, Denmark.

Chahuneau, V., Schlinger, E., Smith, N. A., and Dyer, C. (2013). Translating into Morphologically Rich Languages with Synthetic Phrases. In *Proc. of EMNLP*, pages 1677–1687, Seattle, WA, USA.

Clifton, A. and Sarkar, A. (2011). Combining Morpheme-based Machine Translation with Post-processing Morpheme Prediction. In *Proc. of ACL*, pages 32–42, Portland, OR, USA.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Costa-Jussà, M. R. and Escolano, C. (2016). Morphology generation for statistical machine translation using deep learning techniques. *arXiv preprint arXiv:1610.02209*.

Cotterell, R. and Schütze, H. (2015). Morphological Word-Embeddings. In *Proc. of NAACL*, pages 1287–1292, Denver, CO, USA.

Daiber, J. and Sima'an, K. (2015). Machine Translation with Source-Predicted Target Morphology. In *Proc. of MT Summit*, pages 283–296, Miami, FL, USA.

Ebert, S., Müller, T., and Schütze, H. (2016). LAMB: A Good Shepherd of Morphologically Rich Languages. In *Proc. of EMNLP*, pages 742–752, Austin, TX, USA.

Fishel, M. and Kirik, H. (2010). Linguistically Motivated Unsupervised Segmentation for Machine Translation. In *Proc. of LREC*, Valletta, Malta.

Fraser, A., Weller, M., Cahill, A., and Cap, F. (2012). Modeling Inflection and Word-Formation in SMT. In *Proc. of EACL*, pages 664–674, Avignon, France.

García-Martínez, M., Barrault, L., and Bougares, F. (2016). Factored Neural Machine Translation. In *Proc. of IWSLT*, Seattle, WA, USA.

García-Martínez, M., Barrault, L., and Bougares, F. (2017). *Neural Machine Translation by Generating Multiple Linguistic Factors*, pages 21–31. Springer International Publishing, Cham.

Goldwater, S. and McClosky, D. (2005). Improving Statistical MT through Morphological Analysis. In *Proc. of EMNLP*, pages 676–683, Vancouver, BC, Canada.

Habash, N. and Sadat, F. (2006). Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proc. of NAACL*, pages 49–52, New York City, USA.

Hasan, S., Mansour, S., and Ney, H. (2011). A comparison of segmentation methods and extended lexicon models for Arabic statistical machine translation. *Machine Translation*, pages 1–19.

Heigold, G., Neumann, G., and van Genabith, J. (2016). Neural morphological tagging from characters for morphologically rich languages. *arXiv preprint arXiv:1606.06640*.

Huck, M. and Birch, A. (2015). The Edinburgh Machine Translation Systems for IWSLT 2015. In *Proc. of IWSLT*, pages 31–38, Da Nang, Vietnam.

Huck, M., Braune, F., and Fraser, A. (2017a). LMU Munich's Neural Machine Translation Systems for News Articles and Health Information Texts. In *Proc. of WMT*, pages 315–322, Copenhagen, Denmark.

Huck, M., Riess, S., and Fraser, A. (2017b). Target-side Word Segmentation Strategies for Neural Machine Translation. In *Proc. of WMT*, pages 56–67, Copenhagen, Denmark.

Huck, M., Tamchyna, A., Bojar, O., and Fraser, A. (2017c). Producing Unseen Morphological Variants in Statistical Machine Translation. In *Proc. of EACL*, pages 369–375, Valencia, Spain.

Kann, K. and Schütze, H. (2016). MED: The LMU System for the SIGMORPHON 2016 Shared Task on Morphological Reinflection. In *Proc. of SIGMORPHON*, pages 62–70, Berlin, Germany.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of MT Summit*, Phuket, Thailand.

Koehn, P. and Haddow, B. (2012). Interpolated Backoff for Factored Translation Models. In *Proc. of AMTA*, San Diego, CA, USA.

