Modeling Complement Types in Phrase-Based SMT

Marion Weller-Di Marco[1,2], **Alexander Fraser**[2], **Sabine Schulte im Walde**[1]

[1]Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
[2]Centrum für Informations- und Sprachverarbeitung,
Ludwig-Maximilians-Universität München
{dimarco|schulte}@ims.uni-stuttgart.de fraser@cis.lmu.de

Abstract

We explore two approaches to model complement types (NPs and PPs) in an English-to-German SMT system: A simple abstract representation inserts pseudo-prepositions that mark the beginning of noun phrases, to improve the symmetry of source and target complement types, and to provide a flat structural information on phrase boundaries. An extension of this representation generates context-aware synthetic phrase-table entries conditioned on the source side, to model complement types in terms of grammatical case and preposition choice. Both the simple preposition-informed system and the context-aware system significantly improve over the baseline; and the context-aware system is slightly better than the system without context information.

1 Introduction

SMT output is often incomprehensible because it confuses complement types (noun phrases/NPs vs. prepositional phrases/PPs) by generating a wrong grammatical case, by choosing an incorrect preposition, or by arranging the complements in a meaningless way. However, the choice of complement types in a translation represents important information at the syntax-semantics interface: The case of an NP determines its syntactic function and its semantic role; similarly, the choice of preposition in a PP sets the semantic role of the prepositional phrase.

While the lexical content of a target-language phrase is defined by the source sentence, the exact choice of preposition and case strongly depends on the target context, and most specifically on the target verb. For example, the English verb phrase *to call for sth.* can be translated into German by *etw. erfordern* (subcategorizing a direct-object NP but no preposition) or by *(nach) etw. verlangen* (subcategorizing either a direct-object NP or a PP headed by the preposition *nach*). Differences in grammatical case and syntactic functions between source and target side include phenomena like subject-object shifting: *[I]*$_{\text{SUBJ}}$ *like [the book]*$_{\text{OBJ}}$ vs. *[das Buch]*$_{\text{SUBJ}}$ *gefällt [mir]*$_{\text{OBJ}}$. Here, the English object corresponds to a German subject, whereas the English subject corresponds to the indirect object in the German sentence.

Selecting the wrong complement type or an incorrect preposition obviously has a major effect on the fluency of SMT output, and also has a strong impact on the perception of semantic roles. Consider the sentence *John looks for his book*. When the preposition *for* is translated literally by the preposition *für*, the meaning of the translated sentence *John sucht für sein Buch* shifts, such that *the book* is no longer the object that is searched, but rather a recipient of the search. To preserve the source meaning, the prepositional phrase headed by *for* must be translated as a direct object of the verb *suchen*, or as a PP headed by the preposition *nach*.

Since prepositions tend to be highly ambiguous, the choice of a preposition depends on various factors. Often, there is a predominant translation, such as *for* $\rightarrow$ *für*, which is appropriate in many contexts, but unsuitable in other contexts. Such translation options are often difficult to override, even when there are clues that the translation is wrong. Furthermore, even though prepositions are highly frequent words, there can be coverage problems if a preposition is not aligned with the specific preposition required by the context, due to structural mismatches.

This paper presents two novel approaches to improve the modeling of complement types. A simple approach introduces an abstract representation of "placeholder prepositions" at the beginning of

Proceedings of the First Conference on Machine Translation, Volume 1: Research Papers, pages 43–53,
Berlin, Germany, August 11-12, 2016.

noun phrases on the source and target sides. The insertion of these placeholder prepositions leads to a more symmetric structure and consequently to a better coverage of prepositions, as all NPs are effectively transformed into PPs, and prepositions in one language without a direct equivalent in the other language can be aligned. Furthermore, the placeholder prepositions function as explicit phrase boundaries and are annotated with grammatical case, so they provide flat structural information about the syntactic function of the phrase. The placeholder representation leads to a significant improvement over a baseline system without prepositional placeholders.

Our second approach enhances the abstract placeholder representation, and integrates source-side context into the phrase table of the SMT system to model different complement types. This is done by generating synthetic phrase-table entries containing contextually predicted prepositions. With this process, we aim to (i) improve the preposition choice conditioned on the source sentence, and to (ii) manipulate the scores in the generated entries to favour context-appropriate translations. Generating phrase-table entries allows to create prepositions in contexts not observed in the parallel training data. The resulting phrase-table entries are unique for each context and provide the best selection of translation options in terms of complement realization on token-level. This variant significantly outperforms the baseline, and is slightly better than the system with inserted placeholder prepositions.

2 Related Work

Our work is related to three research areas: using source-side information, previous approaches to model case and prepositions and the synthesis of phrase-table entries.

Source-side information has been applied to SMT before, often for the purpose of word sense disambiguation and improving lexical choice (Carpuat and Wu, 2007; Gimpel and Smith, 2008; Jeong et al., 2010; Tamchyna et al., 2014), but without a focus on synthesis or syntactic-semantic aspects such as subcategorization.

