Combining Bilingual Terminology Mining and Morphological Modeling for Domain Adaptation in SMT

Marion Weller[1,2] **Alexander Fraser**[2] **Ulrich Heid**[3]

[1]Institut für Maschinelle
Sprachverarbeitung
Universität Stuttgart
weller@ims.uni-stuttgart.de

[2]Centrum für Informations-
und Sprachverarbeitung
LMU München
fraser@cis.uni-muenchen.de

[3]Institut f. Informationswissen-
schaft u. Sprachtechnologie
Universität Hildesheim
heid@uni-hildesheim.de

Abstract

Translating in technical domains is a well-known problem in SMT, as the lack of parallel documents causes significant problems of sparsity. We discuss and compare different strategies for enriching SMT systems built on general domain data with bilingual terminology mined from comparable corpora. In particular, we focus on the target-language inflection of the terminology data and present a pipeline that can generate previously unseen inflected forms.

1 Introduction

Adapting statistical machine translation (SMT) systems to a new domain is difficult when the domain lacks sufficient amounts of parallel data, as is the case in many technical or medical domains. SMT systems trained on general language (e.g. government proceedings) face data-sparsity issues when translating texts from such domains, particularly if translating into a morphologically rich language.

In this paper, we compare different strategies to adapt an EN-FR SMT system built on Europarl to a technical domain (wind energy) by making use of term-translation pairs mined from comparable domain-specific corpora. In a first series of experiments, we study two methods of integrating bilingual terminology into a phrase-based SMT system: adding term translation pairs via XML mark-up and as pseudo-parallel training data. In particular, we compare the effects of integrating translation candidates for multi-word terms vs. single-word terms and show that the use of single-word terms can be harmful. Using bilingual terminology in the form of pseudo-parallel data significantly outperforms the the baseline.

However, it also becomes evident that terminology handling requires morphological modeling: when the integrated term-translation pairs are restricted to the inflected forms seen in the (domain-specific) data, this ignores the fact that other forms might be needed when translating. Furthermore, translation-relevant morphological features (e.g. number) must be maintained during the translation process. As a way to address these problems, we present a morphology-aware translation system that treats inflection as a target-side generation problem. Combining the integration of term-translation pairs and the modeling of target-side morphology allows for the generation of unseen word forms and the preservation of translation-relevant features. In the second part of the paper, we describe and discuss a novel pipeline for morphology-aware integration of bilingual terminology. While this system's improvement over the baseline is not statistically significant, our analysis highlights the need for explicit morphological modeling, which, as far as we know, has not been addressed previously.

Issues in translating out-of-domain data. When translating texts of domains that are not well represented by the training data, there are two main problems: (i) data sparsity: many domain-specific words do not appear in the parallel data and thus cannot be translated (e.g. the English term *torque* which does not occur in Europarl), and (ii) polysemy: words can have different meanings when used in general vs. specialized language. For example, the word *boss* means either *manager* or refers to a rivet-type object. In a general language text, the meaning of *manager* is predominant, whereas in a text of a technical domain, that sense is less likely to be correct. Because a translation model trained on general language data learns that *boss* → *manager* is a good translation, this translation is likely to be used when translating data from a technical domain. In order to make previously unknown terms available and to model domain-specific preferences, we enrich the SMT system with domain-specific term-translation pairs that are not contained in the general language parallel data.

Modeling morphology. Another type of data sparsity occurs in translations to languages with rich (noun) inflection, as the parallel training data is unlikely to cover the full inflection paradigms of all words. As a result, some inflected forms are unavailable to the SMT system. This problem increases considerably when translating terms which are not well represented in the parallel training data, as is the case in the domain-adaptation scenario presented in this work. Modeling target-side morphology helps to reduce this kind of data-sparsity: we present a two-step approach, in which we separate the translation process from target-side inflection by first translating into a lemmatized representation, with a post-processing component for generating inflected forms. This simplifies the translation task, as information concerning only the target language has been removed. Also, this two-step approach allows us to generate forms which are not contained in the parallel data, which is of particular interest for domain-adaptation scenarios, where the full inflectional paradigm of term-translation pairs might not even be covered by the domain-specific data used for term mining. Furthermore, this setup allows us to specifically indicate how a term in a given context should be translated. For example, it provides the means to guarantee that a source-language term in plural is translated by the corresponding target-language term in plural, regardless of whether the required inflected form occurs in the training data. Although there are exceptions such as $\textit{furniture}_{SG}$ → $\textit{meubles}_{PL}$, we believe they play a negligible role when translating under-resourced domains.

