Modeling Target-Side Inflection in Neural Machine Translation

Aleš Tamchyna[1,2] and **Marion Weller-Di Marco**[1,3] and **Alexander Fraser**[1]
[1]LMU Munich, [2]Memsource, [3]University of Stuttgart
ales.tamchyna@memsource.com dimarco@ims.uni-stuttgart.de
fraser@cis.lmu.de

Abstract

NMT systems have problems with large vocabulary sizes. Byte-pair encoding (BPE) is a popular approach to solving this problem, but while BPE allows the system to generate any target-side word, it does not enable effective generalization over the rich vocabulary in morphologically rich languages with strong inflectional phenomena. We introduce a simple approach to overcome this problem by training a system to produce the lemma of a word and its morphologically rich POS tag, which is then followed by a deterministic generation step. We apply this strategy for English–Czech and English–German translation scenarios, obtaining improvements in both settings. We furthermore show that the improvement is not due to only adding explicit morphological information.

1 Introduction

Neural machine translation (NMT) has recently become the new state of the art. Despite a large body of recent research, NMT still remains a relatively unexplored territory.

In this work, we focus on one of these less studied areas, namely target-side morphology. NMT systems typically produce outputs word-by-word and at each step, they evaluate the probability of all possible target words. When translating to morphologically rich languages, due to the large size of target-side vocabularies, NMT systems run into scalability issues and struggle with vocabulary coverage.

Byte-pair encoding (BPE, Sennrich et al. (2016b)) is currently perhaps the most successful approach to addressing these problems. However, while BPE allows the system to generate any target-side word (possibly as a concatenation of smaller segments), it does not enable effective generalization over the many different surface forms possible for a single lemma, which had been shown to be useful in phrase-based SMT (Bojar and Kos, 2010).

We see three main problems associated with rich target-side morphology in NMT: (i) NMT systems have no explicit connection between different surface forms of a single target-side lexeme (lemma), leading to data sparsity, (ii) there is no explicit information about morphological features of target-side words, and (iii) NMT systems cannot systematically generate unseen surface forms of known lemmas: while the combination of subword segments obtained with BPE splitting can technically generate new forms, this is not a linguistically informed way to generate new words, and is furthermore restricted to "simple" concatenative word formation processes.

We propose a simple two-step approach to achieve morphological generalization in NMT. In the first step, we use an encoder-decoder NMT system with attention and BPE (Bahdanau et al., 2014; Sennrich et al., 2016b) to generate a sequence of interleaving morphological tags and lemmas. In the second step, we use a morphological generator to produce the final inflected output. This decomposition addresses all three of the problems outlined above:

- the presence of lemmas allows the system to model different inflections jointly and better capture lexical correspondence with the source,
- morphological information is explicit and allows the system to easily learn target-side morpho-syntactic patterns including agreement,

- unseen surface forms can be generated simply by combining a known lemma and a known tag.

While simple, the approach is very effective and leads to significant improvements in translation quality in a medium-resource setting for English-Czech translation. Similarly, experiments in an English–German setting lead to improved translation results and also show that the proposed strategy can be applied to other language pairs.

2 Two-Step NMT

We work within the standard encoder-decoder framework with an attention mechanism as proposed by Bahdanau et al. (2014), using the Nematus implementation (Sennrich et al., 2017). To model target-side morphology, the system is trained on an intermediate representation consisting of interleaved lemmas and morphological tags providing the full set of relevant inflection features. Decoding is followed by a second step which is fully deterministic. We use the predicted pairs of (tag+features, lemma) as input to a morphological generator which outputs the final inflected surface forms. In the rare cases where the generator fails to output any surface form, we simply output the lemma.

Our approach is inspired by the successful results of Nadejde et al. (2017), where the authors interleave target-side words and CCG supertags and observe improvements by learning to also predict the target-side syntax. Our experiments in the English–Czech translation task will, however, show that the improvement we obtain is not a similar effect, but instead requires the improved generalization obtained through mapping inflected forms to their lemmas and the ability to generate correct surface forms.

In this paper, we first apply our tag lemma strategy to an English–Czech translation setting. We show that it is effective and also investigate potential effects of tag prediction interacting with morphological generalization. A second set of experiments concerns English–German translation: here, the focus is rather put on modeling linguistic phenomena, including German word formation. While Czech has a more complex morphology than German, German has the additional problem of compounds that make translation challenging; one system variant thus includes simple compound handling.

