Using subcategorization knowledge to improve case prediction for translation to German

Marion Weller[1] **Alexander Fraser**[2] **Sabine Schulte im Walde**[1]

[1]Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
{wellermn|schulte}@ims.uni-stuttgart.de

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

Abstract

This paper demonstrates the need and impact of subcategorization information for SMT. We combine (i) features on source-side syntactic subcategorization and (ii) an external knowledge base with quantitative, dependency-based information about target-side subcategorization frames. A manual evaluation of an English-to-German translation task shows that the subcategorization information has a positive impact on translation quality through better prediction of case.

1 Introduction

When translating from a morphologically poor language to a morphologically rich language we are faced with two major problems: (i) the richness of the target-language morphology causes data sparsity problems, and (ii) information about morphological features on the target side is not sufficiently contained in the source language morphology.

We address these two problems using a two-step procedure. We first replace inflected forms by their stems or lemmas: building a translation system on a stemmed representation of the target side leads to a simpler translation task, and the morphological information contained in the source and target language parts of the translation model is more balanced. In the second step, the stemmed output of the translation is then inflected: the morphological features are predicted, and the inflected forms are generated using the stem and predicted morphological features.

In this paper, we focus on improving *case* prediction for noun phrases (NPs) in German translations. The NP feature *case* is extremely difficult to predict in German: while the NP features *gender* and *number* are part of the stem or can be derived from the source-side input, respectively, the prediction of *case* requires information about the subcategorization of the entire clause. This is due to German being a less configurational language than English, which encodes grammatical relations (e.g. subject-hood, object-hood, etc.) through the position of constituents. German sentences exhibit a freer constituent order, and thus case is an important indicator of the grammatical functions of noun phrases. Correct case prediction is a crucial factor for the adequacy of SMT output, cf. the example in table 1 providing an erroneously inflected output (this is taken from a baseline "simple inflection prediction" system, cf. section 5.2). The translation of the English input sentence in terms of stems is perfectly acceptable; after the inflection step, however, the translation of NP_4 *ongoing military actions* represents a genitive modifier of the subject NP_2, instead of a direct object NP of the verb *anordnen* (*to order*). The meaning is thus *why the government of the ongoing military actions ordered*, which has only one NP and is completely wrong.

The translation in table 1 needs verb subcategorization information. This is demonstrated by the invented examples (1) and (2):

(1) [Der Mitarbeiter]$_{NPnom}$ hat [den Bericht]$_{NPacc}$ [dem Kollegen]$_{NPdat}$ gegeben.
[The employee]$_{NPnom}$ gave [his colleague]$_{NPdat}$ [the report]$_{NPacc}$

(2) [Der Mitarbeiter]$_{NPnom}$ hat [dem Bericht]$_{NPdat}$ [des Kollegen]$_{NPgen}$ zugestimmt.
[The employee]$_{NPnom}$ agreed [on the report]$_{PP}$ [of his colleague]$_{PP}$

Both inflected sentences rely on the stem sequence [*d Mitarbeiter*] [*d Bericht*] [*d Kollege*] ⟨*verb*⟩, so the case assignment can only be determined by the verb: While *geben (to give)* has a strong preference for selecting a ditransitive subcategorization frame[1], including an agentive subject (nomi-

[1]A ditransitive verb takes a subject and two objects.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 593–603,
Sofia, Bulgaria, August 4-9 2013.

input		[why]$_1$ [the government]$_2$ [ordered]$_3$ [the ongoing military actions]$_4$
output	stemmed	[warum]$_1$ [d Regierung]$_2$ [d anhaltend militärisch Aktion]$_4$ [angeordnet]$_3$
	inflected	[warum]$_1$ [**die** Regierung]$_2$ [**der** anhaltend**en** militärisch**en** Aktion**en**]$_4$ [angeordnet]$_3$

Table 1: Example for case confusion in SMT output when using a simple prediction system.

native case), a benefactive (dative case) and a patient (accusative case), *zustimmen (to agree)* has a strong preference for only selecting an agentive subject (nominative case) and an indirect object theme (dative case). So in the latter case the NP [*d Kollege*] cannot receive case from the verb and is instead the genitive modifier of the dative NP.

While for examples (1) and (2) knowledge about the syntactic verb subcategorization functions is sufficient to correctly predict the NP cases, examples (3) to (6) require subcategorization information at the syntax-semantic interface.

(3) [Der Mitarbeiter]$_{NPnom}$ hat [dem Kollegen]$_{NPdat}$ [den Bericht]$_{NPacc}$ gegeben.

(4) [Der Mitarbeiter]$_{NPnom}$ hat [den Bericht]$_{NPacc}$ [dem Kollegen]$_{NPdat}$ gegeben.

(5) [Dem Kollegen]$_{NPdat}$ hat [der Mitarbeiter]$_{NPnom}$ [den Bericht]$_{NPacc}$ gegeben.