2 Related work

There has been considerable interest in mining translations directly from comparable corpora. A few representative examples are (Daille and Morin, 2005; Haghighi et al., 2008; Daumé III and Jagarlamudi, 2011; Prochasson and Fung, 2011), all of which mine terms using distributional similarity. These approaches tend to favor recall over precision. In contrast, we use a high-precision method consisting in recognizing term candidates by means of part-of-speech patterns with an alignment method relying on dictionary entries (Weller and Heid, 2012).

A second strand of relevant work is the integration of terms into SMT decoding. Hálek et al. (2011) integrated named entity translations mined from Wikipedia using the XML mode of Moses, which creates new phrase table entries dynamically. Pinnis and Skadins (2012) also studied mining named entities, as well as using a high quality terminological database, and added these resources to the parallel training data. We compare these two options (XML vs. added parallel data) and show that adding the terms to the parallel training data leads to better results.

To deal with the issue of obtaining the proper inflection of mined terms, we implemented a morphology-aware English to French translation system that separates the translation task into two steps (translation + inflection generation), following Toutanova et al. (2008) and Fraser et al. (2012).

Formiga et al. (2012) use a component for target-side morphological generation to translate news and web-log data. In contrast to our work, they do not deal with nominal morphology, but model verb inflection: this is important for web-log data, as second-person verb forms rarely appear in Europarl-type training data. Wu et al. (2008) use dictionary entries for adapting a system trained on Europarl to news, but without applying morphological modelling to their EN-FR system. Furthermore, news and also web-log data are considerably more similar to Europarl than technical data.

Our main contribution is that we show how to combine three areas of research: bilingual term mining, using terms in SMT, and generation of inflection for SMT. We describe a novel end-to-end morphology-aware solution for using bilingual term mining in SMT decoding.

3 Bilingual terminology mining

In contrast to parallel corpora, which are difficult to obtain in larger quantities, comparable corpora of a particular domain are relatively easy to obtain. Comparable corpora are expected to have similar content and consequently similar domain-specific terms in both languages and thus constitute a suitable basis for the mining of term-translation pairs.

For both source and target-language, term candidates are extracted based on part-of-speech patterns, focusing on nominal phrases. The resulting sets of term candidates are then aligned.

We use all available domain-specific training data (cf. section 7) for monolingual term extraction on the target language. Source language terms are only extracted for the input data to the SMT system (tuning/test set) because our methods for term integration are restricted to terms contained in the sentences to be translated.

Term alignment. The task of term alignment consists in finding the equivalent of a source language term in a set of target language terms. One method is *pattern-based compositional term-alignment*: all components of a multi-word term are first translated individually using a (general language) dictionary, and then recombined according to handcrafted translation patterns such as
(EN) `noun1 noun2` ↔ (FR) `noun2` *de* `noun1`[1].
As the recombination of individual translations leads to over-generation, the generated translation candidates are filtered against the list of extracted target-language terms. A principal assumption is that the term pairs are semantically transparent and of a similar morpho-syntactic structure. The example for the term *glass fibre* illustrates the process:

(1) individual translation:
`noun1`: *glass* → *verre* (*glass*),
loupe (*magnifying glass*)
`noun2`: *fibre* → *fibre*

(2) recombination[2] of translations:
fibre de verre, *fibre de loupe*

(3) filtering against target terms:
fibre de verre, ~~*fibre de loupe*~~

Source and target terms are not necessarily of the same word class; such shifts are dealt with by simple morphological rules, as shown by the term

energy$_{\mathtt{N}}$ *yield assessment*
→ *estimation du rendement* **énergétique**$_{\mathtt{ADJ}}$
(*assessment of* **energetical**$_{\mathtt{ADJ}}$ *yield*)

Adding the entry *energy* → *énergétique*$_{\mathtt{ADJ}}$ to the dictionary allows to cover morphological variation between source and target terms.