3 Modeling Czech Morphology

Czech is a Slavic language with a rich inflectional morphology. There are seven cases for nouns and adjectives, four genders and two grammatical numbers. Surface forms of verbs follow complex rules as well, as they encode number, person, tense and several other phenomena. Due to its fusional nature, there is a degree of syncretism in Czech – words with different morphological features may share the same surface form.

As such, Czech is a suitable example for evaluating our approach. We use the Czech positional tagset in our work (Hajič and Vidová-Hladká, 1998). Figure 1 illustrates the input and output to our network and the baseline. Figure 2 illustrates the tagset on an example. For Czech morphological analysis, tagging and generation, we use the MorphoDiTa toolkit (Straková et al., 2014), which achieves state-of-the-art results in lemmatization and tagging and its coverage in morphological generation is very high. Morphological generation is based on a lexicon of lemmas and their paradigms and it is fully deterministic.

4 Modeling German Morphology

To obtain the representation of interleaved lemmas and tag+feature sequences for German, we apply a slightly different pipeline than for the English–Czech setting. Instead of representing a word by a simple lemma and a morphological tag, we use a morphological analyzer covering also productive formation processes – the morphologically complex analyses of the lemma ("stem") allow us to easily handle compounds, which pose a considerable challenge when translating into German.

4.1 Linguistic Resources

The key linguistic knowledge sources to model German morphology are the constituency parser BitPar (Schmid, 2004) to obtain morphological analyses in the sentence context, and the morphological tool SMOR (Schmid et al., 2004) to analyze and generate inflected German surface forms.

SMOR is a a morphological analyzer for German inflection and word formation processes implemented in finite state technology. In particular, it also covers productive word formation processes such as compounding or derivation. SMOR functions in two directions: *surface form* $\rightarrow$ *stem+features* and *stem+features* $\rightarrow$ *surface form*. Thus, when preparing the target-side training data,

input:	there are a million different kinds of pizza .
baseline:	existují miliony druhů piz@@ zy .
morphgen:	VB-P—3P-AA— existovat NNIP1——A—- milión NNIP2——A—- druh NNFS2——A—- pizza Z:————- .

Figure 1: Examples of input and output training sequences for the baseline and the proposed system. BPE splits are denoted by "@@".

Category	Value	Description
POS	A	adjective
sub-POS	A	adjective, general
gender	I	masculine inanimate
number	P	plural
case	7	instrumental
possgender	–	*(possessor's gender)*
possnumber	–	*(possessor's number)*
person	–	*(person, verbs)*
tense	–	*(tense, verbs)*
grade	2	comparative degree
negation	A	affirmative (not negated)
voice	–	*(voice, verbs)*
reserve1	–	*(unused)*
reserve2	–	*(unused)*
var	–	*(style, variant)*

Figure 2: Czech positional tagset. Feature values for the word *kulatějšími*, tag `AAIP7----2A----`.

each inflected surface form is analyzed, and then replaced by its stem and respective morphological features, as illustrated for the verb *trifft* below:

surface	*trifft*
stem	`treffen<+V><3><Sg><Pres><Ind>`

For the inflection process after translation, SMOR is used in the reverse direction to output an inflected form when given a stem+feature sequence.

4.2 German Inflectional Features

German has a rich nominal and verbal morphology, and even though it exhibits a relatively high degree of syncretism, it has a high lemma–to–inflected forms ratio. For example, adjectives can have up to 6 different inflected forms, such as *blau, blaue, blaues, blauer, blauen, blauem* ('blue').

Nominal Inflection Unlike in English, where only the feature number is expressed for nouns, German nominal inflection is applied to determiners, adjectives and nouns. The following four features are relevant for nominal inflection:

case	*nominative, accusative, dative, genitive*
gender	*feminine, masculine, neuter*
number	*singular, plural*
str/wk	*strong, weak*

To efficiently handle syncretism, SMOR has the artificial value *NoGend*, that is used when a surface form is the same for all three values of gender; this is typical for plural forms. Similarly, the feature strong/weak[1] does not need to be specified if the surface forms are the same; we thus add the dummy-value `<NA>` to always have a sequence of four values. Words that are subject to nominal inflection are replaced by their SMOR analysis that is split into stem and the tag-feature sequence:

