Modeling Word Formation in English–German Neural Machine Translation

Marion Weller-Di Marco and **Alexander Fraser**
Center for Information and Language Processing
LMU Munich
{dimarco,fraser}@cis.uni-muenchen.de

Abstract

This paper studies strategies to model word formation in NMT using rich linguistic information, namely a word segmentation approach that goes beyond splitting into substrings by considering fusional morphology. Our linguistically sound segmentation is combined with a method for target-side inflection to accommodate modeling word formation. The best system variants employ source-side morphological analysis and model complex target-side words, improving over a standard system.

1 Introduction

A major problem in word-level approaches to MT is a lack of morphological generalization. Both inflectional variants of the same lemma and derivations of a shared word stem are treated as unrelated. For morphologically complex languages with a large vocabulary, this is problematic, especially in low-resource or domain-adaption scenarios.

A simple and widely used approach to reduce a large vocabulary in NMT is *Byte Pair Encoding (BPE)* (Sennrich et al., 2016), which iteratively merges the top-frequent bigrams from initially character-level split words until a set vocabulary size is reached. This strategy is effective, but linguistically uninformed and often leads to sub-optimal segmentation. Also, by only segmenting words into substrings, BPE cannot handle non-concatenative operations, for example:

- umlautung: $Baum_{Sg} \rightarrow B\ddot{a}ume_{Pl}$ (*tree/trees*),
- transitional elements that frequently occur in German compounds: *Grenz|kontroll|politik → Grenze, Kontrolle* (*border control policy*)
- derivation: *abundant ↔ abundance*

In this paper, we apply word segmentation on both the source and target sides that goes beyond merely splitting into exact substrings. This overcomes the issues caused by fusional morphology by accomodating modeling word formation across languages.

Productive word formation can lead to a high number of infrequent word forms even though the morphemes in these words are frequent. A linguistically motivated segmentation method to handle processes such as compounding and derivation allows for better coverage and generalization, both on the word level and on the morpheme level, and also enables the generation of new words. Sound morphological processing on the source and target side aims at learning productive word formation processes during translation, such as the English-German translation pair *ungovernability↔Unregierbarkeit*:

un|PREF govern|V able|SUFF-ADJ ity|SUFF-NOUN ↔ un|PREF regieren|V bar|SUFF-ADJ keit|SUFF-NOUN

Morphological information should not only handle isomorphic translation equivalents as above, but also help to uncover relations between source and target side for structurally different translations.

2 Related work

There is a growing interest in the integration of linguistic information in NMT. For example, Eriguchi et al. (2016) and Bastings et al. (2017) demonstrate the positive impact of source-side syntactic information; Nădejde et al. (2017) report improved translation quality when using syntactic information in form of CCG tags on source and target side.

To address data sparsity, compound modeling has already been proved useful for phrase-based MT, e.g., Koehn and Knight (2003) who model source-side compounds, and Cap et al. (2014) who generate compounds on the target side. For NMT, Huck et al. (2017) apply compound and suffix segmentation using a stemmer. Ataman et al. (2017) reduce complex source-side vocabulary by means of an unsupervised morphology learning algorithm.

Morphological tag	Stem with morph. analysis	inflected form	gloss
[NN-Neut.Acc.Pl.NA]	Umwelt<NN>Kriterium	Umweltkriterien	*ecological criteria*
[APPR-Acc]	für	für	*for*
[ART-Fem.Acc.Sg.St]	die<Def>	die	*the*
[NN-Fem.Acc.Sg.NA]	ermitteln<V>ung<NN><SUFF>	Ermittlung	*detection*
[ADJ-NoGend.Gen.Pl.St]	Schutz<NN>bedürftig<Pos>	schutzbedürftiger	*in need of protection*
[NN-Fem.Gen.Pl.NA]	Meer<NN>Region	Meeresregionen	*marine regions*
[APPR-Dat]	in	in	*in*
[ART-Fem.Dat.Sg.St]	die<Def>	der	*the*
[NN-Fem.Dat.Sg.NA]	tief<Pos><ADJ>See	Tiefsee	*open (lit: deep) sea*

Table 1: Representation of the training data for basic inflection prediction. Corresponding English sentence: *(the ninth meeting should adopt) ecological criteria to determine marine regions in the open sea that need protection.*

Ataman and Federico (2018) forego a traditional morphological analysis of the source language, and instead compose word representations from character trigrams. However, these three papers only apply segmentation on the string level and cannot properly handle fusional morphology.