For the alignment, terms are lemmatized and need to be mapped to the respective inflected forms before being integrated into the SMT system. The translation probabilities are computed based on the relative frequencies of the inflected forms of all translation possibilities in the domain-specific data:

EN	FR	freq	prob
hub height	hauteur du moyeu	14	87.5
	hauteur de moyeu	2	12.5

		MWT	SWT
tuning set	total	440	1014
test set	total	442	1015
tuning set	not in phrase-table	156	18
test set	not in phrase-table	192	15

Table 1: Number of terms (types) for which one or more translation candidates were found.

Table 1 gives an overview of the number of obtained translation pairs for terms extracted from the test/tuning data (cf. section 7); we differentiate between single-word terms (SWTs) and multi-word terms (MWTs). This is motivated by the fact that MWTs provide more context in step (3) and are therefore more likely to be correctly aligned. In the case of SWTs, every translation listed in the dictionary can be output as a valid alignment provided it occurred in the corpus, regardless of context. Table 1 also shows the amount of term-translation pairs not covered by the phrase-table: in the case of MWTs, a reasonable amount of term-translation pairs are new to the system, whereas the number of new SWTs is very low in comparison to the number of found SWT term-translation pairs.

The pattern-based compositional term alignment tends to favor precision over recall. This general outcome is observed in earlier work (Weller and Heid, 2012)[3]; we assume that the findings for DE-EN largely also apply to our EN-FR alignment scenario. Moreover, it is not guaranteed that the translation of a source term occurs in the target-language data when working with comparable corpora. Another problem are structural mismatches of the source term and its target-language equivalent. While the translation occurs in the target language term list, it is of a different morpho-syntactic structure in a way that is not captured by the patterns and morphological rules. Finally, lack of dictionary coverage is also responsible for not finding target-language equivalents. We focus on integrating moderate amounts of good-quality term pairs, motivated by our method for integrating term pairs: our results indicate that the SMT-system is sensitive to incorrect translations, particularly for SWTs.

[1] *de*: French preposition meaning *of*.

[2] Working with translation patterns, non-content words such as prepositions can be easily inserted in this step.

[3] We use alignment patterns adapted from this work.

SMT-output + stem-markup	**pred.** **feat.**	**gen.** **forms**	**post-** **proc.**	**gloss**
`le<+DET>[ART]`	M.Sg	le	**l’**	*the*
`excès<`**M.Sg**`>[N]`	M.Sg	excès	excès	*excess*
`de[P]`	–	de	**d’**	*of*
`énergie<`**F.Sg**`>[N]`	F.Sg	énergie	énergie	*energy*
`peut[VFIN-peut]`	–	peut	peut	*can*
`être[VINF]`	–	être	être	*be*
`vendre[VPP]`	M.Sg	vendu	vendu	*sold*
`à[P]`	–	à	**au**	*to*
`le<+DET>[ART]`	M.Sg	le		*the*
`réseau<`**M.Sg**`>[N]`	M.Sg	réseau	réseau	*grid*

Table 2: Processing steps for the EN input sentence *[...] excess energy can be sold back to the grid.*

4 Inflection prediction system

To build the morphology-aware system, the target-side data (parallel and language model data) is transformed into a stemmed format, based on the annotation of a morphological tagger (Schmid and Laws, 2008). This representation contains translation-relevant feature markup: nouns are marked with *gender* (considered part of the stem) and *number*. Assuming that source-side nouns are translated by nouns with the same *number* value, this feature is indirectly determined by the source-side input. The number markup is thus needed to ensure that the source-side number information is transferred to the target side. For a better generalization, we split portmanteau prepositions into article and preposition (*au* → *à+le*: *to+the*).

For predicting the morphological features of the SMT output (*number* and *gender*), we use a linear chain CRF (Lavergne et al., 2010). In the prediction step, the values specified in the stem-markup (*number* and *gender* on nouns) are propagated over the rest of the phrase, as illustrated in column 2 of table 2. Based on the stems and the morphological features, inflected forms can be generated using a morphological tool for analyzing and generating inflected forms (cf. section 7), as illustrated in column 3. In order to generate correct French surface forms, a post-processing step is required, including the re-insertion of apostrophes and portmanteau merging (*à+le* → *au*), cf. column 4.

5 Integration of term-translation pairs

In this section, we compare two methods to integrate bilingual terminology, using a standard SMT-system (to be referred to as the “inflected” system): using XML-markup and in the form of pseudo-parallel data. In section 6, we discuss the integration of terms into the “morphology-aware” system.