`STEM <+Tag><Gend><Case><Num><St/Wk>`

Verbal Morphology German verbal morphology requires the modeling of these features:

person	*1,2,3*
number	*singular, plural*
tense	*present, past*
mood	*indicative, subjunctive*

These features refer to morphologically expressed properties in a single word; further instances of the feature tense, in particular future tense, are realized as compound tenses. Our modeling of verbal inflection, is restricted to the word-level, and the decision how to combine auxiliaries and full verbs is left to the translation model. Verb forms are represented as follows in the stemmed format:

finite	`STEM <+V><Pers><Num><Tense><Mood>`
participle	`STEM <+V><PPast>`
infinitive	`STEM <+V><Inf>`

4.3 Building the stemmed representation

Table 1 illustrates the process of deriving the fully specified stemmed representation by combining morphological analyses and rich parse tags; the column *infl* indicates whether a word is inflected. As a German surface form can have many possible analyses (cf. below), the parse tags are needed to

[1] Strong/weak inflection is determined by the setting of definite/indefinite articles in combination with the other feature: for example, the NP *das blaue Auto* ('the blue car') is inflected differently when occurring with an indefinite article (*ein blaues Auto*) in the function of subject or direct object.

English sentence	and what you 're seeing here is a cloud of densely packed , hydrogen-sulfide-rich water coming out of a volcanic axis on the sea floor

EN gloss	**DE surface**	**parse-tags**	**infl.**	**fully specified stemmed representation**
and	und	KON	0	`und[KON]`
here	hier	ADV	1	`hier[ADV]`
sees	sieht	VVFIN-Sg	1	`sehen||<+V><3><Sg><Pres><Ind>`
one	man	PIS-Nom.Sg	0	`man[PIS]`
a	eine	ART-Acc.Sg.Fem	1	`eine<Indef>||<+ART><Fem><Acc><Sg><St>`
cloud	Wolke	NN-Acc.Sg.Fem	1	`Wolke||<+NN><Fem><Acc><Sg><NA>`
of	von	APPR-Dat	0	`von[APPR-Dat]`
dense	dichtem	ADJA-Dat.Sg.Neut	1	`dicht<Pos>||<+ADJ><Neut><Dat><Sg><St>`
hydrogen-sulfide-rich	hydrogensulfid-reichem	ADJA-Dat.Sg.Neut	1	`Hydrogen<NN>Sulfid<NN>` `reich<Pos>||<+ADJ><Neut><Dat><Sg><St>`
water	Wasser	NN-Dat.Sg.Neut	1	`Wasser||<+NN><Neut><Dat><Sg><NA>`
,	,	$,	0	`,[$]`
that	das	PRELS-Nom.Sg.Neut	0	`das[PRELS]`
from	aus	APPR-Dat	0	`aus[APPR-Dat]`
a	einer	ART-Dat.Sg.Fem	1	`eine<Indef>||<+ART><Fem><Dat><Sg><St>`
volcanic	vulkanischen	ADJA-Dat.Sg.Fem	1	`vulkanisch||<+ADJ><Pos>` `<NoGend><Dat><Sg><Wk>`
longitudinal axis	Längsachse	NN-Dat.Sg.Fem	1	`längs<ADJ>Achse||<+NN><Fem><Dat><Sg><NA>`
on	an	APPR-Dat	0	`an[APPR-Dat]`
the	dem	ART-Dat.Sg.Masc	1	`die<Def>||<+ART><Masc><Dat><Sg><St>`
sea floor	Meeresboden	NN-Dat.Sg.Masc	1	`Meer<NN>Boden||<+NN><Masc><Dat><Sg><NA>`
oozes	tritt	VVFIN-Sg	1	`treten||<+V><3><Sg><Pres><Ind>`
.	.	$.	0	`.[$]`

Table 1: Example for the fully specified representation used in the NMT system. The double-pipe symbol || indicates the boundary between the word(stem) and the tag with the full set of inflectional features.

disambiguate the morphological analyses.