(6) [Den Bericht]$_{NPacc}$ hat [der Mitarbeiter]$_{NPnom}$ [dem Kollegen]$_{NPdat}$ gegeben.

In all four examples, the verb and the participating noun phrases *Mitarbeiter (employee)*, *Kollege (colleague)* and *Bericht (report)* are identical, and the noun phrases are assigned the same case. However, given that the stemmed output of the translation does not tell us anything about case features, in order to predict the appropriate cases of the three noun phrases, we either rely on ordering heuristics (such that the nominative NP is more likely to be in the beginning of the sentence (the German *Vorfeld*) than the accusative or dative NP, even though all three of these would be grammatical), or we need fine-grained subcategorization information beyond pure syntax. For example, both *Mitarbeiter* and *Kollege* would satisfy the agentive subject role of the verb *geben* better than *Bericht*, and *Bericht* is more likely to be the patient of *geben*.

The contribution of this paper is to improve the prediction of case in our SMT system by implementing and combining two alternative routes to integrate subcategorization information from the syntax-semantic interface: (i) We regard the translation as a function of the source language input, and project the syntactic functions of the English nouns to their German translations in the SMT output. This subcategorization model is necessary when there are several plausible solutions for the syntactic functions of a noun in combination with a verb. For example, both *Mitarbeiter* and *Kollege* are plausible subjects and direct objects of the verb *geben*, so the information about these nouns' roles in the input sentence allows for disambiguation. (ii) The case of an NP is derived from an external knowledge base comprising quantitative, dependency-based information about German verb subcategorization frames and noun modification. The verb subcategorization information is not restricted to syntactic noun functions but models association strength for verb–noun pairs with regard to the entire subcategorization frame plus the syntactic functions of the nouns. For example, the database can tell us that while the verb *geben* is very likely to subcategorize a ditransitive frame, the verb *zustimmen* is very likely to subcategorize only a direct object, next to the obligatory subject **(subcat frame prediction)**. Furthermore, we can retrieve the information that the noun *Bericht* is less likely to appear as subject of *geben* than the nouns *Mitarbeiter* and *Kollege* **(verb–noun subcat case prediction)**. And we can look up that the noun *Aktion* is very unlikely to be a genitive modification of *Regierung* (cf. table 1), while *Kollege* is a plausible genitive modification of *Bericht* (**noun–noun modification case prediction**, cf. example (2)).

In summary, model (i) applies when there are no obvious preferences concerning verb–noun subcategorization or noun–noun modification. Model (ii) predicts case relying on the subcategorization and modification preferences. The combination of our two models approaches a simplified level of semantic role definition but only relies on dependency information that is considerably easier and cheaper to define and obtain than a very high quality semantic parser and/or a corpus annotated with semantic role information. Integrating semantic role information into SMT has been demonstrated by various researchers to improve translation quality (cf. Wu and Fung (2009a), Wu and Fung (2009b), Liu and Gildea (2008), Liu and Gildea (2010)). Our approach is in line with

Wu and Fung (2009b) who demonstrated that on the one hand 84% of verb syntactic functions in a 50-sentence test corpus projected from Chinese to English, and that on the other hand about 15% of the subjects were not translated into subjects, but their semantic roles were preserved across language. These two findings correspond to the expected uses of our models (i) and (ii), respectively.

2 Previous work

Previous work has already introduced the idea of generating inflected forms as a post-processing step for a translation system that has been stripped of (most) target-language-specific features. Toutanova et al. (2008) and Jeong et al. (2010) built translation systems that predict inflected word forms based on a large array of morphological and syntactic features, obtained from both source and target side. Kholy and Habash (2012) and Green and DeNero (2012) work on English to Arabic translation and model gender, number and definiteness, focusing primarily on improving fluency.

Fraser et al. (2012) used a phrase-based system to transfer stems and generated inflected forms based on the stems and their morphological features. For case prediction, they trained a CRF with access to lemmas and POS-tags within a given window. We re-implemented the system by Fraser et al. as a hierarchical machine translation system using a string-to-tree setup. In contrast to the flat phrase-based setting of Fraser et al. (2012), syntactic trees on the SMT output allow us to work with verb–noun structures, which are relevant for case prediction. While the CRF used for case prediction in Fraser et al. (2012) has access to lexical information, it is limited to a certain window size and has no direct information about the relation of verb–noun pairs occurring in the sentence. Using a window of a limited size is particularly problematic for German, as there can be large gaps between the verb and its subcategorized nouns; introducing information about the relation of verbs and nouns helps to bridge such gaps. Furthermore, that model was not able to make effective use of source-side features.

One of the objectives of using an inflection prediction model is morphologically well-formed output. Kirchhoff et al. (2012) evaluated user reactions to different error types in machine translation and came to the result that morphological well-formedness has only a marginal impact on the comprehensibility of SMT output in the case of English-Spanish translation. As already discussed, German *case* is essential to the meaning of the sentence, so this result will not hold for German output.