Addressing morphology in NMT, Banerjee and Bhattacharyya (2018) combine BPE with a morphological analyzer to "guide" the segmentation of surface forms into substrings. Their approach does not result in morphemes, for example *googling → googl|ing*, which does not match with *google*, while in our work we match such morphemes. Tamchyna et al. (2017) present an NMT system to generate inflected forms on the target side, with a focus on overcoming data-sparsity caused by inflection. Their work contains a simple experiment on compound splitting with promising initial results that encouraged us to systematically explore word formation, including compounding, in NMT.

To model word formation, we investigate (i) source-side tags for shallow syntactic information; (ii) target-side segmentation relying on a rich morphological analysis; and (iii) source-side segmentation strategies also relying on a tool for morphological analysis. We show that combining these strategies improves translation quality.

Our contribution is a segmentation strategy that includes modeling non-concatenative processes, by implementing an English morphological analyzer suitable for this task, and by exploiting an existing tool for German, in order to obtain a consistent morphological sub-word representation.

3 Modeling target-side morphology

Our strategy to model word formation operates on lemma level as this allows for a better generalization than using surface forms. To model target-side inflection, we follow the simple lemma-tag generation approach by Tamchyna et al. (2017), but we improved the lemma representation to better support modeling word formation, and also implement a novel source-side morphological representation.

Lemma-tag generation (existing strategy): In a pre-processing step, all inflected forms of the target-side training data are replaced by pairs of the lemma and its rich morphological tag. In a post-processing step, the system's output is re-inflected by generating inflected forms using the morphological tool SMOR (Schmid et al., 2004). Table 1 depicts the process of inflecting tag and lemma pairs (columns 1, 2) into surface forms (column 3).

New selection of lemma analyses: SMOR is a finite-state based tool covering inflection and derivation; it outputs all possible analysis paths, i.e. analyses at different levels of granularity. While not much attention is paid to the lemma selection in Tamchyna et al. (2017), a carefully selected lemma-internal representation is crucial for modeling word formation, as it provides the basis for segmentation across morphemes. To obtain optimal analyses, we follow Koehn and Knight (2003), and use the frequencies of observed non-complex words (we ignore bound morphemes). We select the analysis with the highest geometric mean of the components' frequencies, which gives a preference to words occurring more frequently in the data. The modified selection strategy favors more complex analyses; table 2 shows some examples.

Old	atomwaffenfrei<+ADJ>
New	Atom<NN>Waffe<NN>frei<+ADJ>
	nuclear weapon free
Old	Forschung<NN>Ergebnis<+NN>
New	forschen<V>ung<NN><SUFF>Ergebnis<+NN>
	research result
Old	gefährlich<+ADJ>
New	Gefahr<NN>lich<SUFF><+ADJ>
	danger -ous

Table 2: Lemma analyses with better internal analysis. Note that the derivational morpheme *-ung* makes the verb *to research* into the noun *research*.

word	analysis
conspiracy	conspire|V acy|SUFF/N/e
conspiratorial	conspire|V ator|SUFF/N/e ial|SUFF/ADJ/-
conspirator	conspire|V ator|SUFF/N/e
conspire	conspire|V
ceremonial	ceremony|N ial|SUFF/ADJ/y
ceremony	ceremony|N
acquired	acquire|V ed|SUFF/ADJ/e
acquirer	acquire|V er|SUFF/N/e
acquire	acquire|V
acquisition	acquire|V ition|SUFF/N/s→re
acquisitive	acquire|V itive|SUFF/ADJ/s→re
acquisitiveness	acquire|V itive|SUFF/ADJ/s→re ness|SUFF/N/-

Table 3: English morphological analysis: the rightmost string on *SUFF* segments denotes a string operation, such as the removal of *e* in *conspire + acy*.

EN Morph-Markup-Split
enthusiasm <N> tic<SUFF_ADJ> ally<SUFF_ADV>
explode <V> ion<SUFF_N>
inevitable <ADJ> ly<SUFF_ADV>

EN Morph-noMarkup-Split
enthusiasm tic<SUFF_ADJ> ally<SUFF_ADV>
explode ion<SUFF_N>
inevitable ly<SUFF_ADV>

EN Morph-noMarkup-noSplit
enthusiasmtic<SUFF_ADJ>ally<SUFF_ADV>
explodeion<SUFF_N>
inevitablely<SUFF_ADV>

Table 4: Representation variants for the words *enthusiastically*, *explosion* and *inevitably*. This annotation replaces the lemma in the lemma-tag representation.