Prepositions are difficult to translate and responsible for many errors, as has been shown in many evaluations of machine translation. For example, Williams et al. (2015) presented a detailed error analysis of their shared task submissions, listing the number of missing/wrong content and function words. For the language pair English–German, the combined number of *missing/wrong/added prepositions* is one of the most observed error types. Agirre et al. (2009) were among the first to use rich linguistic information to model prepositions and grammatical case in Basque within a rule-based system, leading to an improved translation quality for prepositions. Their work is extended by Shilon et al. (2012) with a statistical component for ranking translations. Weller et al. (2013) use a combination of source-side and target-side features to predict grammatical case on the SMT output, but without taking into account different complement types (NP vs. PP). Weller et al. (2015) predict prepositions as a post-processing step to a translation system in which prepositions are reduced to placeholders. They find, however, that the reduced representation leads to a general loss in translation quality. Experiments with annotating abstract information to the placeholders indicated that grammatical case plays an important role during translation. We build on their observations, but in contrast with generating prepositions in a post-processing step, prepositions in our work are accessible to the system during decoding, and the phrase-table entries are optimized with regard to the source-sentence. Finnish is a highly inflective language with a very complex case and preposition system. Tiedemann et al. (2015) experimented with pseudo-tokens added to Finnish data to account for the fact that Finnish morphological markers (case) often correspond to a separate English word (typically a preposition). Due to the complexity of Finnish, only a subset of markers is considered. The pseudo-tokens are applied to a Finnish–English translation system, but a manual evaluation remains inconclusive about the effectiveness of their method. For the preposition-informed representation in our work, we adapt both source and target language to obtain more isomorphic parallel data. Also, we translate *into* the morphologically rich language, which requires morphological modeling with regard to, e.g., grammatical case and portmanteau prepositions (cf. section 3) to ensure morphologically correct output.

Synthetic phrases have been implemented by Chahuneau et al. (2013) to translate into morphologically rich languages. They use a discriminative model based on source-side features (dependency information and word clusters) to predict inflected target words based on which phrase-table entries

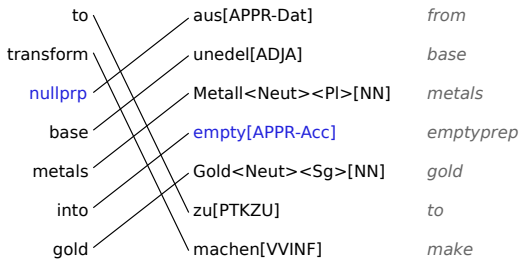

Figure 1: Example for preposition-informed representation with empty placeholders heading NPs.

are generated. They report an improvement in translation quality for several language pairs. In contrast, our approach concentrates on the generation of closed-class function words to obtain the most appropriate complement type given the source sentence. This includes generating word sequences not observed in the training data, i.e. adding/changing prepositions for a (different) PP or removing prepositions to form an NP. A task related to synthesizing prepositions is that of generating determiners, the translation of which is problematic when translating from a language like Russian that does not have definiteness morphemes. Tsvetkov et al. (2013) create synthetic translation options to augment the phrase-table. They use a classifier trained on local contextual features to predict whether to add or remove determiners for the target-side of translation rules. In contrast with determiners, which are local to their context, we model and generate function words with semantic content which are subject to complex interactions with verbs and other subcategorized elements throughout the sentence.

3 Inflection Prediction System

We work with an inflection prediction system which first translates into a stemmed representation with a component for inflecting the SMT output in a post-processing step. The stemmed representation contains markup (POS-tags and number/gender on nouns and case on prepositions, as can be seen in figure 1) which is used as input to the inflection component. Inflected forms are generated based on the morphological features *number*, *case*, *gender* and *strong/weak*, which are predicted on the SMT output using a sequence model and a morphological tool (cf. section 6.1). Modeling morphology is necessary when modifying German prepositions, as they determine grammatical case and changing a preposition might require to adapt the inflection of the respective phrase, too. Portmanteau prepositions (contracted forms of preposition and determiner) are split during the synthesizing and translation process, and are merged after the inflection step. For more details about modeling complex morphology, see for example Toutanova et al. (2008), Fraser et al. (2012) or Chahuneau et al. (2013).

4 Preposition-Informed Representation

Our first approach introduces a simple abstract representation that inserts pseudo-preposition markers to indicate the beginning of noun phrases. This representation serves two purposes: to adjust the source and target sides for structural mismatches of different complement types, and to provide information about syntactic functions and semantic roles via the annotation of grammatical case.

Placeholders for empty prepositions are inserted at the beginning of noun phrases in both the source and target language. Figure 1 provides an example of the training data with two structural mismatches: the PP on the source side *into gold* corresponds to the NP `Gold<Sg>[NN]` on the target side, and the NP on the source side (*base metals*) corresponds to the PP `aus unedel Metall` on the target side. Without the placeholders at the beginning of noun phrases, the word alignment for these phrases contains either unaligned overt prepositions[1], or imprecise one-to-many alignments containing prepositions such as "*into gold* → `Gold<Sg>[NN]`", which are wrong in many contexts.