3 Translation pipeline

This section presents an overview of our two-step translation process. In the first step, English input is translated to German stems. In the second step, morphological features are predicted and inflected forms are generated based on the word stems and the morphological features. In subsections 3.1 to 3.4, we present the simple version of the inflection prediction system; our new features are described in sections 4.2 and 4.3.

3.1 Stemmed representation/feature markup

We first parse the German side of the parallel training data with BitPar (Schmid, 2004). This maps each surface form appearing in normal text to a stem and morphological features (case, gender, number). We use this representation to create the stemmed representation for training the translation model. With the exception of stem-markup (discussed below), all morphological features are removed from the stemmed representation. The stem markup is used as part of the input to the feature prediction; the basic idea is that the given feature values are picked up by the prediction model and then propagated over the phrase.

Nouns, as the head of NPs and PPs, are annotated with *gender* and *number*. We consider gender as part of the stem, whereas the value for number is derived from the source-side: if marked for number, singular/plural nouns are distinguished during word alignment and then translated accordingly. Prepositions are also annotated with case; many prepositions are restricted to only one case, some are ambiguous and allow for either *dative* or *accusative*. Other words which are subject to feature prediction (e.g. adjectives, articles) are reduced to their stems with no feature markup, as are all remaining words. As sole exception, we keep the inflected forms of verbs (verbal inflection is not modelled). In addition to the translation model, the target-side language model, as well as the reference data for parameter tuning use this representation.

3.2 Building a stemmed translation model

We use a hierarchical translation system. Instead of translating phrases, a hierarchical system extracts translation rules (Galley et al., 2004) which allow the decoder to provide a tree spanning over the translated sentence. In order to avoid sparsity during rule extraction, we use a string-to-tree setup, where only the target-side part of the data is parsed. Translation rules are of the following form:

$[X]_1$ allows $[X]_2$ → $[NP]_1$ $[NP]_2$ erlaubt
$[X]_1$ allows $[X]_2$ → $[NP]_1$ erlaubt $[NP]_2$

This example illustrates how rules can cover the different word ordering possibilities in German.

PP nodes are annotated with their respective case, as well as with the lemma of the preposition they contain. In our experiments, this enriched annotation has small improvements over the simpler setting with only head categories (details omitted). This outcome, in particular that adding the lemma of the preposition to the PP node helps to improve translation quality, has been observed before in tree restructuring work for improving translation (Huang and Knight, 2006).

3.3 Feature prediction and generation of inflected forms

In this section we discuss our focus, which is prediction of case, but also the prediction of number, gender and strong/weak adjectival inflection. The latter feature is German-specific; its values[2] (strong/weak) depend on the combination of the other features, as well as on the type of determiner (e.g. definite/indefinite/none).

Morphological features are predicted on four separate CRF models, one for each feature. The models for case, number and gender are independent of another, whereas the model for adjectival inflection requires information about these features, and is thus the last one to be computed, taking the output of the 3 other models as part of its input. In contrast, the adjectival inflection model in Fraser et al. (2012) is independent from the other features. Each model has access to stems, POS-tags and the feature to be modelled within a window of four positions to the right and the left of the current position[3].

Table 2 illustrates the different steps of the inflection process: the markup (number and gender on nouns) in the stemmed output of the SMT system is part of the input to the respective feature prediction. For gender and number, the values given on the stems of the nouns are then propagated over the phrase. While the case of prepositional phrases is determined by the case annotation on prepositions, the case of nominal phrases is computed only based on the respective contexts. After predicting all morphological features, the information required to generate inflected forms is complete: based on the stems and the features, we use the morphological tool SMOR (Schmid et al., 2004) for the generation of inflected forms.

One general problem with feature-prediction is that the ill-formed SMT output is not well represented by the training data which consists of well-formed sentences. This problem was also mentioned by Stymne and Cancedda (2011) and Kholy and Habash (2012). They deal with this problem by translating the training data and annotating it with the respective features, and then adding this new data set to the original training data. As this method comes with its own problems, such as transferring the morphological annotation to not necessarily isomorphically translated text, we do not use translated data as part of the training data. Instead, we limit the power of the CRF model through experimenting with the removal of features, until we had a system that was robust to this problem.

3.4 Dealing with word formation issues

To reduce data sparsity, we split portmanteau prepositions. Portmanteaus are compounds of prepositions and articles, e.g. *zur* = *zu der* (*to the*). Being components of nominal phrases, they have to agree in all morphological features with the rest of the phrase. As only some combinations of articles and prepositions can form a portmanteau, the decision of whether to merge prepositions and articles is made after feature prediction. Since our focus is case prediction, we do not do special modelling of German compounds.

4 Using subcategorization information

Within the area of (automatic) lexical acquisition, the definition of lexical verb information has been a major focus, because verbs play a central role for the structure and the meaning of sentences and

[2]Note that the values for strong/weak inflection are not always the same over the phrase, but follow a certain pattern depending on the settings of case, number and gender.