4 Simple English morphological analysis

We implemented a simple morphological analyzer that is generally based on Koehn and Knight (2003), in that a word is segmented into strings that are already observed in the training data. Our method additionally relies on tag information (similar to the compound splitter of Weller-Di Marco (2017)), and on a hand-crafted set of prefixes and suffixes in combination with rules such as $i \rightarrow y$ to handle non-concatenative transitions as in *beautiful* → *beauty\|N ful\|SUFF-ADJ*.

The segmentation is based on statistics derived from tagged and lemmatized data. This has several advantages: (i) the lemma and tag information restricts the possible operations (e.g., *-ion* as suffix is only applicable to nouns); (ii) there is no need to handle inflection; (iii) the tag provides a flat morpho-syntactic structure of the segmented word.

The analysis first identifies a potential prefix by finding a combination with a prefix in the training data, for example *deactivation\|N → de\|PREF activation\|N*. The tag restriction at this step is important to maintain the word class of the original word, and to avoid analyses such as *decent\|ADJ → de\|PREF cent\|N*. The remaining part undergoes splitting into either word+suffix (e.g., *activation\|N → activate\|V ion\|SUFF-N*) or a combination of two words (e.g., *evildoer\|N → evil\|N doer\|N*) until no further segmentation can be found. In case of several possibilities, the analysis whose components lead to the highest geometric mean is selected.

Table 3 illustrates how the morphological segmentation makes the word parts accessible such that they match with other occurrences of the word.

The splitter in its present form is rather aggressive and tends to oversplit. While it is often assumed that this is not harmful in MT (e.g., Koehn and Knight (2003)), we have not investigated the impact of oversplitting vs. undersplitting.

5 Data representation in NMT

The morphological analyses provide a straightforward basis for the segmentation experiments.

German: The lemma-tag approach (*oldLemTag*) is contrasted to the system variant with new lemma selection (*newLemTag*). For the segmentation experiments (*newLemTagSplit*), we apply compound splitting, such as `Gold<NN>Preis<NN>` → `Gold<NN> Preis<NN>` (*gold price*). Also, nominalization, e.g., `regieren<V> ung<NN><SUFF>` (*govern ment*), is segmented, but different adjective suffixes (such as *-lich*) are kept attached. Generally, we found that variation of the splitting granularity of adjective suffixes does not have a large impact.

English: We first look at a representation where lemma-tag pairs replace surface forms (*LemTag*). To evaluate the effect of morphological information, we compare the three settings in table 4 that also rely on the lemma-tag representation: the tags convey inflectional information, but the lemma is replaced by its morphological analysis.

In *Morph-Markup-Split*, words are split following the analysis, with tags indicating word-internal structure. *Morph-noMarkup-Split* is the same, but without word-internal tags. The annotation of prefixes/suffixes (*-ion<SUFF-N>*) is always kept.

In addition to explicit splitting, we consider a variant where lemmas are replaced by the unsplit morphological analysis (*Morph-noMarkup-noSplit*), and all segmentation is done by BPE, which can now access actual words (*enthusiasm* instead of **enthusias*) that already occur in the training data. This representation is conceptionally similar to the German lemma-tag representation.

System	Source (EN)	Target (DE)	Small	Medium	Large2M	Large4M
1	plain	plain	21.77	26.60	28.66	33.71
2	plain	oldLemTag	22.25	26.96	28.87	33.97
3	plain	newLemTag	22.47	27.05	28.61	33.90
4	LemTag	newLemTag	23.32	27.36	28.88	34.28
5	LemTag	newLemTagSplit	22.55	27.22	29.07	34.21
6	LemTag MorphMarkup-Split	newLemTag	21.85	26.90	29.33	33.96
7	LemTag Morph-noMarkup-Split	newLemTag	22.86	27.05	29.20	34.10
8	LemTag Morph-noMarkup-noSplit	newLemTag	22.82	27.18	29.18	34.12
9	LemTag MorphMarkup-Split	newLemTagSplit	22.25	27.12	29.39	34.38
10	LemTag Morph-noMarkup-Split	newLemTagSplit	22.53	26.90	29.10	34.12
11	LemTag Morph-noMarkup-noSplit	newLemTagSplit	23.13	27.55	29.42	34.19

Table 5: Experimental results in case-sensitive BLEU for 4 training settings.