vulkanischen

```
vulkanisch<+ADJ><Pos><Neut><Gen><Sg>
vulkanisch<+ADJ><Pos><Masc><Acc><Sg>
vulkanisch<+ADJ><Pos><Masc><Gen><Sg>
vulkanisch<+ADJ><Pos><NoGend><Acc><Pl><Wk>
vulkanisch<+ADJ><Pos><NoGend><Dat><Pl>
vulkanisch<+ADJ><Pos><NoGend><Dat><Sg><Wk>
vulkanisch<+ADJ><Pos><NoGend><Gen><Pl><Wk>
vulkanisch<+ADJ><Pos><NoGend><Nom><Pl><Wk>
vulkanisch<+ADJ><Pos><Fem><Gen><Sg><Wk>
```

The stem and the tag-feature sequence (or the bare tag for non-inflected words) are separated, allowing the model to learn lexical relations between source- and target-side separately from target-side morpho-syntactic patterns. As the addition of tags effectively doubles the length of German sentences, we also add tags (obtained with treetagger, Schmid (1994)) on the source-side to balance the source/target side sentence lengths.

4.4 Reduction of Vocabulary Size

One of the main objectives of the two-step approach is to reduce the target-side vocabulary size. Table 2 shows the most frequent fragments on the end of words obtained through BPE splitting on the German surface data – while it is difficult to generalize without the actual context, most tend to be inflectional suffixes. While this type of splitting does make sense, it also seems that there is some redundancy, and a systematic generalization is impossible. Furthermore, a mere segmentation of surface forms does not cover non-concatenative phenomena such as "Umlautung": for example, the concatenation of *Haus-* (lemma: 'house') and *-er* (typical plural suffix) does not result in the correct plural form (*Häuser*) – thus, two "lemmas" are required to guarantee correct inflections of words that undergo Umlautung when working with surface forms. Table 3 shows the reduction of vocabulary in the stemmed representation: replacing inflected forms with their stems leads to

freq	part	freq	part	freq	part
2469	ten	1257	sten	1077	ern
2157	te	1214	es	1077	-
1738	en	1169	ter	1058	den
1607	er	1148	gen	1040	s
1474	ung	1 078	ischen	1015	ungen

Table 2: The most frequent fragments on word ends after BPE from the German surface data.

	vocabulary size	vocabulary size w/ BPE
DE surface data	121.892	22.712
DE morph	97.587	21.663
DE morph-split	68.533	21.892

Table 3: Overview of vocabulary size in the German TED data (BPE: Byte Pair Encoding).

a considerable reduction of the vocabulary size; compound splitting leads to a further reduction.

4.5 Simple Compound Handling

Another factor contributing to a high vocabulary size is the productivity of German compounds; in SMT, compound handling has been found to improve translation quality, e.g. Stymne et al. (2011) and Cap et al. (2014). In addition to inflectional morphology, SMOR also provides a derivational analysis, including splitting into compound parts: for example, the compound *Häuser|markt* ('house market') is analyzed as `Haus<NN>Markt<+NN><...>`. In particular, the modifier is represented by its base form `Haus`, covering the non-concatenative process of "Umlautung" (*Haus* ↔ *Häuser*).

In the stemmed representation, this may already present an indirect advantage, as compounds fragmented through BPE splitting can match other stemmed occurrences of that word. An obvious idea at this point is to go a step further and add compound splitting to the pre-processing of the German data. Using the SMOR annotation, compounds are split at mid-word adjective and noun borders. For example, the word *Meeres|boden* ('sea bottom') from table 1 is split into two subwords separated by the modifier's tag:

```
Meer §§<NN>§§ Boden <+NN><...>
```

This notation separates lexical parts from SMOR markup, thus allowing the model to learn compound patterns. After translation, the compound stems are concatenated and then inflected.

On the English side, it is assumed that the equivalents of compounds are already separate words. For this system variant, however, the English side was slightly simplified by aggressive hyphen splitting, and replacing nouns and verbs by their lemma form, accompanied by a tag indicating the type of inflection. Our hope is that this representation will be more parallel to the compound-split representation in German.

corpus	sents	src tokens	tgt tokens
train	114k	2309k	1908k
test2012	1385	25150	20682
test2013	1327	28454	24107

Table 4: Sizes of English-Czech corpora.

5 Experimental Evaluation

In this section, we describe our experiments with English-Czech and English-German translation.

5.1 Czech

We use the IWSLT training and test sets in English-Czech experiments[2]. The training set consists of transcribed TED talks as collected in the WIT3 corpus (Cettolo et al., 2012). We use IWSLT test set 2012 as the held-out set and the 2013 test set for evaluation. Table 4 summarizes the basic data statistics.