[3]Preliminary experiments showed that larger windows do not improve translation quality.

SMT output	predicted features	inflected forms	gloss
beeinflussen<VVFIN>	–	beeinflussen	*influence*
d<ART>	Fem.Acc.Sg.St	die	*the*
politisch<ADJ>	Fem.Acc.Sg.Wk	politische	*political*
Stabilität<NN><**Fem**><**Sg**>	Fem.Acc.Sg.Wk	Stabilität	*stability*

Table 2: Overview of the inflection process: the stem markup is highlighted in the SMT output.

discourse. On the one hand, this has led to a range of manually or semi-automatically developed lexical resources focusing on verb information, such as the Levin classes (Levin, 1993), VerbNet (Kipper Schuler, 2006), FrameNet[4] (Fillmore et al., 2003), and PropBank (Palmer et al., 2005). On the other hand, we find automatic approaches to the induction of verb subcategorization information at the syntax-semantics interface for a large number of languages, e.g. Briscoe and Carroll (1997) for English; Sarkar and Zeman (2000) for Czech; Schulte im Walde (2002a) for German; Messiant (2008) for French. This basic kind of verb knowledge has been shown to be useful in many NLP tasks such as information extraction (Surdeanu et al., 2003; Venturi1 et al., 2009), parsing (Carroll et al., 1998; Carroll and Fang, 2004) and word sense disambiguation (Kohomban and Lee, 2005; McCarthy et al., 2007).

4.1 Extracting subcategorization information

As described in the introductory section, we make use of two[5] major kinds of subcategorization information. Verb–noun tuples referring to specific syntactic functions within verb subcategorization **(verb–noun subcat case prediction)** are integrated with an associated probability for *accusative* (direct object), *dative* (indirect object) and *nominative* (subject).[6] Further to the subject and object noun phrases, the subcategorization information provides quantitative triples for verb–preposition–noun pairs, thus predicting the case of NPs within prepositional phrases (we do this only when the prepositions are ambiguious, i.e., they could subcategorize either a dative or an accusative NP). In addition to modelling subcategorization information, it is also important to differentiate between subcategorized noun phrases (such as object or subject), and noun phrases

[4]Even though the FrameNets approach does not only include knowledge about verbal predicates, the actual lexicons are skewed towards verb behaviour.

[5]The third kind of information, **subcat frame prediction** is implicit, since verb–noun tuples rely on specific frames.

[6]Genitive objects can also occur in German verb subcategorization frames, but this is extremely rare and verb-specific and thus not considered in our model.

	V-SUBJ	**V-OBJ**$_{Acc}$	**V-OBJ**$_{Dat}$
EP	454,350	332,847	53,711
HGC	712,717	329,830	160,377
Both	1,089,492	607,541	206,764

Table 3: Number of verb-noun types extracted from Europarl (EP) and newspaper data (HGC).

that modify nouns **(noun–noun modification case prediction)**. Typically, these NP modifiers are genitive NPs. To this end, we integrate noun-noun$_{Gen}$ tuples with their respective frequencies. These preferences for a certain function (i.e. subject, object or modifier) are passed on to the system at the level of nouns and integrated into the CRF through the derived probabilities.

The tuples and triples are obtained from dependency-parsed data by extracting all occurrences of the respective relations; table 3 gives an overview of the number of extracted tuple types. For the subcategorization information, the verb-noun tuples (verb-subject, verb-object$_{Acc}$, verb-object$_{Dat}$) are then grouped as follows:

tuple	gloss	Acc	Dat	Nom
Schema$_N$ folgen$_V$	*pattern follow*	0	322	19

We compute the probabilities for the verb-noun tuple to occur in the respective functions based on the relative frequencies. In the case of *Schema$_N$ folgen$_V$*, we find that the function of *Schema* as dative object is predominant (*to follow a pattern*), but it can also occur in the subject position (*the pattern follows*). The fact that two functions are possible for this noun are reflected in their probabilities. The probabilities are discretized into 5 buckets ($B_{p=0}$, $B_{0<p\leq 0.25}$, $B_{0.25<p\leq 0.5}$, $B_{0.5<p\leq 0.75}$, $B_{0.75<p\leq 1}$). In contrast, noun modification in noun-noun$_{Gen}$ construction is represented by co-occurrence frequencies.[7]

[7]The frequencies are bucketed to the powers of ten, i.e. $f = 1, 2 \leq f \leq 10, 11 \leq f \leq 100$, etc. and also $f = 0$: this representation allows for a more fine-grained distinction in the low-to-mid frequency range, providing a good basis for the decision of whether a given noun-noun pair is a true noun-noun$_{Gen}$ structure or just a random co-occurrence of two nouns.