Using XML input to add translation options. One way to integrate term-translation pairs into an SMT system is to list translation options with their translation probabilities for a word or word sequence in the input sentence by means of XML-markup. This approach has been applied by Hálek et al. (2011) (cf. section 2) to translations of named entities mined from Wikipedia in an English-Czech SMT system. In contrast, we integrate translation pairs of nominal phrases: this requires modelling features that are dependent on the source-side (e.g. number) which is not to the same extent necessary for names. Named entities are in many cases easier to deal with than terminology, as they are usually the same on the source side, even though their inflection can vary, e.g. in the form of *case*-markers, which depend on the target-language. This means that source-side information plays a negligible role, whereas for nominal phrases, number information (as contained in the stem markup) is important for the generation of inflected forms.

For the integration of term translation pairs, potential source terms are identified in the input sentence using the same pattern-based approach as for monolingual term identification (cf. section 3). Longer terms are preferred in the case of several annotation possibilities in order to provide the system with long translations, but also to avoid that phrasal units are interrupted: *[wind*$_N$ *energy*$_N$*] site*$_N$ vs. *[wind*$_N$ *energy*$_N$ *site*$_N$*]*.

We compare the effects of integrating multi-word and single-word terms vs. only multi-word terms. As a variant, only term-translation pairs of which the source-side term does not occur in the phrase table are integrated: assuming that the translation model already has more reliable statistics for terms in the phrase-table, only term-translation pairs that are not covered by the parallel data are used. Particularly for SWTs, this drastically reduces the amount of term-translation pairs. When restricting the integration to “new” terms, however, the problem of polysemy (e.g. *boss* → *manager* or *rivet-type object*) is not resolved. In such cases, it is even likely that the wrong sense, i.e. the general language meaning, is output by the translation system. Nevertheless, this variant leads to the best results.

As term alignment is based on lemmas, a mapping between surface forms and lemmas is needed: first, inflected EN surface forms are projected to their lemmas, which are then aligned to FR lemmas. Then, the aligned target-side lemmas are mapped

Input	`clean the <term translation=′′fer au rotor||pale de rotor||pales de rotor ||pale du rotor||pales du rotor′′ prob=′′0.0385||0.0385||0.2692||0.1153|| 0.5384′′> rotor blades </term> with a mild soap and water .`
Baseline	`nettoyage du rotor des lames de savon avec une légère et de l′ eau .` *cleaning of the rotor of the blades (of a knife) of soap with a mild and water.*
With terms	`nettoyer les pales du rotor avec un savon modérée et de l′ eau .` *cleaning the blades of the rotor with a moderate soap and water.*
Reference	*Nettoyez les pales du rotor au savon doux et à l'eau.*

Table 3: Adding translation options for the term *rotor blades* to the input sentence.

to the respective inflected forms observed in the domain-specific corpus. As a result, some of the inflected forms can be incorrect in terms of *number* by mapping the lemma to both singular and plural forms, regardless of the input term. Filtering for number in this step is useful only to a limited extent, as it will prevent a translation entirely if the inflected forms of the required *number* value do not occur in the domain-specific data. While a good translation in the wrong number is clearly better than no translation, it is still desirable to have the possibility to model *number*: we consider this a strong motivation for a morphology-aware integration of terminology.

Another crucial point is the language model data which needs to contain the target-language terms offered to the translation model. As all target language terms are extracted from a domain-specific corpus, this data is used in the language model.

The example in table 3 illustrates how the system benefits from the translations for the term *rotor blades* in the input sentence: while FR *pale* (blades on a wind mill) occurs once in the parallel data, there is no alignment to EN *blade*. As a result, *blades* is translated as *lames* (blades on a knife). Providing the translation options leads to the correct translation of *blades* → *pales* in the context of the term *rotor blades*. In addition, the system with terminology information produces a well-formed French sentence in contrast to the meaningless output of the baseline system, because the correct translation allows for matching a plausible word sequence with the language model.

Adding terms to parallel data. In our experiments, adding translation options via XML markup did not work as well as hoped for; this is in line with the findings of Hálek et al. (2011): adding translation pairs directly into the SMT system can be too intrusive, causing more harm than benefit. We tested a different approach: the term-translation pairs are added as a pseudo parallel corpus to the parallel training data. Adding each term-translation pair once is not likely to help if the word is ambiguous and already occurs in the parallel data with its general language translation. Instead, term translation pairs are added according to their frequency in the target-side corpus. As before, all observed inflected forms are listed as possible translations.