The placeholder prepositions lead to a cleaner word alignment: the inserted empty preposition on the source side (in *nullprp base metals*) is aligned to the overt preposition `aus` on the target side, whereas the overt source preposition in *into gold* can be aligned to an empty preposition on the target side. As a consequence of the improved word alignment, the resulting system has a better coverage of individual prepositions, and the amount of prepositions being lumped together with an adjacent word via alignment is reduced. In addition, the placeholder between `Metall` and `Gold` provides an explicit phrase boundary between a PP and a direct object NP. The annotation with grammatical case provides information about the syntactic function of a phrase, such as a subject (`EMPTY-Nom`) or a direct object (`EMPTY-Acc`). For PPs, the case repre-

[1] We use the term *overt prepositions* for actually present prepositions, as opposed to "empty" prepositions.

sentence 1: nullprp beginners look **for weapons** in different ways .

sentence 2: nullprp screenshot of the site that accepts nullprp orders **for weapons** .

	1 **NP/PP src**	2 **tag src**	3 **word src**	4 **func src**	5 **head src**	6 **head trg**	7 **parent src**	8 **parV src**	9 **parV trg**	10 **parN src**	11 **parN trg**	**best-5 predicted**	
sentence 1	PP	IN	for	prep	weapon	Waffe	V	look	–	–	–	**nach-Dat**	0.349
												empty-Acc	0.224
												empty-Nom	0.206
												von-Dat	0.067
												für-Acc	0.064
sentence 2	PP	IN	for	prep	weapon	Waffe	N	–	–	order	–	**für-Acc**	0.559
												empty-Nom	0.184
												von-Dat	0.087
												nach-Dat	0.078
												empty-Acc	0.053

Table 1: Source and target side features for the prediction of placeholders in the phrase *for weapons* → `PREP Waffe<Pl>[NN]` in two sentences, using the top-5 five predictions; appropriate prepositions are bold. The prediction model corresponds to model (2) in table 7.

sents an indicator whether a preposition is part of a directional (accusative) or a locational (dative) PP.

5 Synthetic Phrase-Table Entries

Our second, extended approach generates synthetic phrases from intermediate generic placeholders. We combine source-side and target-side features to synthesize phrase-table entries that are unique for the respective source-side context.

5.1 Motivation and Example

The preposition-informed representation presents a straightforward solution to handle different structures on the source and target side. However, there are two remaining issues: first, the distribution of translation probabilities might favour a complement realization that is invalid for the respective context; and second, the required preposition might not even occur in the parallel training data as a translation of the source phrase. As a solution to these problems, we explore the idea of synthesizing phrase-table entries, in order to adjust the translation options to token-level requirements in a way that allows to take into account relevant information from the entire source sentence.

As a basis for the prediction of synthetic phrase-table entries, all empty and overt prepositions are replaced with a generic placeholder `PREP`. In the prediction step, generic placeholders are transformed into an overt or an empty preposition. Every phrase can thus be inflected as either PP or NP, depending on the sentence context. The format of the synthesized phrases corresponds to that of the preposition-informed system, with one major difference: for each source phrase, a unique set of target-phrases (possibly with new word sequences) is generated to provide an optimal set of translation options on token level.

Table 1 illustrates the first step of the process: the two sentences above the table both contain the phrase *for weapons*, which occur in different contexts. The predominant literal translation of *for* is *für*, which is however only correct in the second sentence, modifying the noun *order*. In the context of the verb *look*, the preposition *nach* or the empty preposition are correct. Thus, for the underlying target phrase `PREP Waffe<Pl>[NN]`, different prepositions need to be available for different contexts: for the first sentence, the intermediate placeholder entry should yield `nach Waffe<Pl>[NN]` and `EMPTY-Acc Waffe<Pl>[NN]`; for the second sentence, it should yield `für Waffe<Pl>[NN]` (bold in table 1). In particular, it is possible to generate target entries that have not been observed in the training data in combination with the source phrase. This is, for example, the case for `EMPTY-Acc Waffe<Pl>[NN]` which does not occur as a possible translation option of *for weapons* in the preposition-informed system.