	Gloss	Stem	Tag	Acc	Dat	Nom	Verb	Gen	N1	Gold
1	*companies*	Unternehmen<NN>	*NN*	0.00	0.00	1.00	erhalten	–	–	*Nom*
2	*should*	sollten<VVFIN>	*VVFIN*	–	–	–	–	–	–	–
3	*financial*	finanziell<ADJ>	*ADJ*	–	–	–	–	–	–	*Acc*
4	*funding*	Mittel<NN>	*NN*	1.00	0.00	0.00	erhalten	–	–	*Acc*
5	*for*	für APPR<Acc>	*PRP*	–	–	–	–	–	–	–
6	*the*	d<ART>	*ART*	–	–	–	–	–	–	*Acc*
7	*introduction*	Einführung<NN>	*NN*	–	–	–	–	–	–	*Acc*
8	*new*	neu<ADJ>	*ADJ*	–	–	–	–	–	–	*Gen*
9	*technologies*	Technologie<NN>	*NN*	–	–	–	–	100	Einführung<NN>	*Gen*
10	*obtain*	erhalten<VVINF>	*VVINF*	–	–	–	–	–	–	–

Table 4: Adding subcategorization information into SMT output. (EN input: *companies should obtain financial funding for the introduction of new technologies*). On the right, the correct labels are given.

4.2 Integrating subcategorization knowledge

There are two possibilities to integrate subcategorization information into the case prediction model: (i) It can be integrated into the data set using the tree-structure provided by the decoder. Here, verb-noun tuples are extracted from VP and S structures, and then the probabilities for the different functions are looked up. Similarly, for two adjacent NPs, the occurrence frequencies of the respective two nouns are looked up in the list of noun-noun$_{Gen}$ constructions. (ii) The subcategorization information can be integrated based on the verb-noun tuples obtained by using tuples obtained from source-side dependencies.

The classification task of the CRF consists in predicting a sequence of labels: case values for NPs/PPs or no value otherwise, cf. table 4. The model has access to the basic features *stem* and *tag*, as well as the new features based on subcategorizaion information (explained below), using unigrams within a window of up to four positions to the right and the left of the current position, as well as bigrams and trigrams for stems and tags (current item + left and/or right item).

An example for integrating subcategorization features is given in table 4. The first word *Unternehmen* (*companies*) is annotated as subject of *erhalten* (*obtain*) with probability 1, and *Mittel* (*funding*) is annotated as direct object of *erhalten* with probability 1. The word *Technologie* (*technology*) has been marked as a candidate for a genitive in a noun-noun$_{Gen}$ construction[8]; the co-occurrence frequency of the tuple *Einführung-Technologie* (*introduction - technology*) lies in the bucket 11...100.

In addition to the probability/frequency of the respective functions, we also provide the CRF with bigrams containing the two parts of the tuple,

[8]There is no annotation on *Einführung* as the preposition *für* is always in accusative case.

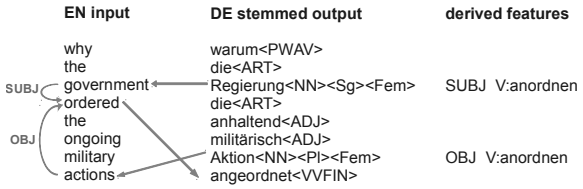

Figure 1: Deriving features from dependency-parsed English data via the word alignment.

i.e. verb+noun or the two nouns of possible noun-noun$_{Gen}$ constructions. As can be seen in the example in table 4, the subject (line 1) and the verb (line 10) are far apart from each other. By providing the parts of the tuple as unigrams, bigrams or trigrams to the CRF, all relevant information is available: verb, noun and the probabilities for the potential functions of the noun in the sentence. In addition to bridging the long distance between verbs and subcategorized nouns, a very common problem for German, this type of precise information also helps to close the gap between the well-formed training data and the broken SMT-output as it replaces to a certain extent the target-language context information (n-grams of stems or lemmas within a small window).

4.3 Integrating source-side features

For predicting case in SMT output, information about an NP's function in the input sentence is essential. Syntax-semantic functions can be isomorphic (e.g., English subjects and objects may have the same function in a German translation), but this is not necessarily the case. Despite this, an important advantage of integrating source-side features is that the well-formed source-side text can be reliably parsed, whereas SMT output is often disfluent and cannot be reliably parsed.

The English features are obtained from dependency-parsed data (Choi and Palmer, 2012). The relevant annotation of the parser is transferred

to the SMT output via word alignment. We focus on English subjects, direct objects and noun-of-noun structures (often equivalent to noun-noun$_{Gen}$ phrases on the German side): these structures are generally likely to correspond to each other within source and target language. In contrast to the subcategorization-based information, the difference between well-formed training data and disfluent SMT output tends to work to our benefit here: while the parallel sentences of the training data were manually translated with the objective to produce good target-language sentences, the syntactic structures of the source and target sentences are often diverging. In contrast, the SMT system often produces more isomorphic translations, which is helpful for annotating source-side features on the target language.