Koehn, P. and Hoang, H. (2007). Factored Translation Models. In *Proc. of EMNLP-CoNLL*, pages 868–876, Prague, Czech Republic.

Labeau, M., Löser, K., and Allauzen, A. (2015). Non-lexical neural architecture for fine-grained POS Tagging. In *Proc. of EMNLP*, pages 232–237, Lisbon, Portugal.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In *Proc. of NAACL*, pages 260–270, San Diego, CA, USA.

Lee, Y.-S. (2004). Morphological Analysis for Statistical Machine Translation. In *Proc. of NAACL*, pages 57–60, Boston, MA, USA.

Matthews, P. H. (1991). Morphology (Cambridge Textbooks in Linguistics). *Cambridge University*.

Minkov, E., Toutanova, K., and Suzuki, H. (2007). Generating Complex Morphology for Machine Translation. In *Proc. of ACL*, pages 128–135, Prague, Czech Republic.

Mueller, T., Schmid, H., and Schütze, H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In *Proc. of EMNLP*, pages 322–332, Seattle, WA, USA.

Müller, T., Cotterell, R., Fraser, A., and Schütze, H. (2015). Joint Lemmatization and Morphological Tagging with Lemming. In *Proc. of EMNLP*, pages 2268–2274, Lisbon, Portugal.

Niehues, J. and Cho, E. (2017). Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning. In *Proc. of WMT*, pages 80–89, Copenhagen, Denmark.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*, pages 311–318, Philadelphia, PA, USA.

Passban, P., Liu, Q., and Way, A. (2017). Providing Morphological Information for SMT Using Neural Networks. *PBML*, 108:271–282.

Pinnis, M., Krišlauks, R., Deksne, D., and Miks, T. (2017). *Neural Machine Translation for Morphologically Rich Languages with Improved Sub-word Units and Synthetic Data*, pages 237–245. Springer.

Popović, M. and Ney, H. (2004). Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In *Proc. of LREC*, pages 1585–1588, Lisbon, Portugal.

Popović, M., Vilar, D., Ney, H., Jovičić, S., and Šarić, Z. (2005). Augmenting a Small Parallel Text with Morpho-Syntactic Language Resources for Serbian-English Statistical Machine Translation. In *Proc. of the ACL Workshop on Building and Using Parallel Texts*, pages 41–48, Ann Arbor, MI, USA.

Santos, C. D. and Zadrozny, B. (2014). Learning Character-level Representations for Part-of-Speech Tagging. In *Proc. of ICML*, pages 1818–1826, Bejing, China.

Sennrich, R. et al. (2017). Nematus: a Toolkit for Neural Machine Translation. In *Proc. of EACL*, pages 65–68, Valencia, Spain.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proc. of ACL*, pages 1715–1725, Berlin, Germany.

Stymne, S., Holmqvist, M., and Ahrenberg, L. (2008). Effects of Morphological Analysis in Translation between German and English. In *Proc. of WMT*, pages 135–138, Columbus, OH, USA.

Tamchyna, A., Weller-Di Marco, M., and Fraser, A. (2017). Modeling Target-Side Inflection in Neural Machine Translation. In *Proc. of WMT*, pages 32–42, Copenhagen, Denmark.

Toutanova, K., Suzuki, H., and Ruopp, A. (2008). Applying Morphology Generation Models to Machine Translation. In *Proc. of ACL*, pages 514–522, Columbus, OH, USA.

Vania, C. and Lopez, A. (2017). From Characters to Words to in Between: Do We Capture Morphology? In *Proc. of ACL*, pages 2016–2027, Vancouver, Canada.

Williams, P. and Koehn, P. (2011). Agreement Constraints for Statistical Machine Translation into German. In *Proc. of WMT*, pages 217–226, Edinburgh, Scotland.

Williams, P. and Koehn, P. (2014). Using Feature Structures to Improve Verb Translation in English-to-German Statistical MT. In *Proc. of HyTra*, pages 21–29, Gothenburg, Sweden.