6 Morphology-aware integration of term-translation pairs

The setup described in the previous sections has two shortcomings: the data might not provide the full inflection paradigm of the terms, and it is not possible to model features such as *number*: integrating stemmed terms to the inflection prediction system allows us to handle these two problems as the number information of a source-term can simply be transferred as number markup to the stemmed translation candidate and specific forms not occurring in the data used for term mining can be generated using a morphological resource.

For the terminology integration into a morphology-aware translation system, we opted for the variant of adding pseudo parallel data to the training data of the SMT system as this led to the best results in the previous experiments. First, the aligned terms are transferred to the stemmed representation. For the number markup, the source-side is tagged and the *number* values are transferred to the corresponding stems based on the alignment patterns (cf. section 3). In this step, the number markup in the generated target-side text is determined by transfer from the source-side. In comparison, the number markup in the "original" parallel data (Europarl) is given by the target-side, i.e. the parse-annotation.

Generating target phrases depending on the requirements of the source-side, i.e. creating unseen forms, can lead to stem+markup combinations that do not occur in the data used to build the language model. Words not contained in the language model score very badly during decoding and are thus effectively not available to the SMT system. In order to make all stems accessible, the generated pseudo parallel data is added to the language model data.

An alternative way to avoid the generation of forms not represented in the language model consists in foregoing number markup. Instead of keeping it through the translation in form of stem markup, number information can be reinstated in the feature prediction step using source-side features. However, this creates two new problems: first, the representation without number markup loses discriminatory power[4]. For example, there is no way to guarantee subject-verb agreement without number information on nouns. The second problem is that parallel domain-specific data is needed to train the models for feature prediction. While we believe that removing number markup in the translation step is a sounder way to deal with target-side morphology in this application, we leave this extension of our model to future work due to the practical problems that arise with this.

7 Data and resources

Our experiments are carried out on an EN-FR standard phrase-based Moses[5] system which is adapted to the domain of wind energy. As a basis for terminology mining, we compiled a target-language corpus for that domain. This included documents obtained by automatic crawling (de Groc, 2011), and manually obtained data from various web-sites. In total, the corpus consists of 161.367 sentences (4.136.751 words). For the tuning/test data, we manually collected and sentence-aligned parallel texts from various internet resources, including manuals for setting up/maintaining wind energy towers, multi-lingual scientific journal articles and data about regulations and administrative aspects. The resulting 1290 parallel sentences were evenly divided into test/tuning sets.

The parallel training data for the EN-FR SMT system consists of 2.159.501 sentences (Europarl and News data from the 2013 WMT shared task). For the language model, we used a combination of the FR part of the parallel data and the wind energy corpus. As the domain-specific corpus is considerably smaller, we built individual language models for each corpus and interpolated them using weights optimized on the tuning data following the approach of Schwenk and Koehn (2008).

For the feature prediction, we used the Wapiti toolkit (Lavergne et al., 2010) to train CRFs on combinations of the wind corpus and the FR part of the parallel data. The CRF has access to the basic features *stem* and *POS tag* as well as *gender* and *number* within a window of 5 positions to each side of the current word.

The morphological analysis of the French training data is obtained using RFTagger, which is designed for annotating fine-grained morphological tags (Schmid and Laws, 2008). For generating inflected forms based on stems and morphological features, we use an extended version of the finite-state morphology FRMOR (Zhou, 2007). FRMOR is a morphology tool similar to SMOR (Schmid et al., 2004), which allows to analyze and generate inflected word forms. The term alignment requires a general language dictionary[6] from which we use the 36,963 1-to-1 entries.

8 Experiments and results

We present results for the integration of bilingual terminology into an inflected system and a morphology-aware translation system.

Integrating terminology into the inflected system. An easy way to adapt an SMT system to a new domain consists in adding language model data of that domain. This does not help with the problem of out-of-vocabulary words, but it can enhance translations with low probabilities and provide plausible contexts for the generated sentences. The systems in row 1 in table 4 show that adding domain-specific data leads to a considerable increase in BLEU; all further systems in table 4 use this enlarged language model and are compared to baseline *b*.