Figure 1 shows the process of integrating source-side features: for each German noun that is aligned with an English noun labelled as subject or direct object, this annotation is transferred to the target-side. Using the English dependency structures, the verb subcategorizing the respective noun is identified, and via the alignment, the equivalent German verb is obtained. Similarly, candidates for noun-noun$_{Gen}$ structures are identified by extracting and aligning English noun-of-noun phrases.

5 Experiments and evaluation

In this section, we present experiments using different feature combinations. We also present a manual evaluation of our best system which shows that the new features improve translation quality.

5.1 Data and experimental setup

We use the hierarchical translation system that comes with the Moses SMT-package and GIZA++ to compute the word alignment, using the "grow-diag-final-and" heuristics. The rule table was computed with the default parameter setting for GHKM extraction (Galley et al., 2004) in the implementation by Williams and Koehn (2012).

Our training data contains 1,485,059 parallel sentences[9]; the German part of the parallel data is used as the target-side language model. The dev and test sets (1025/1026 lines) are wmt-2009-a/b.

For predicting the grammatical features, we used the Wapiti Toolkit (Lavergne et al., 2010).[10] We train four CRFs on data prepared as shown in section 3. The corpora used for the extraction of subcategorization tuples were Europarl and German newspaper data (200 million words). We choose this particular data combination in order to provide data that matches the training data, as well as to add new data of the test set's domain (news). The German part of Europarl was dependency-parsed with Bohnet (2010), and subcategorization information was extracted as described in Scheible et al. (2013); the newspaper data (HGC - Huge German Corpus) was parsed with Schmid (2000), and subcategorization information was extracted as described in Schulte im Walde (2002b).

5.2 Results

We report results of two types of systems (table 5): first, a regular translation system built on surface forms (i.e., normal text) and second, four inflection prediction systems. The first inflection prediction system (1) uses a simple case prediction model, whereas the remaining systems are enriched with (2) subcategorization information (cf. section 4.2), (3) source-side features (cf. section 4.3), and (4) both source-side features and subcategorization information. In (2) and (4), the subcategorization information was included using tuples obtained from source-side dependencies[11]. The simple prediction system corresponds to that presented in section 3; for all inflection prediction systems, the same SMT output and models for number, gender and strong/weak inflection were used; thus the only difference with the simple prediction system is the model for case prediction.

We present three types of evaluation: BLEU scores (Papineni et al., 2001), prediction accuracy on clean data and a manual evaluation of the best system in section 5.3.

Table 5 gives results in case-insensitive BLEU. While the inflection prediction systems (1-4) are significantly[12] better than the surface-form system (0), the different versions of the inflection systems are not distinguishable in terms of BLEU; however, our manual evaluation shows that the new features have a positive impact on translation quality.

[9] English/German data released for the 2009 ACL Workshop on Machine Translation shared task.

[10] To eliminate irrelevant features, we use L1 regularization; the regularization parameter is optimized on held out data.

[11] Using tuples extracted from the target-side parse tree (produced by the decoder) results in a BLEU score of 14.00.

[12] We used Kevin Gimpel's implementation of pairwise bootstrap resampling with 1000 samples.

	0	1	2	3	4
	surface system	**simple prediction**	**subcat. features (tuples from EN side)**	**source-side features**	**source-side + subcat. featues**
BLEU	13.43	14.02	14.05	14.10	14.17
Clean	–	85.05 %	85.65 %	85.61 %	85.81 %

Table 5: Results of the simple prediction vs. three systems enriched with extra features.

One problem with using BLEU as an evaluation metric is that it is a precision-oriented metric and tends to reward fluency rather than adequacy (see (Wu and Fung, 2009a; Liu and Gildea, 2010)). As we are working on improving adequacy, this will not be fully reflected by BLEU. Furthermore, not all components of an NP do necessarily change their inflection with a new case value; it might happen that the only indicator for the case of an NP is the determiner: *er sieht [den alten Mann]$_{NPacc}$* (*he sees the old man*) vs. *er folgt [dem alten Mann]$_{NPdat}$* (*he follows the old man*). While the case marking of NPs is essential for comprehensibility, one changed word per noun phrase is hardly enough to be reflected by BLEU.

An alternative to study the effectiveness of the case prediction model is to evaluate the prediction accuracy on parsed clean data, i.e. not on SMT output. In this case, we measure (using the dev set) how often the case of an NP is predicted correctly[13]. In all cases, the prediction accuracy is better for the enriched systems. This shows that the additional features improve the model, but also that a gain in prediction accuracy on clean data is not necessarily related to a gain in BLEU. We observed that the more complex the model, the less robust it is to differences between the test data and the training data. Related to this problem, we observed that high-order n-gram POS/lemma-based features in the simple prediction (sequences of lemmas and tags) are given too much weight in training and thus make it difficult for the new features to have a larger impact, so we restricted the n-gram order of this type of feature to trigrams.