We use the Nematus toolkit for training the NMT systems (Sennrich et al., 2017). We run BPE training on both sides of the training data with 49500 splits. We set the vocabulary size to 50000 word types. The embedding size is set to 500, the dimension of the hidden layer is 1024. We optimize the model using Adam (Kingma and Ba, 2014) and we use the default early stopping criterion in Nematus. We do not apply drop-out anywhere in the model. Following Nadejde et al. (2017), we set the maximum sequence length to 50 for the baseline and to 100 for systems which produce interleaved outputs.

Our *baseline* system is a standard Nematus setup with the parameters described above. We refer to our two-step setup as *morphgen* from now on. For comparison, we also evaluate a third setting where we train the system to output sequences of morphological tags interleaved with the surface

[2] http://workshop2016.iwslt.org

system	BLEU (dev)	BLEU (test)
baseline	12.60	12.89
morphgen	**14.05**	**14.57**
serialization	11.49	12.07

Table 5: English-Czech: BLEU scores of NMT system variants.

forms. We refer to this contrastive experiment as *serialization* – our aim is to tease apart the possible benefit of explicitly predicting target-side morphological tags from the improvements due to morphological generalization.

Note that BPE is applied in all system variants. However, due to a reduced vocabulary size in the *morphgen* setting, the splits are uncommon and morphological tags are never split (this is an effect of BPE, not a hard constraint).

Because NMT system results can vary significantly due to randomness in initialization and training, we run system training end-to-end for each variant three times. We then select the best run based on BLEU as measured on the development set (test2012) and then evaluate it on the final test set (test2013).

Importantly, the network was able to learn the correct structure for both *morphgen* and *serialization* systems. The outputs are well-formed sequences of interleaving tags and lemmas/forms.

Table 5 shows the obtained results. In our main experiment, our two-step system achieves a substantial improvement of roughly 1.7 BLEU points, showing that two-step in the neural context works for English to Czech translation for this data size.

In the serialization experiment, we see that, surprisingly, the *serialization* system does not outperform the baseline setup. This stands in contrast to the use of CCG supertags by Nadejde et al. (2017), which was effective in this framework. The result there showed that using CCG supertags which handle syntactic generalization helps produce a better sequence of surface forms. We attribute our result to the trade-off between providing the system with explicit morpho-syntactic information (which is weaker information than CCG supertags) and increasing the sequence length (which complicates training). It is possible that with larger training data, *serialization* might still outperform the baseline, but our main result has shown that morphological generalization on this data size is beneficial.

	baseline	morphgen	Δ
IWSLT	12.89	14.57	1.68
250k	14.87	17.51	2.64
500k	16.96	20.05	**3.09**
1M	18.07	20.95	2.88
2M	20.04	**22.31**	2.27

Table 6: English-Czech: BLEU scores of systems with larger parallel training data.

Scaling to Larger Data The observed improvements are certainly at least partially due to reduced data sparsity: because Czech is a morphologically rich language, there is a high number of distinct surface forms. We help the system generalize by essentially dividing the information that surface forms carry into two different "streams": one for morpho-syntax (tags) and the other for semantics (lemmas).

One possible concern with the proposed approach is the ability to scale to larger training data. Data sparsity could be such a major issue only when training data are small and once we scale up, the observed benefits might disappear as the system gets more robust statistical estimates for the individual surface forms.

We run a targeted experiment with larger sizes of parallel training data to determine whether the improvements hold. We always use the main training set described above but additionally, we add a random sample from the CzEng 1.0 parallel corpus (Bojar et al., 2012) to achieve training data sizes of 250 thousand up to 2 million parallel sentences (total).

Table 6 shows the results. We observe the highest difference in the 500k setting (over 3 BLEU points absolute) and while the improvement decreases slightly as we add more data, the difference is still around 2.3 BLEU points even in the largest evaluated setting, which is an encouraging result.

Note that due to the increased computational cost, scores for larger system variants are only based on a single training run.

Analysis and Discussion We now further analyze our two-step system, *morphgen*, in the IWSLT data setting. We first look at cases where the generator failed to produce the surface form. We found only a handful of cases; these mostly involved unknown proper names (Braper, Hvanda).

In just four cases, the tag proposed by the network was not compatible with the lemma (i.e., the network made an error).