5.2 Prediction Features

Table 1 shows the set of source-side and target-side features used to train a maximum entropy classifier for the prediction task. As phrase-table entries are often short, we rely heavily on source-side features centered around the placeholder preposition. Via dependency parses (Choi and Palmer, 2012), relevant information is gathered in the source sentence. Source information comes from the entire sentence, and may go beyond the phrase boundary, whereas

		Target	$p(e\|f)$
Prep-Informed		für[Acc]Waffe<Fem><Pl>[NN]	0.333
		nach[Dat] Waffe<Fem><Pl>[NN]	0.148
		für[Acc] nuklear<Pos>[ADJA]	0.037
		für[Acc] militärisch<Pos>[ADJA]	0.037
		für[Acc] die<+ART>[ART]	0.037
Synthetic Phrases	sentence 1	nach[Dat] Waffe<Fem><Pl>[NN]	0.192 ✓
		empty[Acc] Waffe<Fem><Pl>[NN]	0.131 ✓
		empty[Nom] Waffe<Fem><Pl>[NN]	0.121
		für[Acc] Waffe<Fem><Pl>[NN]	0.094
		von[Dat] Waffe<Fem><Pl>[NN]	0.038
	sentence 2	für[Acc] Waffe<Fem><Pl>[NN]	0.336 ✓
		empty[Nom] Waffe<Fem><Pl>[NN]	0.101
		von[Dat] Waffe<Fem><Pl>[NN]	0.045
		nach[Dat] Waffe<Fem><Pl>[NN]	0.041
		die<+ART>[ART] Waffe<Fem><Pl>[NN]	0.037

Table 2: The top-5 synthetic phrases according to $p(e|f)$ for the phrase *for weapons* based on the predictions from table 1. Phrases marked with ✓are correct in the respective context.

the target-side context is restricted to the phrase.

The source-side features comprise the type of the aligned phrase (1), the tag (2) and the word (3), as well as the syntactic function of that phrase in the source sentence (4: subj, obj, prep), and the governed noun (5: *weapon*). Furthermore, the word (verb (8) or noun (10)) governing the aligned preposition is identified and used as a feature alongside with its tag information (7: V/N). The content words from the source side, head-src (5) and parent-V/N (9,11) are then projected to the target side, if present in the phrase. In addition, up to three words to the left or right of the placeholder provide target-side context, depending on the length of the target phrase. From these features, information about the verb and the syntactic role in the source sentence are probably most important. While the content of an NP (e.g., *to order weapons/cake/etc.*) is not necessarily relevant to determine the realization of a placeholder[2], the training also relies on *feature n-grams* such as noun-verb tuples or preposition-noun-verb triples, which contain important information about subcategorizational preferences.

As training data for this model, we use all extracted source/target/alignment triples containing a relevant preposition from the preposition-informed system; the preposition with case annotation is used as the label. We record which sentence was used to extract each phrase in order to obtain the token-level source-side context. For the prediction task, the model is applied to phrase-table entries obtained on the placeholder representation: For each n-gram in the source sentence, the relevant phrase-table entries are identified and the respective features are extracted from the source sentence. Based on the top-5 predictions, along with the prediction scores, context-dependent phrase-table entries are generated. Since the complement realization also depends on lexical decisions in the target sentence (such as the verb), there are often several valid options and there is no possibility to decide for *one* particular realization without the actual target sentence context during the prediction step. We thus work with the set of n-best predictions to provide a *selection* of probable phrase-table entries given the source-sentence.

In this model, each preposition to be predicted is treated as one instance; this means that each preposition is predicted independently. In the case of several prepositions occurring in a single phrase, we consider all permutations of the respective n-best predictions.

5.3 Building the Phrase Table

To build the phrase-table with synthesized target phrases, we start by building a phrase table on data with generic placeholders, using the word alignments from the preposition-informed system. The entries are then separated into two groups: entries with and without placeholders. Entries without placeholders do not need any further processing, and are kept for the final phrase table, including translation probabilities and lexical weights. Phrase-table entries whose target side contains a placeholder are then selected to undergo the prediction step.

A prediction for all phrases is not feasible, so we restrict the table to the top-20 entries according to $p(e|f)$. This filtering is applied to the phrase table of the preposition-informed system; the phrase-table entries containing generic placeholders are then selected accordingly. With this process of phrase selection, the synthetic-phrase system and the preposition-informed system rely on the same set of underlying phrase-table entries.

5.4 Scores in Phrase and Reordering Table

A phrase table typically contains the translation probabilities $p(f|e)$ and $p(e|f)$, as well as the lexical probabilities $lex(f|e)$ and $lex(e|f)$. For the newly generated entries, new scores have to be computed: the lexical weight of a phrase can be

[2]Our experiments indicated that using features (5) and (6) as individual features tends to be harmful, whereas in combination with other features they provide useful information.

calculated based on the lexical weights of the individual words. In contrast, the translation probability of a newly generated phrase cannot be calculated. We consider the translation probability from the placeholder representation table as an approximate translation probability independent of the actual preposition; the classifier (ME) score indicates how well a particular preposition fits into the target-phrase. We present three variants to estimate the translation probabilities and then explore several ways to use the scores as features to be optimized by MERT training.

SCORE-VARIANT 1: The placeholder translation probability and the ME scores are used as separate features. An indicator feature counts the predicted prepositions. Non-synthesized phrases get a pseudo ME-score of 1, and $exp(0)$ for the indicator feature. In the case of $n > 1$ prepositions, the ME scores are multiplied, and the indicator feature is set to $exp(n)$.