Moses' XML mode offers two possibilities: forcing the SMT system to use the given translations (*exclusive*) or allowing for an optional usage (*inclusive*). As preliminary experiments, as well as the findings of Hálek et al. (2011), showed that the inclusive setting leads to better results, we only report BLEU scores for this variant[7]. We compare two versions: providing only the translations of multi-word terms (MWTs) and providing the translations

[4] See also experiments on re-inflecting surface forms ("Method 1") in Toutanova et al. (2008).

[5] `http://www.statmt.org/moses`

[6] from `www.dict.cc` and `www.freelang.net`

[7] Particularly for SWTs, forcing the system to use the provided translations using the exclusive setting can very much hurt performance as it goes against Moses' tendency to use long translation units.

	system	BLEU
1	Baseline a: general LM	18.93
	Baseline b: +domain-spec. LM	21.59
2	XML-markup (MWT + SWT)	20.56
	XML-markup (MWT)	20.71
3	XML-markup-filt. (MWT + SWT)	21.68
	XML-markup-filt. (MWT)	21.57
4	Added parallel (MWT + SWT)	21.68
	Added parallel (MWT)	21.87
	Added parallel (MWT + filt. SWT)	22.03*
	Added parallel filt. (MWT + SWT)	21.96*

Table 4: Results for integration of terminology into an inflected EN-FR translation system. (*: significantly better than baseline b at a 0.05 level)

of both multi-word and single-word terms (SWTs). This is motivated by the assumption that adding translations of single words is likely to be more harmful as it is to some extent incompatible with Moses' tendency to prefer longer phrases.

The translation probabilities of term-translation pairs given in the XML markup usually are considerably higher than the ones in the phrase-table and might thus have an undue advantage, particularly when assuming that the statistics in the phrase-table are more reliable for terms that are not restricted to the domain. Furthermore, the generated translations of multi-word terms are more likely to be correct as they provide more context in the alignment step. While the system with only MWTs is slightly better, both variants are worse than the baseline (row 2 in table 4). Restricting the added term-translation pairs to those where the source-phrase does not occur in the phrase-table helps, but does not outperform the baseline (row 3 in table 4). Here, using both MWTs and SWT leads to a slightly better score, presumably because the added SWTs are unknown to the system and even a translation by a one-word phrase is beneficial.

Integrating bilingual terminology in the form of pseudo-parallel data leads to the best results (row 4 in table 4). Again, restricting the data to MWTs is slightly better than using all term-translation pairs. The score for the MWT-only system (21.87) is on the verge of being statistically significantly better than baseline *b*. Adding single-word translations which do not occur in the phrase-table leads to a statistically significant improvement (22.03), as does filtering both SWTs and MWTs (21.96).

	system	CRF trained on	BLEU
1	Baseline	wind+news	21.47
		wind+europarl	21.54
2	MWT^a	wind+europarl	21.77
3	MWT + SWT^c	wind+europarl	21.11
4	MWT + filt. SWT^b	wind+europarl	21.74
5	filt. (MWT + SWT)^b	wind+europarl	21.48

Table 5: Adding pseudo parallel data to the training data for a morphology-aware system. *a*: LM from baseline system; *b*: MWT translations added to LM data; *c*: MWT+SWT translations added to LM data.

Integrating terminology into the morphology-aware system. The score of the morphology-aware system (21.54) is comparable to that of the inflected system (21.59), as shown in table 5. The importance of in-domain training data for the CRF is illustrated by the results obtained when training the CRF on wind+news (318.112 sentences) and on wind+europarl (2.161.367 sentences): even though the second training set is considerably larger, there is basically no gain in BLEU. Considering this outcome, we assume that more in-domain training data for the CRF would lead to better overall results.

In order to make better use of the in-domain training data, singletons were replaced by their part-of-speech tag[8]. However, the stem feature considerably contributes to the prediction result: this is illustrated by the results in table 5, where a CRF trained on a combination of Europarl and wind energy data is only marginally better in terms of BLEU than a system trained on a much smaller amount of general language data and data of the wind energy domain.

It is important to keep in mind that the CRF is trained on fluent data whereas the SMT output is heavily disfluent. As a result, there is a mismatch between ill-formed translation output and the well-formed data used to train the CRF; the gap between training data and the text for which features are to be predicted gets larger with increasing difficulty of the translation task, as is the case here.