5.3 Manual evaluation of the best system

In order to provide a better understanding of the impact of the presented features, in particular to see whether there is an improvement in adequacy, we carried out a manual evaluation comparing system (4) with the simple prediction system (1). From the set of different sentences between the simple prediction system and the enriched system (144 of 1026), we evaluated those where the English input sentence was between 8 and 25 words long (46 sentences in total). We specifically restricted the test set in order to provide sentences which are less difficult to annotate, as longer sentences are often very disfluent and too hard to rate. Most of the sentences in the evaluation set differ only in the realization of one NP. For comparing the two systems, the sentences were presented in random order to 3 native speakers of German.

		enriched preferred	**simple preferred**	**equal**
(a)	**person 1**	23	11	12
	person 2	21	8	17
	person 3	26	11	9
(b)	**person 1**	23	5	18
	person 2	21	11	14
	person 3	29	8	9
(c)	**agreement**	17	2	6

Table 6: Manual evaluation of 46 sentences: without (a) and with (b) access to EN input, and the annotators' agreement in the second part (c).

The evaluation consists of two parts: first, the participants were asked to decide which sentence is better without being given the English input (this measures fluency). In the second part, they should to mark that sentence which better reproduces the content of the English input sentence (this measures adequacy). The test set is the same for both tasks, the only difference being that the English input is given in the second part. The results are given in table 6. Summarizing we can say that the participants prefer the enriched system over the simple system in both parts; there is a high agreement (17 cases) in decisions over those sentences which were rated as *enriched better*.

When looking at the pairwise inter-annotator agreement for the task of annotating the test-set with the 3 possible labels *enriched preferred*, *simple preferred* and *no preference*, we find that the annotators P1 and P2 have a substantial agreement

[13]The numbers in table 5 are artificially high and downplay the difference as they also include cases which are very easy to predict, such as nouns in PPs where only one value for case is possible. We measure how many case labels were correctly predicted, not correct inflected forms.

1	**input**	hundreds of policemen were on alert , and [a helicopter]$_{Subj}$ circled the area with searchlights .
	simple	Hunderte von Polizisten auf Trab , und [einen Helikopter]$_{Acc}$ eingekreist das Gebiet mit searchlights .
	enriched	Hunderte von Polizisten auf Trab , und [ein Helikopter]$_{Nom}$ eingekreist das Gebiet mit searchlights .
2	**input**	while 38 %percent put [their trust]$_{Obj}$ in viktor orbán .
	simple	während 38 % [ihres Vertrauens]$_{Gen}$ schenken in Viktor Orbán .
	enriched	während 38 % [ihr Vertrauen]$_{Acc}$ schenken in Viktor Orbán .
3	**input**	more than $ 100 billion will enter [the monetary markets]$_{Obj}$ by means of public sales .
	simple	mehr als 100 Milliarden Dollar werden durch öffentlichen Verkauf [der Geldmärkte]$_{Gen}$ treten .
	enriched	mehr als 100 Milliarden Dollar werden durch öffentlichen Verkauf [die Geldmärkte]$_{Acc}$ treten .

Table 7: Output from the simple system (1) and the enriched system (4).

in terms of Kappa ($\kappa = 0.6184$), whereas the agreement of P3 with P1/P2 respectively leads to lower scores ($\kappa = 0.4467$ and $\kappa = 0.3596$). However, the annotators tend to agree well on sentences with the label *enriched preferred*, but largely disagree on sentences labelled as either *simple preferred* or *no preference*. The number of decisions where all three annotators agree on a label when given the English input is listed in table 6(c): for example, only two sentences were given the label *baseline is better* by all three annotators. This outcome shows how difficult it is to rate disfluent SMT output. For evaluating the case prediction system, the distinction between *enriched preferred* and *enriched dispreferred* is the most important question to answer. Redefining the annotation task to annotating only two values by grouping the labels *simple preferred* and *no preference* into one annotation possibility leads to $\kappa = 0.7391$, $\kappa = 0.4048$ and $\kappa = 0.5652$.

5.4 Examples

Table 7 shows some examples for output from the simple system and the system using source-side and subcategorization features. In the first sentence, the subject NP *a helicopter* was inflected as a direct object in the simple system, but as a subject in the enriched system, which was preferred by all three annotators. In the second sentence, the NP *their trust*, i.e. a direct object of *put*, was incorrectly predicted as genitive-modifier of *38 %* (i.e. *38 % of their trust*) in the simple system. The enriched system made use of the preference for accusative for the pair *Vertrauen schenken* (*place trust*), correctly inflecting this NP as direct object. Interestingly, only two annotators preferred the enriched system, whereas one was undecided. The third sentence illustrates how difficult it is to rate case marking on disfluent SMT output: there are two possibilities to translate *enter the money market*; the direct equivalent of the English phrase (*den Geldmarkt*$_{Acc}$ *betreten*), or via the use of a prepositional phrase (*auf den Geldmarkt*$_{Acc}$ *treten*: "*to step into the money market*"). The SMT-output contains a mix of both, i.e. the verb *treten* (instead of *betreten*), but without the preposition, which cannot lead to a fully correct inflection. While the inflection of the simple system (a genitive construction meaning *the public sales of the money market*) is definitely wrong, the inflection obtained in the enriched system is not useful either, due to the structure of the translation[14]. This difficulty is also reflected by the annotators, who gave twice the label *no preference* and once the label *enriched better*.