In order to determine where the improvement comes from, we analyze the number of novel surface forms produced by the system. We find that indeed, unseen word forms *are* generated by the system but not nearly as many as we expected: only 125 novel tokens were found in the test set (114 word types). Out of these, 14 forms are confirmed by the reference sentences (note that the unconfirmed words may still be correct within the system output).

It seems that the system mostly benefits from the decomposition that we proposed – Czech lemmas are more easily mapped to source-side English words than the many inflected forms associated with each lemma. The interleaving tags then help explicitly train the morpho-syntactic structure of the sentences and allow the second step to deterministically generate the final translations. While morphological generalization does indeed occur, it is not the source of most of the observed improvement. When we use surface forms together with the annotations (in our serialization experiment), we see no improvement.

Finally, we report the results of a blind manual annotation contrasting outputs of *baseline* and *morphgen*. For each instance, the annotator had access to the reference translation and both outputs. The task was to rank which translation is better or to mark both as equal quality. The annotator analyzed 200 sentences. In 130 cases, the translations were judged as equal. Out of the remaining 70 sentences, the *morphgen* system was marked as better in 48 cases and the baseline won in 22 cases.

5.2 German

The initial English–German experiments are evaluated on IWSLT training and test data, which consists of transcribed TED talks. The system is optimized on the 2012 dev-set (1165 sentences), and tested on the 2013 test-set (1363 sentences) and the 2014 test-set (1305 sentences). The training data consists of 184.879 parallel sentences, after filtering out sentences shorter than 5 or longer than 50 words, as well as sentences that could not be parsed. Prior to training the NMT system, the (stemmed) source- and target-data undergo BPE splitting (29500 splits), in order to keep the vocabulary within the predefined limit.

TED'13	**run-1**	**run-2**	**avg.**
baseline	19.87	20.15	20.01
morph-gen	20.73	20.98	20.86
morph-gen-split	20.88	21.18	21.03

TED'14	**run-1**	**run-2**	**avg.**
baseline	19.02	18.68	18.85
morph-gen	20.01	19.93	19.97
morph-gen-split	20.07	20.76	20.42

Table 7: English–German: lowercased BLEU for two test sests (1363 and 1305 sentences).

	baseline	**morph-gen**	**morph-gen-split**
250k	18.75	20.55	20.51
500k	21.39	22.79	23.00

Table 8: English–German: lowercased BLEU for newstest'16 (2169 sentences) trained on 250k and 500k sentences news-mix data.

The translation experiments are carried out with the Nematus toolkit (Sennrich et al., 2017), using the training parameters as displayed below, in combination with the default early stopping criterion in Nematus:

vocab	30k	dropout	yes
dim_word	500	dropout_emb	0.2
dim	1024	dropout_hid	0.2
lrate	0.0001	dropout_src	0.1
opt	adam	dropout_trg	0.1
maxlen	50(100)		

The sentence length is set to 50 for the baseline system, and extended to 100 for the morph-gen systems, because the addition of the morphological tags doubles the sentence length.

Table 7 shows the results for the English–German translation experiments, averaged over two training runs: on both test sets, the system generating inflected forms based on stems and features is better than the baseline.

Despite SMOR's complicated structure, the resulting stems are generally well-formed; for uninflectable stems (mostly made-up words such as `Parunelogramm<+NN><Neut><Gen><Sg>`), the markup is simply removed.

The addition of compound splitting leads to a minor further improvement. We consider this a promising result, indicating that segmentation using the rich information provided by SMOR can be helpful; we plan to explore this further in future work.

Generation of novel words A closer look at the translation output reveals that there are indeed new word forms generated by the *morph-gen* system. For the TED'13 set, for example, the *morph-gen* system output a total of 261 words that are not in the training data or the English input sentence. Of these, 112 are names or nonsense words produced by concatenating BPE segments[3]. The other 149 words are morphologically well-formed, though not necessarily semantically sound (e.g. *Schokoladenredakteur*: 'chocolate editor' as proposed translation for 'smart-ass editor') or appropriate in the translation context. Thus, we compared the novel words with the reference translations: 23 words (21 nouns, 2 adjectives) were found in the reference of the respective sentence. Of course, this under-estimates the number of useful new creations, as a valid translation does not necessarily need to match exactly with the reference. For the *morph-gen-split* system, only 27 matches with the reference were found in a set of 328 unseen forms.