6 Experiments

We compare four training settings: *small* (248,730 sentences: news-commentary), *large2M* (1,956,444 sentences: Europarl + news-commentary), *large4M* (4,116,215 sentences: Europarl + news-commentary + CommonCrawl) and *medium* (1M sentences) where the medium corpus consists of the news-commentary corpus and the first ~750k sentences of Europarl. We use WMT'15 as dev-set (2169 sentences) and WMT'16 as test-set (2999 sentences).

The lemma-tag approach doubles the sentence length by inserting tags. To avoid overly long sentences, the training data was first filtered to sentences of length 50, and after that, sentences more than 60 words long after BPE splitting were removed (e.g., sentences containing mostly foreign language words split nearly at character level).

Data pre-processing The baseline was trained on plain surface forms (tokenized and true-cased).

For the German lemma-tag system, we used Bit-Par (Schmid, 2004) to obtain morphological features in the sentence context, and SMOR (Schmid et al., 2004) for morphological analysis. For English, we used TreeTagger (Schmid, 1994). The English morphological analyzer for the small, medium and large2M systems was trained on the large2M data, the analyzer for the large4M system was trained on the full ~4M lines.

All systems (baseline and lemma-tag variants) underwent BPE segmentation ("joint" BPE of source/target side) with 30k merging operations.

encoder	transformer	initial-learning-rate	0.0002
decoder	transformer	label-smoothing	0.1
batch-type	word	transf.-dropout-act	0.1
batch-size	4096	transf.-dropout-attention	0.1
num-layers	6	transf.-dropout-prepost	0.1
max-seq-len	max sent len	checkpoint-frequency	3000

Table 6: Sockeye parameters for the Transformer model.

Training For the experiments, we used a *Transformer* NMT model (Sockeye toolkit: Hieber et al. (2017)). Table 6 shows the training parameters.

6.1 System variants

Table 5 shows different representation variants on the source and target side, as outlined in section 5. Generally, the lemma-tag systems are better than a standard NMT system; there is not much difference between the old (Tamchyna et al., 2017) and the new version (lines 2 and 3). Source-side lemma-tag pairs improve the small and medium settings when paired with non-split German data; split German data works better for the Large2M system. Both variants perform similarly for the Large4M system (lines 4 and 5). English word-internal markup improves the Large2M system, both with split and unsplit German data (lines 6 and 9), and leads to the best result when combined with split German data in the Large4M setting (line 9). The variants in lines 7 and 8 (split/unsplit morphological analysis) produce similar results when translating to non-split German data. Interestingly, with explicit splitting on the German side (lines 10 and 11), the non-split English data performs considerably better for the small/medium/large2M settings, leading to the best results overall for these data settings.

There seems to be a tendency that explicit splitting on both sides harms the smaller settings, possibly because translating at morpheme level requires more training data. Similarly, the English word-internal markup might introduce a complexity that only the larger systems can handle. On the other hand, using the non-split morphological analysis is less intrusive, but potentially useful at the BPE segmentation step by providing better access to sub-words. However, the best variants use explicit segmentation on the target side – this makes the question "to split or not to split" difficult to answer.

System	Small	Medium	Large2M	Large4M
1	17.17	22.12	24.32	30.85
2	18.86	22.85	25.13	31.67
3	19.16	23.13	24.46	31.52
4	19.30	22.75	24.42	31.70
5	18.45	23.05	24.68	32.56
6	18.19	22.92	24.93	32.23
7	19.43	23.06	25.22	31.74
8	18.79	22.55	24.68	32.15
9	18.79	22.67	24.68	31.95
10	18.06	22.44	24.98	32.25
11	18.75	23.09	24.46	31.99

Table 7: BLEU scores for a medical-domain test set.

Maybe *always* splitting at a certain level is not the right approach, but rather a more context-sensitive segmentation strategy would be desirable.

6.2 Application to out-of-domain data

In low-resource scenarios, such as translating data of a particular domain, the problems caused by inflectional variants and forms created through derivation are typically aggravated. Applying a system trained on general language, but with a component to handle inflection and word formation, to an out-of-domain test set constitutes an interesting use case. We use a test set[1] (Haddow et al., 2017) from the medical domain (1931 sentences), containing health information aimed at the general public and summaries of scientific studies. Table 7 shows the results for the different system variants. For all data settings, the lemma-tag variants are better than the surface form baselines. There are no clear tendencies for a best-performing strategy across all settings, but English morphological analysis seems to contribute less, whereas English lemma-tag information (lines 4, 5) leads to overall good results.