6 Conclusion

We illustrated the necessity of using external knowledge sources like subcategorization information for modelling case for English to German translation. We presented a translation system making use of a subcategorization database together with source-side features. Our method is language-independent with regard to the source language; furthermore, no language-specific high-quality semantic annotation is needed for the target language, but the data required to model the subcategorization preferences can be obtained using standard NLP techniques. We showed in a manual evaluation that the proposed features have a positive impact on translation quality.

Acknowledgements

This work was funded by the DFG Research Project *Distributional Approaches to Semantic Relatedness* (Marion Weller), the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde), as well as by the Deutsche Forschungsgemeinschaft grant *Models of Morphosyntax for Statistical Machine Translation* (Alexander Fraser).

[14]Furthermore, with *treten* being polysemous, *die Geldmärkte treten* can also mean *to kick the money markets*.

References

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING) 2010*, pages 89–97, Beijing, August.

Ted Briscoe and John Carroll. 1997. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC.

John Carroll and Alex C. Fang. 2004. The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, pages 107–114, Sanya City, China.

John Carroll, Guido Minnen, and Ted Briscoe. 1998. Can Subcategorisation Probabilities Help a Statistical Parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*, Montreal, Canada.

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*.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

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.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a Translation Rule? In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL)*.

Spence Green and John DeNero. 2012. A Class-Based Agreement Model for Generating Accurately Inflected Translations. pages 146–155.

Bryant Huang and Kevin Knight. 2006. Relabeling Syntax Trees to Improve Syntax-Based Machine Translation Quality. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*.

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

Ahmed El Kholy and Nizar Habash. 2012. Translate, Predict or Generate: Modeling Rich Morphology in Statistical Machine Translation. In *European Association for Machine Translation*.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, Computer and Information Science.

Katrin Kirchhoff, Daniel Capurro, and Anne Turner. 2012. Evaluating User Preferences in Machine Translation Using Conjoint Analysis. In *European Association for Machine Translation*.

Upali S. Kohomban and Wee Sun Lee. 2005. Learning Semantic Classes for Word Sense Disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 34–41, Ann Arbor, MI.

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)*, pages 504–513. Association for Computational Linguistics, July.

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press.

Ding Liu and Daniel Gildea. 2008. Improved Tree-to-String Transducers for Machine Translation. In *ACL Workshop on Statistical Machine Translation*.

Ding Liu and Daniel Gildea. 2010. Semantic Role Features for Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING) 2010*.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised Acquisition of Predominant Word Senses. *Computational Linguistics*, 33(4):553–590.

Cédric Messiant. 2008. A Subcategorization Acquisition System for French Verbs. In *Proceedings of the Student Research Workshop at the 46th Annual Meeting of the Association for Computational Linguistics*, pages 55–60, Columbus, OH.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated Resource of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

Anoop Sarkar and Daniel Zeman. 2000. Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Germany.

Silke Scheible, Sabine Schulte im Walde, Marion Weller, and Max Kisselew. 2013. A Compact but Linguistically Detailed Database for German Verb Subcategorisation relying on Dependency Parses from a Web Corpus. In *Proceedings of the 8th Web as Corpus Workshop*, Lancaster, UK. To appear.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: a German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*.

Helmut Schmid. 2000. LoPar: Design and Implementation. Arbeitspapiere des Sonderforschungsbereichs 340 'Linguistic Theory and the Foundations of Computational Linguistics' 149, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors.

Sabine Schulte im Walde. 2002a. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas de Gran Canaria, Spain.

Sabine Schulte im Walde. 2002b. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas de Gran Canaria, Spain.

Sara Stymne and Nicola Cancedda. 2011. Productive Generation of Compound Words in Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Machine Translation*.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15, Sapporo, Japan.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL): Human Language Technologies*.

Giulia Venturi1, Simonetta Montemagni, Simone Marchi, Yutaka Sasaki, Paul Thompson, John McNaught, and Sophia Ananiadou. 2009. Bootstrapping a Verb Lexicon for Biomedical Information Extraction. In Alexander Gelbukh, editor, *Linguistics and Intelligent Text Processing*, pages 137–148. Springer, Heidelberg.

Philip Williams and Phillipp Koehn. 2012. GHKM-Rule Extraction and Scope-3 Parsing in Moses. In *Proceedings of the 7th Workshop on Statistical Machine Translation, ACL*.

Dekai Wu and Pascale Fung. 2009a. Can Semantic Role Labeling Improve SMT? In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT)*.

Dekai Wu and Pascale Fung. 2009b. Semantic Roles for SMT: A Hybrid two-pass Model. In *Proceedings of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies Conference (NAACL-HLT)*.