SCORE-VARIANT 2: Variant 1 is extended with the product of the placeholder translation probabilities and the ME score, to account for cases where lexically bad translation options received a high ME score and thus are boosted erroneously.

SCORE-VARIANT 3A: We consider the placeholder translation probability as the probability of a phrase to contain *some* preposition and use it as the basis to calculate a score for the phrase to contain the *predicted* preposition, using the ME score. Note, however, that the prediction score does not provide the probability of the target phrase representing a translation of the source phrase, but only how well the predicted preposition fits into the target phrase; this leads to potentially high ME scores for bad translation options. For this reason, we "dampen" the prediction score with the lexical probability as an indicator for the quality of the source-target pair, resulting in the following formula:

$$P_{prep}(e|f) = P_{PlaceHolder}(e|f) * (ME + lex(e|f))$$

where ME is the prediction score and $P_{PlaceHolder}$ is the translation probability based on the placeholder representation. lex is the lexical probability based on the phrase containing the generated prepositions. In a variant (3b), the resulting translation probability scores are then normalized such that they sum to 1 with the entries without prepositions, whose probability mass remains unchanged and corresponds to that in the preposition-informed system. This aims at obtaining a "real" probability distribution with context-dependent scores for phrases containing prepositions that is as close as possible to that in the preposition-informed system: probabilities of phrases without prepositions remain the same, whereas the scores for the generated phrases are normalized to share the remaining probability mass given a source phrase.

In variants 1 and 2, the ME-based scores are used as additional features to the lexical and placeholder translation probabilities, whereas in variant 3, new phrase-translation probabilities are computed based on the placeholder probabilities and the prediction scores to replace the placeholder probabilities. Table 2 shows the generated entries and the scores for $p(e|f)$ according to score variant 3b for the predictions from table 1; suitable translation options are marked with ✓. For sentence 1, the two possible variants *nach* and *empty* are top-ranked, whereas the top entry from the preposition-informed system, *für*, is unlikely to be selected in this context. For sentence 2, the top-ranked preposition *für* is even more likely than in the preposition-informed system. The entries for both sentence 1 and sentence 2 show that the previous two top-ranked candidates (`für Waffe<Pl>[NN]` and `nach Waffe<Pl>[NN]`) are now expanded and take up the top-5 positions for sentence 1 and the top-4 positions for sentence 2. As a result, the lexically invalid options on positions 3-5 from the preposition-informed system are disfavoured.

For the reordering table, we use the statistics from the placeholder representation. We assume that no changes in the reordering are caused by modifying the complement type or modifying prepositions; this assumption was verified experimentally (details are omitted).

6 Experiments and Results

We compare the preposition-informed system with the synthetic-phrases system where we explore different ways to integrate the synthetic phrases.

6.1 Experimental Setup

All systems were built using the Moses phrase-based framework. We used 4.592.139 parallel sentences aligned with GIZA++ for translation model training, and 45M sentences (News14+parallel data) to build a 5-gram language model. We used NewsTest13 (3000 sentences) for development and NewsTest14 (3003 sentences) as test set. These

System		BLEU
baseline-1	Surface forms	19.17
baseline-2	Stemmed	19.35
prep-informed system (P-1)	Stemmed + ∅-CASE	19.76
prep-informed system (P-2)	Stemmed + ∅-CASE-top-20	19.73

Table 3: Scores for baselines and preposition-informed system.

System	Features used for MERT tuning	BLEU
SP-1	SCORE-VARIANT-1	19.76
SP-2	SCORE-VARIANT-2	19.83
SP-3a	SCORE-VARIANT-3	19.80
SP-3b	SCORE-VARIANT-3, norm. $P_{prep}(e\|f)$	19.86*

Table 4: Variants of the synthetic-phrases system. * marks significant improvement over system P-2 (with pair-wise bootstrap resampling with sample size 1,000 and a p-value of 0.05)

datasets are from the WMT2015 shared task.

To predict the four morphological features number, gender, case and strong/weak for inflecting the stemmed output, we trained 4 CRF sequence models on the target-side of the parallel data. These features are predicted as a sequence of labels (i.e. case/number/etc of consecutive words in an NP/PP) at sentence level. For the prediction of the placeholder prepositions, we trained a maximum entropy model on the parallel training data. In contrast to the morphological features, each preposition in a phrase is predicted independently. For all models, we used the toolkit Wapiti (Lavergne et al., 2010). The German data was parsed with BitPar (Schmid, 2004) and German inflected forms were generated with the morphological resource SMOR (Schmid et al., 2004).

6.2 Baselines

We consider two baselines:

BASELINE-1: a standard phrase-based translation system trained on surface forms without any form of morphological modeling.

BASELINE-2: a system with morphological modeling, as described in section 3. Portmanteau prepositions are split into preposition and article prior to translation and merged in a post-processing step. Otherwise, prepositions are not modeled.