6.3 Examples

Table 8 shows two examples to illustrate the effect of morphological analysis. In (a), the baseline translates the noun *foolishness* as adjective, whereas the morphologically enriched system chooses a valid translation. Looking at the representation of *foolishness* after BPE segmentation, the baseline's *fool@@ is@@ hn@@ ess* is not particularly meaningful, whereas the representation of system 6 is [NN] *fool* <N> *ish*<SUFF_ADJ> *ness*<SUFF_N>, which provides a better basis for translation.

In (b), from the medical test set, the baseline fails to translate *coagulation*. Below, the BPE representations of *coagulation* (f=19) and *coagulate*

In	For all his foolishness Ed Miliband knew who his enemies were.
S1	Trotz seines törichten Ed Miliband wusste er, wer seine Feinde waren. *For his foolish Ed Miliband knew he, who his enemies were.*
S6	Trotz all seiner Dummheit wusste Ed Miliband, wer seine Feinde waren.
Ref	Bei all seiner Verrücktheit wusste Ed Miliband, wer seine Feinde waren.

(a)

In	Current testing for TIC normally involves coagulation tests on the patient's blood.
S1	Aktuelle Tests für TIC beinhalten normalerweise Coagulationstests am Blut des Patienten.
S10	Bei der aktuellen TIC-Prüfung handelt es sich in der Regel um Gerinnungstests am Blut des Patienten.
Ref	Aktuelle Erprobung von TIC beinhaltet normalerweise Gerinnungstests der Blut des Patienten.

(b)

Table 8: Translation examples from the Large2M setting (a) and the Large4M setting (b).

(f=3) in the respective systems are compared:

Surface (System 1)	Tag	morph. annotated (System 10)
co@@ ag@@ ulation	[NN]	co@@ ag@@ ulate ion<SUFF_N>
co@@ ag@@ ulate	[VB]	co@@ ag@@ ulate

Even with BPE segmentation, the representation in System 10 is more general than in the surface system, and in particular allows matching with e.g., *coagulate*. Similarly, *Gerinnungstest* (*coagulation test*) is represented as *ger@@ innen*<V> *ung*<NN><SUFF> *Test*<NN>, allowing to combine statistics of the verb *gerinnen* and the noun *Gerinnung*. Thus, better generalization, paired with tag information, enables the morphology-informed systems to make better use of the training data.

7 Conclusion

We showed that morphologically sound segmentation that considers non-concatenative processes in order to obtain a consistent representation of subwords improves translation. The findings of our experiments provide important insights for translating morphologically rich languages, and are particularly important for low-resource settings.

References

Duygu Ataman and Marcello Federico. 2018. Compositional Representation of Morphologically-Rich Input for Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Paper)*, pages 305–311, Melbourne, Australia.

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 2017*, pages 331–342, Prague, Czech Republic.

Tamali Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 55–60, New Orleans. Association for Computational Linguistics.

Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph Convolutional Encoders for Syntax-Aware Neural Machine Translation. In *Proceedings of EMNLP 2017*, pages 1957–1967, Copenhagen, Denmark.

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*, pages 579–587, Gothenburg, Sweden. Association for Computational Linguistics.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-Sequence Attentional Neural Machine Translation. In *Proceedings of ACL 2016*, pages 823–833, Berlin, Germany.

Barry Haddow, Alexandra Birch, Ondřej Bojar, Fabienne Braune, Colin Davenport, Alex Fraser, Matthias Huck, Michal Kašpar, Květoslava Kovaříková, Josef Plch, Anita Ramm, Juliane Ried, James Sheary, Aleš Tamchyna, Dušan Variš, Marion Weller, and Phil Williams. 2017. HimL: Health in my Language. In *Proceedings of the EAMT 2017 User Studies and Project/Product Descriptions*, page 33, Prague, Czech Republic.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side Word Segmentation Strategies for Neural Machine Translation. In *Proceedings of the Second ACL Conference on Machine Translation (WMT)*, pages 56–67, Copenhagen, Denmark.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 187–194, Budapest, Hungary. Association for Computational Linguistics.

Maria Nădejde, 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.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

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

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)*, pages 1263–1266, Lisbon, Portugal.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*.

Aleš Tamchyna, Marion Weller-Di Marco, and Alexander Fraser. 2017. Modeling Target-Side Inflection in Neural Machine Translation. In *Proceedings of the Second ACL Conference on Machine Translation (WMT)*, pages 32–42, Copenhagen, Denmark.

Marion Weller-Di Marco. 2017. Simple Compound Splitting For German. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE17)*, Valencia, Spain.