Different Domain and Larger Corpus To assess the influence of domain and corpus size, we also evaluate the approach to model German morphology in a larger news corpus setting. To obtain a training corpus that is diverse, but still restricted in size, we combined randomly selected sentences (between 5-50 words) from the 4 parallel corpora provided for EN–DE translation at the WMT'17 shared task[4] (selected in equal parts from Europarl, CommonCrawl, News-Commentary and RapidCorpus), resulting in a set of 250k and 500k sentences The model is optimized on newstest'15 and evaluated on newstest'16; table 8 shows the results for the surface form baseline and the morphological generation systems with and without compound handling. As for the TED data set, the morphological generation systems outperforms the systems trained on surface data, but the improvement for the system trained on 500k sentences is slightly lower than for the system trained on 250k sentences. The systems with additional compound splitting obtained the same result as the basis morphological generation system (250k), or were slightly better (500k). With regard to the effectiveness of compound handling, it is difficult to draw a clear conclusion, but, looking also at the results obtained in the TED setting, it seems that there is a tendency that compound handling leads to a slight improvement. As compounding is a productive word formation process that is challenging to cover even in large corpora, compound handling might be useful also when using larger data training corpora.

[3]Into this category, we also count non-wellformed generations by SMOR caused by incorrect transitional elements in compounds, e.g. *Oszillation**en**generator* vs. *Oszillation**s**generator*.

[4]http://www.statmt.org/wmt17/translation-task.html

6 Related Work

Generation of unseen morphological variants has been tackled in various ways in the context of phrase-based models and other SMT approaches. Notably, two-step SMT was proposed to address this problem (Toutanova et al., 2008; Bojar and Kos, 2010; Fraser et al., 2012). In two-step SMT, a separate prediction model (such as a linear-chain CRF) is used to either directly predict the surface form (as in Toutanova et al. (2008)) or used to predict the grammatical features, following which morphological generation is performed (as in Bojar and Kos (2010); Fraser et al. (2012)). Our work differs from their work in that we do not use a separate prediction model, but instead rely on predicting the lemmas and surface-forms as a single sequence in a neural machine translation model.

Huck et al. (2017b) recently proposed an approach related to two-step MT where the unseen surface forms are added as synthetic phrases directly in the system phrase table and a context-aware discriminative model is applied to score the unseen variants. Unlike our work, the authors report diminishing improvements as training data grows larger. Our approach learns a more robust underlying model thanks to the reduced data sparsity. Unlike Huck et al. (2017b), our improvements are therefore not only due to the ability to generate words which were not seen in the training data.

Factored translation models (Koehn and Hoang, 2007) can deal with unseen word forms thanks to generation steps. One of the original goals of factored MT was in fact the scenario where the system produces lemmas and tags and then a generation step could be used to produce the inflected forms. Factored models failed to achieve this goal due to lemmas and tags being predicted independently, leading to many invalid combinations, and due to the involved combinatorial explosion.

García-Martínez et al. (2016) attempt to include target-side factors in neural MT. Unlike our simple technique, their approach requires modifications

to the network architecture. The authors work with English-French translation and they report mixed results.

Another successful attempt to learn novel inflections in SMT is back-translation (Bojar and Tamchyna, 2011). By using an MT system trained to translate *lemmas* in the opposite direction, it is possible to create synthetic parallel data which contain unseen word forms of known lemmas on the target side. There are two main downsides to this approach. The first is that the source language contains translation errors, which may affect translation quality. The second is that the substitution of different surface forms for the same target language lemma may result in incoherent translations, where the context no longer agrees with the chosen surface form. Sennrich et al. (2016a) propose to use back-translation in NMT to include language modeling data, but the "inverse" NMT system is not able to translate unseen target word forms (no lemmatization is done) and therefore this method does not learn novel inflections. Applying BPE splitting can technically lead to new inflected word forms, but this requires an appropriate segmentation into base form and inflectional suffixes which might not always be the case, in particular for infrequent words.

A very similar method to our two-step setting was independently proposed for use in a natural language generation (NLG) pipeline for morphologically rich languages (Dušek, 2017). However, in this scenario, the approach was not better than a baseline which operated on surface forms.