6.3 Results

The preposition-informed system contains overt prepositions and empty prepositions annotated with grammatical case at the beginning of noun phrases, as described in section 4. Empty prepositions are simply deleted from the SMT output after translation before generating inflected forms. The introduction of empty prepositions into the training data leads to statistically significant improvements in BLEU over both the surface system (baseline-1) and the inflection prediction system (baseline-2), cf. Table 3. Furthermore, restricting the phrase-table to the top-20 entries according to $p(e|f)$ (system P-2) does not decrease performance.

Table 4 shows the results for the variants of the synthetic-phrases systems, which all significantly outperform baseline-2. Even though the difference is small, the best system (SP-3b) is significantly better than system P-2, the preposition-informed system using the top-20 translation table entries. It is, however, not significantly better than system P-1, which uses all phrase-table entries. This is reasonable considering that SP-3b is built from placeholder entries based on the same phrase inventory as system P-2.

The system with the lowest score (SP-1) uses lexical and placeholder phrase probabilities combined with the ME prediction scores and the count feature. System SP-2, extended with the product of the phrase translation probability and the ME score, yields a slightly better result. For system SP-3, in which new phrase-translation probabilities replace the placeholder probabilities, we compare a version with and without normalized $p(e|f)$ scores: the normalization leads to a best overall score; all synthetic-phrases systems score in a similar range, however.

7 Discussion

In this section, we summarize the results and in particular, discuss the use of newly generated phrases. We also attempt to analyze potential side-effects on the phrase table and present additional experiments to better handle these effects.

7.1 Summary of Results

The insertion of placeholder prepositions leads to an improvement over both baselines due to the cleaner alignment enabled by the more similar source and target sides. Furthermore, the empty prepositions can function as phrase boundaries and provide "flat structural" information in the form of annotated grammatical case.

The synthetic-phrases approach aims at generating a context-sensitive variant of the preposition-

	SP-1	SP-2	SP-3a	SP-3b
new	1489	1507	1391	1398
regular	38132	34541	35101	33571

Table 5: Number of newly generated and regular phrase-table entries used to translate the test set (3003 sentences).

informed system that is able to generate new entries if needed. We explored different score settings, either as separate features (variants 1/2) or combined into a translation probability score in (variant 3). While all variants perform similarly, the best system is significantly better than the preposition-informed system built on the top-20 phrase-table entries. This shows that the proposed method of synthetic phrases indeed improves translation quality. However, the difference is very small and only applies to one pair of system variants, which makes it difficult to draw a solid conclusion.

7.2 Use of Newly Generated Phrases

An important property in the presented method is the ability to generate new phrases. Table 5 shows the distribution of phrases used to translate the test set. For the 3003 sentences, roughly 1500 new phrases have been applied; on average, this corresponds to about one new phrase in one out of two sentences. Given that function words usually are thought to be well-covered in NLP training data, this number is substantial.

The following example illustrates how newly generated translation options can improve translations by closing coverage gaps. Table 6 shows the translations for an input sentence (EN) of the preposition-informed system P-2 and the synthetic-phrases system SP-2. The two outputs are identical and both correct, except for the wrong preposition *zur* in system P-2. To translate the sentence with the synthetic-phrases system, these new translation options[3] have been used:

the deutsche bahn → die ∅-Nom deutsche Bahn
to improve ∅ the → **auf-Acc** eine Verbesserung ∅-Gen der
railway line in → Eisenbahnlinie in-Dat

In particular, the phrase pair "*to improve ∅ the* → ***auf-Acc** eine Verbesserung ∅-Gen der*" enables a translation with the correct preposition. Due to the segmentation of the sentence, the English verb *hope* is translated as part of another phrase, which excludes a translation as one unit such as *hope to* → *hoffen auf*. Furthermore, there is a structural shift between the source side phrase "*hope to improve*$_{\text{VERB}}$", and the German sentence with the structure "*hofft* PREP *Verbesserung*$_{\text{NOUN}}$". The incorrect *zu* in the preposition-informed system would be a valid connection to a following verb, but cannot be used to introduce a PP in this context.

[3]Shown in inflected format for better readability.

EN	nullprp the deutsche bahn hopes to improve nullprp the kinzigtal railway line in the coming year.
P-2	die deutsche Bahn hofft **zur** Verbesserung der kinzigtal Eisenbahnlinie im kommenden Jahr.
SP-2	die deutsche Bahn hofft **auf eine** Verbesserung der kinzigtal Eisenbahnlinie im kommenden Jahr.

Table 6: Improved translation output by applying a newly generated translation option.

7.3 Side-Effects on the Phrase-Table

A recurring problem in the synthetic-phrases system are lexically wrong translations that are boosted due to unreasonably high ME scores in comparison to lexically more correct options. In particular, this is the case when infrequent words occur within a lexically wrong translation, which also happens to have lexical and phrase translation probabilities in a similar range as better translation candidates. When predicting prepositions for such phrases, the ME model is often overly confident and outputs comparatively high prediction scores based on an insufficient amount of training examples[4].