Effects caused by sparse data do also affect the language model data: forms which are not contained in the parallel data cannot be produced by the translation system. In order to deal with out-of-vocabulary words, stem markup+tags are stripped of all those words in the language model data that do not occur in the parallel data. This enables the SMT system to score unknown words (e.g. names) in the language model, but also leads to side-effects due to sparsity: for example, the French term *rotors* occurs once in the parallel data and is correctly stemmed as `rotor<Masc.Pl>[N]` , while all occurrences of *rotor* in the singular form

[8] Experiments with replacing out-of-vocabulary words by a special tag were also not effective in terms of BLEU.

are stripped of the markup and treated as a name and thus do not undergo the inflection process.

As the method of adding term translation pairs to the parallel data led to the best results for the inflected system, we opted for this method for the integration of terms into the morphology-aware system. While the MWT-only system (2 in table 5) gets a better score than the baseline (1 in table 5) (21.77 vs. 21.54 using the larger CRF), the difference is not statistically significant. In contrast to the results for the inflected system, adding the set of SWTs filtered against the phrase-table slightly decreases BLEU, whereas adding all SWTs leads to a considerable decrease in BLEU. We assume that this outcome is partially caused by a problem with the language model: while all generated target terms are added to the language model data, they are not embedded in the context of a sentence, or, if also adding SWTs (system 3 in table 5), not even in the context of a term.

9 Conclusion

We presented different approaches to integrate bilingual terminology of a technical domain into an SMT system. First, we compared two integrating methods (providing translation options vs. term-translation pairs as pseudo-parallel data) and studied the effects of using only multi-word terms in comparison to both single-word and multi-word terms. Then, we applied the best term integration strategy to a morphology-aware translation system.

With the inflected system, we obtained a significant improvement over the baseline when adding terms as pseudo-parallel data. Our evaluation also clearly showed that Moses' XML mode has considerable problems in dealing with single-word terms. Furthermore, we highlighted the need for explicit modeling of morphological features for the integration of bilingual terminology.

While the morphology-aware system enriched with term pairs was not able to outperform the baseline on a statistically significant level, it outlines a pipeline that tackles two central problems of adapting translation systems to under-resourced domains: (i) preservation of translation-relevant features and (ii) generation of previously unseen inflected forms.

10 Acknowledgements

This work was funded by the DFG Research Projects "Distributional Approaches to Semantic Relatedness" and "Models of Morpho-Syntax for Statistical Machine Translation".
The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n. 248005.

References

Daille, B. and E. Morin. 2005. French-English terminology extraction from comparable corpora. In *Proceedings of IJCNLP 2005*.

Daumé III, H. and J. Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of ACL 2011*.

de Groc, C. 2011. Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *International Conferences on Web Intelligence and Intelligent Agent Technology*.

Formiga, L., A. Hernández, J. Mariño, and E. Monte. 2012. Improving English to Spanish out-of-domain translations by morphology generalization and generation. In *Proceedings of AMTA 2012*.

Fraser, A., M. Weller, A. Cahill, and F. Cap. 2012. Modeling Inflection and Word-Formation in SMT. In *Proceedings of EACL 2012*.

Haghighi, A., P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL 2008*.

Hálek, O., R. Rosa, A. Tamchyna, and O. Bojar. 2011. Named entities from wikipedia for machine translation. In *Proceedings of the Conference on Theory and Practice of Information Technologies*.

Lavergne, T., O. Cappé, and F. Yvon. 2010. Practical very large scale CRFs. In *Proceedings of ACL 2010*.

Pinnis, M. and R. Skadins. 2012. MT adaptation for under-resourced domains - what works and what not. In *Proceedings of HLT - the baltic Perspective*.

Prochasson, E. and P. Fung. 2011. Rare word translation extraction from aligned comparable documents. In *Proceedings of ACL 2011*.

Schmid, H. and F. Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of COLING 2008*.

Schmid, H., A. Fitschen, and U. Heid. 2004. SMOR: a German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of LREC 2004*.

Schwenk, H. and P. Koehn. 2008. Large and diverse language models for statistical machine translation. In *Proceedings of IJCNLP 2008*.

Toutanova, K., H. Suzuki, and A. Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *Proceedings of ACL-HLT 2008*.

Weller, M. and U. Heid. 2012. Analyzing and aligning german compound nouns. In *Proceedings of LREC 2012*.

Wu, Hua, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of COLING 2008*.

Zhou, Z. 2007. Entwicklung einer französischen Finite-State-Morphologie. University of Stuttgart.