Finally, there has been further more recent work on alternatives to using BPE segmentation for NMT. Ataman et al. (2017) looked at segmentation for Turkish, which is an agglutinative language. Huck et al. (2017a) presents an approach for segmenting German with a focus on compound splitting and splitting suffixes off of stems using a stemmer, which may allow generalization in a similar way to our work. It would be interesting to compare with these approaches in future work.

7 Conclusion

In this work we showed that a simple setup, interspersing lemmas and rich morphological tags, followed by deterministic generation of the resulting surface form, results in impressive gains in NMT of English to Czech. Applying the technique to an English to German system also resulted in considerable improvements. For English–German, the addition of compound handling yielded promising results. Furthermore, among the novel word forms for German, most were compounds – as compounding is a very productive process, this is also a challenging problem when using larger corpora. Exploring strategies for better segmentation and compound handling is an interesting task that we plan to investigate further.

We believe that while simple, this technique effectively addresses the fundamental problems of rich target-side morphology: (i) sparse data and lack of connection between different forms of a single target lexeme, (ii) lack of explicit morphological information, and (iii) inability to generate unseen forms of known lexemes. Our results indicate that most of the improvement comes from the first two properties.

Perhaps a modified training criterion could be used to encourage the system to generalize more; in the standard setting, the system probably learns to strongly condition the lemma on the tag and avoids the risk of generating new pairs. In the situations where a novel form is required, the system may either bypass this by producing a synonymous word or paraphrase, or it might simply produce an ungrammatical form of the correct lemma. This phenomenon deserves more examination which we leave to future work.

We further analyzed the serialization scenario, showing that the effect here is not due to training the system to also predict morphological tags, which is in contrast with the result of Nadejde et al. (2017). It is likely that the two approaches are complementary, the rich information in CCG supertags could bring additional benefit to the morphological generalization that we perform. We plan to investigate this in future work.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402 (HimL). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 640550).

References

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. In *Proceedings of EAMT*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473. http://arxiv.org/abs/1409.0473.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proc. of LREC*. ELRA, pages 3921–3928.

Ondřej Bojar and Kamil Kos. 2010. 2010 failures in english-czech phrase-based mt. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, Stroudsburg, PA, USA, WMT '10, pages 60–66. http://dl.acm.org/citation.cfm?id=1868850.1868855.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, pages 330–336. http://www.aclweb.org/anthology/W11-2138.

Fabienne Cap, Alexander Fraser, Marion Weller, and Aoife Cahill. 2014. How to Produce Unseen Teddy Bears: Improved Morphological Processing of Compounds in SMT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Göteborg, Sweden, pages 579–587.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*. Trento, Italy, pages 261–268.

Ondřej Dušek. 2017. *Novel methods in natural language generation for spoken dialogue systems*. Ph.D. thesis.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling Inflection and Word-Formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon, France, pages 664–674. http://www.aclweb.org/anthology/E12-1068.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation. *CoRR* abs/1609.04621. http://arxiv.org/abs/1609.04621.

Jan Hajič and Barbora Vidová-Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of the COLING - ACL Conference*. pages 483–490.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017a. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation (WMT)*. Copenhagen, Denmark.

Matthias Huck, Aleš Tamchyna, Ondřej Bojar, and Alexander Fraser. 2017b. Producing unseen morphological variants in statistical machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, Prague, Czech Republic, pages 868–876. http://www.aclweb.org/anthology/D/D07/D07-1091.

Maria Nadejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. Syntax-aware neural machine translation using CCG. *CoRR* abs/1702.01147. http://arxiv.org/abs/1702.01147.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, pages 44–49.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the International Conference on Computational Linguistics*. Geneva, Switzerland, pages 162–168.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal, pages 1263–1266.

R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. Valerio Miceli Barone, J. Mokry, and M. Nădejde. 2017. Nematus: a Toolkit for Neural Machine Translation. *ArXiv e-prints* .

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation

models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*. http://aclweb.org/anthology/P/P16/P16-1009.pdf.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*. http://aclweb.org/anthology/P/P16/P16-1162.pdf.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 13–18. http://www.aclweb.org/anthology/P/P14/P14-5003.pdf.

Sara Stymne, Nicola Candedda, and Lars Ahrenberg. 2011. Generation of Compound Words in Statistical Machine Translation into Compounding Languages. *Computational Linguistics* 39(4):1067–1108.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, Columbus, Ohio, pages 514–522. http://www.aclweb.org/anthology/P/P08/P08-1059.