Consider as an example the English phrase *for bags* and two of its translation options: "PREP *Taschen*" ('bags') and "PREP *Müllsäcke*" ('garbage bags'), which have similar translation and lexical probabilities. In the ME training data, there are only very few occurrences of PREP *Müllsäcke*. As a result, the ME very confidently reproduces the seen training instances with a score around 0.9 for the top-ranked preposition. In comparison, the predictions for PREP *Taschen* are more balanced due to more occurrences of this word, with a score of around 0.55 for the top-ranked preposition. Thus, the incorrect *für Müllsäcke* option is boosted by its prediction score and consequently gets chosen by the synthetic-phrases system.

Lexical features, e.g., in verb-noun tuples, are important for the prediction power of our ME model. However, the example above illustrates how infrequent words can be harmful. We addressed

[4]Note that the model must be trained on parallel data only as it makes use of source-side features.

	SP-1	SP-2	SP-3a	SP-3b
(1) no infreq nouns	19.59	19.85	19.71	19.94*
(2) reduced data	19.82	19.58	19.73	19.64

Table 7: Results when filtering out infrequent nouns in the ME training data (1) or reducing the amount of source-target-alignment triples used for ME training (2). * marks significant improvement over system P-2.

this problem by weighting down the prediction scores using lexical and/or phrase translation probabilities. In addition, we also experimented with replacing infrequent words with dummy tokens to still benefit from lexical information while excluding insufficiently represented words. The first line in table 7 shows the results for prediction models trained on data where infrequent nouns (*freq* $<$ *25*) occurring in the NP/PP (features 5 and 6 in table 1) are omitted when training the prediction ME. The general outcome is similar to the experiments reported in table 4, with variants SP-2 and SP-3b being slightly better. The result for system SP-3b is the overall best result. This suggests that a careful representation of infrequent lexical items in training data benefits the prediction quality.

In an attempt to reduce the training data to relevant entries, we restricted the source-target-alignment triples used to train the prediction ME to those occurring in the top-20 filtered table. Thus, all entries in the phrase-table are covered by the model, while infrequent and non-relevant training instances are mostly omitted. The results are listed in the second line in table 7; however, this model leads to generally worse results than the previous ones. We assume that removing a subset of training triples leads to a somewhat unbalanced training set.

7.4 Distribution in Phrase Table

Another, potentially negative, effect on the phrase distribution in the phrase table stems from integrating the n-best predictions per place-holder entry: an already dominant translation option can be further reinforced if it does not only represent the top-most translation option (as in the preposition-informed or place-holder table), but can be expanded to several entries. An equally valid, but less probable translation option is then less accessible if its prediction scores are in the same range, as this translation is then dispreferred by its translation scores and has to compete with several entries stemming from the original top translation option. Consider the example of the phrase "*expand nullprp their*": in the preposition-informed system, the lexically correct translation *erweitern* `EMPTY-Acc` *ihre* is ranked third according to $p(e|f)$, with two meaningless translations (only determiner or only preposition) as the two top-ranked translations, which is already a bad starting point for translation of the verb. In the synthetic-phrases system, "descendants" of the previously top-2 meaningless translations now are expanded and fill the positions 1-5, resulting in the correct translation option being ranked 6th.

	prep-informed		synth-phrases	
	missing	wrong	missing	wrong
verbs	32	11	23	10
nouns	2	15	2	17
prepositions	6	6	3	8
gram. case	–	4	–	3

Table 8: Manual error analysis of 50 randomly selected sentences.

This effect can also be positive by promoting lexically correct translation options (in cases where the leading translation is correct, but is closely followed by a less suited translation). For example, it can be seen in the example in table 2 where the lexically incorrect phrases are moved to lower positions. However, it might also happen that literal translations are preferred over less common senses in cases of word sense ambiguities. A small manual evaluation (cf. next section) showed that slightly more verbs are translated with the synthetic phrases system. Verbs in English-to-German translation are often omitted during translation; the effect of enhancing literal translations might be responsible for the observed tendency to translate more verbs.

The different score variants explored in the previous section aim to find a combination that considers these factors, but the results show that it is a difficult task to account for all possible interactions.

7.5 Manual Evaluation

We carried out a small manual evaluation for 50 sentences (length 10-20 words) randomly chosen from system SP-3b in table 7, the best overall system, in comparison to the preposition-informed system P-2. Two native speakers annotated errors concerning missing or incorrect verbs, nouns and prepositions, as well as incorrect grammatical case. Table 8 depicts the outcome: The number of errors found in the categories *preposition* and *grammatical case* are similar for both systems. A slight improvement

EN	this is mainly *due to the higher contribution* from the administrative budget ...
P-2	das ist hauptsächlich ***auf*** *die höheren Beiträge* aus dem Verwaltungshaushalt ...
SP-3b	das ist vor allem ***wegen*** *den höheren Beiträgen* aus dem Verwaltungshaushalt ...

Table 9: Example for unclear error categories.

is found, however, for the number of translated verbs, which are known to be generally difficult for the language pair English-to-German. We assume that this is due to a tendency to strengthen literal translations, from which verbs might benefit as they are generally less well represented in the phrase-table.

Note, however, that there are other relevant factors that this manual evaluation does not take into account, such as, e.g., the overall structure of the sentence. Furthermore, the evaluation of verbs and its subcategorized elements is often difficult as there might be several valid options for annotation, which is illustrated by the example in table 9. The translations of the two systems are nearly identical, except for the prepositions heading the translation for *due to the higher contribution* (and consequently the realization of grammatical case in the respective phrases, which is correct given the respective preposition). The sentence produced by the synthetic-phrases system is correct, preserving the structure of the English sentence by translating *due to* as *wegen+Dative* (*wegen+Genitive* would be correct, too.). Thus, replacing the preposition *auf* and adjusting the grammatical case in the sentence produced by the preposition-informed system would lead to the same, valid, translation. However, the preposition *auf* strongly triggers the reader to expect the verb *zurückführen (auf)* ('to attribute (to)') which also would lead to a valid translation. Such cases make the evaluation of prepositions and complement types difficult, as the error category (*missing verb* or *wrong preposition*) is not always clear.

8 Conclusion and Future Work

We compared two approaches for modeling complement types in English-to-German SMT. Our experiments showed that explicit information about different complement types (insertion of empty placeholders) leads to improved SMT quality. The results of the synthetic-phrases system are slightly better than those of the preposition-informed system, with two variants being significantly better. As the differences are rather small and apply only to some system pairs, it is difficult to draw a clear conclusion concerning the effectiveness of the synthetic-phrases method. Our analysis showed, however, that newly generated phrases are indeed used within the systems and help to improve translation quality. We consider this a confirmation that the generation of synthetic phrases for handling subcategorization is a sound approach.

In future work, we plan to explore models that predict the complete target *phrase* given the source phrase and subcategorization-relevant features instead of predicting the *preposition* in a target phrase.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402 (HimL), from the European Research Council (ERC) under grant agreement No 640550, from the DFG grants *Distributional Approaches to Semantic Relatedness* and *Models of Morphosyntax for Statistical Machine Translation (Phase Two)* and from the DFG Heisenberg Fellowship SCHU-2580/1.

References

Eneko Agirre, Aitziber Atutxa, Gorka Labaka, Mikel Lersundi, Aingeru Mayor, and Kepa Sarasola. 2009. Use of Rich Linguistic Information to Translate Prepositions and Grammatical Cases to Basque. In *Proceedings of the 13th Annual Conference of the EAMT*, Barcelona, Spain.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic.

Victor Chahuneau, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. Translating into Morphologically Rich Languages with Synthetic Phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, Washington.

Jinho D. Choi and Martha Palmer. 2012. Getting the Most out of Transition-Based Dependency Parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling Inflection and Word-

Formation in SMT. In *Proceedings of the the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France.

Kevin Gimpel and Noah A. Smith. 2008. Rich Source-Side Context for Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio.

Minwoo Jeong, Kristina Toutanova, Hisami Suzuki, and Chris Quirk. 2010. A Discriminative Lexicon Model for Complex Morphology. In *In Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas.*, Denver, Colorado.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: a German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings LREC 2004*, Lisbon, Portugal.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the International Conference on Computational Linguistics*.

Reshef Shilon, Hanna Fadida, and Shuly Wintner. 2012. Incorporating Linguistic Knowledge in Statistical Machine Translation: Translating Prepositions. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, EACL 2012*, Avignon, France.

Aleš Tamchyna, Fabienne Braune, Alexander Fraser, Marine Carpuat, Hal Daumé III, and Chris Quirk. 2014. Integrating a Discriminative Classifier into Phrase-based and Hierarchical Decoding. In *The Prague Bulletin of Mathematical Linguistics, Number 101*, pages 29–41.

Jörg Tiedemann, Filip Ginter, and Jenna Kanerva. 2015. Morphological Segmentation and OPUS for Finnish-English Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *Proceedings of ACL08-HLT*, Columbus, Ohio.

Julia Tsvetkov, Chris Dyer, Lori Levin, and Archna Bhatia. 2013. Generating English Determiners in Phrase-Based Translation with Synthetic Translation Options. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria.

Marion Weller, Alexander Fraser, and Sabine Schulte im Walde. 2013. Using Subcategorization Knowledge to Improve Case Prediction for Translation to German. In *Proceedings of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Marion Weller, Alexander Fraser, and Sabine Schulte im Walde. 2015. Target-Side Generation of Prepositions for SMT. In *Proceedings of EAMT 2015*, Antalya, Turkey.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, and Philpp Koehn. 2015. Edinburgh's Syntax-Based Systems at WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal.