Modeling Inflection and Word-Formation in SMT

Alexander Fraser[*] **Marion Weller**[*] **Aoife Cahill**[†] **Fabienne Cap**[*]

[*]Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
D–70174 Stuttgart, Germany
{fraser,wellermn,cap}@ims.uni-stuttgart.de

[†]Educational Testing Service
Princeton, NJ 08541
USA
acahill@ets.org

Abstract

The current state-of-the-art in statistical machine translation (SMT) suffers from issues of sparsity and inadequate modeling power when translating into morphologically rich languages. We model both inflection and word-formation for the task of translating into German. We translate from English words to an underspecified German representation and then use linear-chain CRFs to predict the fully specified German representation. We show that improved modeling of inflection and word-formation leads to improved SMT.

1 Introduction

Phrase-based statistical machine translation (SMT) suffers from problems of data sparsity with respect to inflection and word-formation which are particularly strong when translating to a morphologically rich target language, such as German. We address the problem of inflection by first translating to a stem-based representation, and then using a second process to inflect these stems. We study several models for doing this, including: strongly lexicalized models, unlexicalized models using linguistic features, and models combining the strengths of both of these approaches. We address the problem of word-formation for compounds in German, by translating from English into German word parts, and then determining whether to merge these parts to form compounds.

We make the following new contributions: (i) we introduce the first SMT system combining inflection prediction with synthesis of portmanteaus and compounds. (ii) For inflection, we compare the mostly unlexicalized prediction of linguistic features (with a subsequent surface form generation step) versus the direct prediction of surface forms, and show that both approaches have complementary strengths. (iii) We combine the advantages of the prediction of linguistic features with the prediction of surface forms. We implement this in a CRF framework which improves on a standard phrase-based SMT baseline. (iv) We develop separate (but related) procedures for inflection prediction and dealing with word-formation (compounds and portmanteaus), in contrast with most previous work which usually either approaches both problems as inflectional problems, or approaches both problems as word-formation problems.

We evaluate on the end-to-end SMT task of translating from English to German of the 2009 ACL workshop on SMT. We achieve BLEU score increases on both the test set and the blind test set.

2 Overview of the translation process for inflection prediction

The work we describe is focused on generalizing phrase-based statistical machine translation to better model German NPs and PPs. We particularly want to ensure that we can generate novel German NPs, where what we mean by novel is that the (inflected) realization is not present in the parallel German training data used to build the SMT system, and hence cannot be produced by our baseline (a standard phrase-based SMT system). We first present our system for dealing with the difficult problem of inflection in German, including the inflection-dependent phenomenon of portmanteaus. Later, after performing an extensive analysis of this system, we will extend it to model compounds, a highly productive phenomenon in German (see Section 8).

The key linguistic knowledge sources that we use are morphological analysis and generation of German based on SMOR, a morphological analyzer/generator of German (Schmid et al., 2004) and the BitPar parser, which is a state-of-the-art parser of German (Schmid, 2004).

2.1 Issues of inflection prediction

In order to ensure coherent German NPs, we model linguistic features of each word in an NP. We model *case*, *gender*, and *number* agreement and whether or not the word is in the scope of a determiner (such as a definite article), which we label *in-weak-context* (this linguistic feature is necessary to determine the *type of inflection* of adjectives and other words: *strong, weak, mixed*). This is a diverse group of features. The *number* of a German noun can often be determined given only the English source word. The *gender* of a German noun is innate and often difficult to determine given only the English source word. *Case* is a function of the slot in the subcategorization frame of the verb (or preposition). There is agreement in all of these features in an NP. For instance the *number* of an article or adjective is determined by the head noun, while the *type of inflection* of an adjective is determined by the choice of article.

We can have a large number of surface forms. For instance, English *blue* can be translated as German *blau*, *blaue*, *blauer*, *blaues*, *blauen*. We predict which form is correct given the context. Our system can generate forms not seen in the training data. We follow a two-step process: in step-1 we translate to *blau* (the stem), in step-2 we predict features and generate the inflected form.[1]

2.2 Procedure

We begin building an SMT system by parsing the German training data with BitPar. We then extract morphological features from the parse. Next, we lookup the surface forms in the SMOR morphological analyzer. We use the morphological features in the parse to disambiguate the set of possible SMOR analyses. Finally, we output the "stems" of the German text, with the addition of markup taken from the parse (discussed in Section 2.3).

We then build a standard Moses system translating from English to German stems. We obtain a sequence of stems and POS[2] from this system, and then predict the correct inflection using a sequence model. Finally we generate surface forms.

2.3 German Stem Markup

The translation process consists of two major steps. The first step is translation of English words to German stems, which are enriched with some inflectional markup. The second step is the full inflection of these stems (plus markup) to obtain the final sequence of inflected words. The purpose of the additional German inflectional markup is to strongly improve prediction of inflection in the second step through the addition of markup to the stems in the first step.

In general, all features to be predicted are stripped from the stemmed representation because they are subject to agreement restrictions of a noun or prepositional phrase (such as *case* of nouns or all features of adjectives). However, we need to keep all morphological features that are not dependent on, and thus not predictable from, the (German) context. They will serve as known input for the inflection prediction model. We now describe this markup in detail.

Nouns are marked with *gender* and *number*: we consider the *gender* of a noun as part of its stem, whereas *number* is a feature which we can obtain from English nouns.

Personal pronouns have number and gender annotation, and are additionally marked with *nominative* and *not-nominative*, because English pronouns are marked for this (except for *you*).

Prepositions are marked with the *case* their object takes: this moves some of the difficulty in predicting *case* from the inflection prediction step to the stem translation step. Since the choice of case in a PP is often determined by the PP's meaning (and there are often different meanings possible given different case choices), it seems reasonable to make this decision during stem translation.

Verbs are represented using their inflected surface form. Having access to inflected verb forms has a positive influence on case prediction in the second

[1] E.g., *case*=nominative, *gender*=masculine, *number*=singular, *in-weak-context*=true; inflected: *blaue*.

[2] We use an additional target factor to obtain the coarse POS for each stem, applying a 7-gram POS model. Koehn and Hoang (2007) showed that the use of a POS factor only results in negligible BLEU improvements, but we need access to the POS in our inflection prediction models.

input	decoder output	inflected	merged
in	in<APPR><Dat>	in	**im**
	die<+ART><Def>	dem	
contrast	Gegensatz<+NN><Masc><Sg>	Gegensatz	Gegensatz
to	zu<APPR><Dat>	zu	**zur**
the	die<+ART><Def>	der	
animated	lebhaft<+ADJ><Pos>	lebhaften	lebhaften
debate	Debatte<+NN><Fem><Sg>	Debatte	Debatte

Table 1: Re-merging of prepositions and articles after inflection to form portmanteaus, *in dem* means *in the*.

step through subject-verb agreement.

Articles are reduced to their stems (the stem itself makes clear the definite or indefinite distinction, but lemmatizing involves removing markings of *case*, *gender* and *number* features).

Other words are also represented by their stems (except for words not covered by SMOR, where surface forms are used instead).

3 Portmanteaus

Portmanteaus are a word-formation phenomenon dependent on inflection. As we have discussed, standard phrase-based systems have problems with picking a definite article with the correct case, gender and number (typically due to sparsity in the language model, e.g., a noun which was never before seen in dative case will often not receive the correct article). In German, portmanteaus increase this sparsity further, as they are compounds of prepositions and articles which must agree with a noun.

We adopt the linguistically strict definition of the term portmanteau: the merging of two function words.[3] We treat this phenomena by splitting the component parts during training and re-merging during generation. Specifically for German, this requires splitting the words which have German POS tag APPRART into an APPR (preposition) and an ART (article). Merging is restricted, the article must be *definite*, *singular*[4] and the preposition can only take *accusative* or *dative* case. Some prepositions allow for merging with an article only for certain noun genders, for example the preposition in_{Dative} is only merged with the following article if the following noun is of masculine or neuter gender. The definite article must be inflected before making a decision about whether to merge a preposition and the article into a portmanteau. See Table 1 for examples.

4 Models for Inflection Prediction

We present 5 procedures for inflectional prediction using supervised sequence models. The first two procedures use simple N-gram models over fully inflected surface forms.

1. Surface with no features is presented with an underspecified input (a sequence of stems), and returns the most likely inflected sequence.

2. Surface with case, number, gender is a hybrid system giving the surface model access to linguistic features. In this system prepositions have additionally been labeled with the case they mark (in both the underspecified input and the fully specified output the sequence model is built on) and gender and number markup is also available.

The rest of the procedures predict morphological features (which are input to a morphological generator) rather than surface words. We have developed a two-stage process for predicting fully inflected surface forms. The first stage takes a stem and predicts morphological features for that stem, based on the surrounding context. The aim of the first stage is to take a stem and predict four morphological features: *case*, *gender*, *number* and *type of inflection*. We experiment with a number of models for doing this. The second stage takes the stems marked with morphological features (predicted in the first stage) and uses a morphological generator to generate the full surface form. For the second stage, a modified version of SMOR (Schmid et al., 2004) is used, which, given a stem annotated with morphological features, generates exactly one surface form.

We now introduce our first linguistic feature prediction systems, which we call joint sequence models (JSMs). These are standard language models, where the "word" tokens are not represented as surface forms, but instead using POS and features. In testing, we supply the input as a sequence in underspecified form, where some of the features are specified in the stem markup (for instance, POS=Noun, *gender*=masculine, *number*=plural), and then use Viterbi search to find the most probable fully specified form (for instance, POS=Noun, *gender*=masculine, *number*=plural,

[3] Some examples are: *zum* (to the) = *zu* (to) + *dem* (the) [German], *du* (from the) = *de* (from) + *le* (the) [French] or *al* (to the) = *a* (to) + *el* (the) [Spanish].

[4] This is the reason for which the preposition + article in Table 2 remain unmerged.

output decoder	input prediction	output prediction	inflected forms	gloss
`haben<VAFIN>`	haben-V	haben-V	haben	*have*
`Zugang<+NN><Masc><Sg>`	NN-Sg-Masc	NN-Masc.Acc.Sg.in-weak-context=false	Zugang	*access*
`zu<APPR><Dat>`	APPR-zu-Dat	APPR-zu-Dat	zu	*to*
`die<+ART><Def>`	ART-in-weak-context=true	ART-Neut.Dat.Pl.in-weak-context=true	den	*the*
`betreffend<+ADJ><Pos>`	ADJA	ADJA-Neut.Dat.Pl.in-weak-context=true	betreffenden	*respective*
`Land<+NN><Neut><Pl>`	NN-Pl-Neut	NN-Neut.Dat.Pl.in-weak-context=true	Ländern	*countries*

Table 2: Overview: inflection prediction steps using a single joint sequence model. All words except verbs and prepositions are replaced by their POS tags in the input. Verbs are inflected in the input ("haben", meaning "have" as in "they have", in the example). Prepositions are lexicalized ("zu" in the example) and indicate which *case* value they mark ("Dat", i.e., Dative in the example).

case=nominative, *in-weak-context*=true).[5]

3. Single joint sequence model on features. We illustrate the different stages of the inflection prediction when using a joint sequence model. The stemmed input sequence (cf. Section 2.3) contains several features that will be part of the input to the inflection prediction. With the exception of verbs and prepositions, the representation for feature prediction is based on POS-tags.

As *gender* and *number* are given by the heads of noun phrases and prepositional phrases, and the expected *type of inflection* is set by articles, the model has sufficient information to compute values for these features and there is no need to know the actual words. In contrast, the prediction of *case* is more difficult as it largely depends on the content of the sentence (e.g. which phrase is object, which phrase is subject). Assuming that verbs and prepositions indicate subcategorization frames, the model is provided crucial information for the prediction of case by keeping verbs (recall that verbs are produced by the stem translation system in their inflected form) and prepositions (the prepositions also have case markup) instead of replacing them with their tags.

After having predicted a single label with values for all features, an inflected word form for the stem and the features is generated. The prediction steps are illustrated in Table 2.

4. Using four joint sequence models (one for each linguistic feature). Here the four linguistic feature values are predicted separately. The assumption that the different linguistic features can be predicted independently of one another is a reasonable linguistic assumption to make given the additional German markup that we use. By splitting the inflection prediction problem into 4 component parts, we end up with 4 simpler models which are less sensitive to data sparseness.

Each linguistic feature is modeled independently (by a JSM) and has a different input representation based on the previously described markup. The input consists of a sequence of coarse POS tags, and for those stems that are marked up with the relevant feature, this feature value. Finally, we combine the predicted features together to produce the same final output as the single joint sequence model, and then generate each surface form using SMOR.

5. Using four CRFs (one for each linguistic feature). The sequence models already presented are limited to the *n*-gram feature space, and those that predict linguistic features are not strongly lexicalized. Toutanova et al. (2008) uses an MEMM which allows the integration of a wide variety of feature functions. We also wanted to experiment with additional feature functions, and so we train 4 separate linear chain CRF[6] models on our data (one for each linguistic feature we want to predict). We chose CRFs over MEMMs to avoid the label bias problem (Lafferty et al., 2001).

The CRF feature functions, for each German word w_i, are in Table 3. The common feature functions are used in all models, while each of the 4 separate models (one for each linguistic feature) includes the context of only that linguistic feature. We use $L1$ regularization to eliminate irrelevant feature functions, the regularization parameter is optimized on held out data.

[5] Joint sequence models are a particularly simple HMM. Unlike the HMMs used for POS-tagging, an HMM as used here only has a single emission possibility for each state, with probability 1. The states in the HMM are the fully specified representation. The emissions of the HMM are the stems+markup (the underspecified representation).

[6] We use the Wapiti Toolkit (Lavergne et al., 2010) on 4 x 12-Core Opteron 6176 2.3 GHz with 256GB RAM to train our CRF models. Training a single CRF model on our data was not tractable, so we use one for each linguistic feature.

Common	lemma$_{w_{i-5}\ldots w_{i+5}}$, tag$_{w_{i-7}\ldots w_{i+7}}$
Case	case$_{w_{i-5}\ldots w_{i+5}}$
Gender	gender$_{w_{i-5}\ldots w_{i+5}}$
Number	number$_{w_{i-5}\ldots w_{i+5}}$
in-weak-context	in-weak-context$_{w_{i-5}\ldots w_{i+5}}$

Table 3: Feature functions used in CRF models (feature functions are binary indicators of the pattern).

5 Experimental Setup

To evaluate our end-to-end system, we perform the well-studied task of news translation, using the Moses SMT package. We use the English/German data released for the 2009 ACL Workshop on Machine Translation shared task on translation.[7] There are 82,740 parallel sentences from news-commentary09.de-en and 1,418,115 parallel sentences from europarl-v4.de-en. The monolingual data contains 9.8 M sentences.[8]

To build the baseline, the data was tokenized using the Moses tokenizer and lowercased. We use GIZA++ to generate alignments, by running 5 iterations of Model 1, 5 iterations of the HMM Model, and 4 iterations of Model 4. We symmetrize using the "grow-diag-final-and" heuristic. Our Moses systems use default settings. The LM uses the monolingual data and is trained as a five-gram[9] using the SRILM-Toolkit (Stolcke, 2002). We run MERT separately for each system. The recaser used is the same for all systems. It is the standard recaser supplied with Moses, trained on all German training data. The dev set is wmt-2009-a and the test set is wmt-2009-b, and we report end-to-end case sensitive BLEU scores against the unmodified reference SGML file. The blind test set used is wmt-2009-blind (all lines).

In developing our inflection prediction systems (and making such decisions as n-gram order used), we worked on the so-called "clean data" task, predicting the inflection on stemmed reference sentences (rather than MT output). We used the 2000 sentence dev-2006 corpus for this task.

Our contrastive systems consist of two steps, the first is a translation step using a similar Moses system (except that the German side is stemmed, with the markup indicated in Section 2.3), and the second is inflection prediction as described previously in the paper. To derive the stem+markup representation we first parse the German training data and then produce the stemmed representation. We then build a system for translating from English words to German stems (the stem+markup representation), on the same data (so the German side of the parallel data, and the German language modeling uses the stem+markup representation). Likewise, MERT is performed using references which are in the stem+markup representation.

To train the inflection prediction systems, we use the monolingual data. The basic surface form model is trained on lowercased surface forms, the hybrid surface form model with features is trained on lowercased surface forms annotated with markup. The linguistic feature prediction systems are trained on the monolingual data processed as described previously (see Table 2).

Our JSMs are trained using the SRILM Toolkit. We use the SRILM disambig tool for predicting inflection, which takes a "map" that specifies the set of fully specified representations that each underspecified stem can map to. For surface form models, it specifies the mapping from stems to lowercased surface forms (or surface forms with markup for the hybrid surface model).

6 Results for Inflection Prediction

We build two different kinds of translation system, the baseline and the stem translation system (where MERT is used to train the system to produce a stem+markup sequence which agrees with the stemmed reference of the dev set). In this section we present the end-to-end translation results for the different inflection prediction models defined in Section 4, see Table 4.

If we translate from English into a stemmed German representation and then apply a unigram stem-to-surface-form model to predict the surface form, we achieve a BLEU score of 9.97 (line 2). This is only presented for comparison.

The baseline[10] is 14.16, line 1. We compare this with a 5-gram sequence model[11] that predicts

[7] http://www.statmt.org/wmt09/translation-task.html

[8] However, we reduced the monolingual data (only) by retaining only one copy of each unique line, which resulted in 7.55 M sentences.

[9] Add-1 smoothing for unigrams and Kneser-Ney smoothing for higher order n-grams, pruning defaults.

[10] This is a better case-sensitive score than the baselines on wmt-2009-b in experiments by top-performers Edinburgh and Karlsruhe at the shared task. We use Moses with default settings.

[11] Note that we use a different set, the "clean data" set, to determine the choice of n-gram order, see Section 7. We use

surface forms without access to morphological features, resulting in a BLEU score of 14.26. Introducing morphological features (*case* on prepositions, *number* and *gender* on nouns) increases the BLEU score to 14.58, which is in the same range as the single JSM system predicting all linguistic features at once.

This result shows that the mostly unlexicalized single JSM can produce competitive results with direct surface form prediction, despite not having access to a model of inflected forms, which is the desired final output. This strongly suggests that the prediction of morphological features can be used to achieve additional generalization over direct surface form prediction. When comparing the simple direct surface form prediction (line 3) with the hybrid system enriched with *number*, *gender* and *case* (line 4), it becomes evident that feature markup can also aid surface form prediction.

Since the single JSM has no access to lexical information, we used a language model to score different feature predictions: for each sentence of the development set, the 100 best feature predictions were inflected and scored with a language model. We then optimized weights for the two scores LM (language model on surface forms) and FP (feature prediction, the score assigned by the JSM). This method disprefers feature predictions with a top FP-score if the inflected sentence obtains a bad LM score and likewise disfavors low-ranked feature prediction with a high LM score. The prediction of *case* is the most difficult given no lexical information, thus scoring different prediction possibilities on inflected words is helpful. An example is when the *case* of a noun phrase leads to an inflected phrase which never occurs in the (inflected) language model (e.g., *case*=genitive vs. *case*=other). Applying this method to the single JSM leads to a negligible improvement (14.53 vs. 14.56). Using the n-best output of the stem translation system did not lead to any improvement.

The comparison between different feature prediction models is also illustrative. Performance decreases somewhat when using individual joint sequence models (one for each linguistic feature) compared to one single model (14.29, line 6).

The framework using the individual CRFs for each linguistic feature performs best (14.72, line 7). The CRF framework combines the advantages of surface form prediction and linguistic feature prediction by using feature functions that effectively cover the feature function spaces used by both forms of prediction. The performance of the CRF models results in a statistically significant improvement[12] ($p < 0.05$) over the baseline. We also tried CRFs with bilingual features (projected from English parses via the alignment output by Moses), but obtained only a small improvement of 0.03, probably because the required information is transferred in our stem markup (also a poor improvement beyond monolingual features is consistent with previous work, see Section 8.3). Details are omitted due to space.

1	baseline	14.16
2	unigram surface (no features)	9.97
3	surface (no features)	14.26
4	surface (with case, number, gender features)	14.58
5	1 JSM morphological features	14.53
6	4 JSMs morphological features	14.29
7	4 CRFs morphological features, lexical information	14.72

Table 4: BLEU scores (detokenized, case sensitive) on the development test set wmt-2009-b

We further validated our results by translating the blind test set from wmt-2009, which we have never looked at in any way. Here we also had a statistically significant difference between the baseline and the CRF-based prediction, the scores were 13.68 and 14.18.

7 Analysis of Inflection-based System

Stem Markup. The first step of translating from English to German stems (with the markup we previously discussed) is substantially easier than translating directly to inflected German (we see BLEU scores on stems+markup that are over 2.0 BLEU higher than the BLEU scores on inflected forms when running MERT). The addition of case to prepositions only lowered the BLEU score reached by MERT by about 0.2, but is very helpful for prediction of the case feature.

Inflection Prediction Task. Clean data task results[13] are given in Table 5. The 4 CRFs outperform the 4 JSMs by more than 2%.

a 5-gram for surface forms and a 4-gram for JSMs, and the same smoothing (Kneser-Ney, add-1 for unigrams, default pruning).

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

[13] 26,061 of 55,057 tokens in our test set are ambiguous. We report % surface form matches for ambiguous tokens.

Model	Accuracy
unigram surface (no features)	55.98
surface (no features)	86.65
surface (with case, number, gender features)	91.24
1 JSM morphological features	92.45
4 JSMs morphological features	92.01
4 CRFs morphological features, lexical information	94.29

Table 5: Comparing predicting surface forms directly with predicting morphological features.

training data	1 model	4 models
7.3 M sentences	92.41	91.88
1.5 M sentences	92.45	92.01
100000 sentences	90.20	90.64
1000 sentences	83.72	86.94

Table 6: Accuracy for different training data sizes of the single and the four separate joint sequence models.

As we mentioned in Section 4, there is a sparsity issue at small training data sizes for the single joint sequence model. This is shown in Table 6. At the largest training data sizes, modeling all 4 features together results in the best predictions of inflection. However using 4 separate models is worth this minimal decrease in performance, since it facilitates experimentation with the CRF framework for which the training of a single model is not currently tractable.

Overall, the inflection prediction works well for *gender*, *number* and *type of inflection*, which are local features to the NP that normally agree with the explicit markup output by the stem translation system (for example, the *gender* of a common noun, which is marked in the stem markup, is usually successfully propagated to the rest of the NP). Prediction of *case* does not always work well, and could maybe be improved through hierarchical labeled-syntax stem translation.

Portmanteaus. An example of where the system is improved because of the new handling of portmanteaus can be seen in the dative phrase *im internationalen Rampenlicht* (in the international spotlight), which does not occur in the parallel data. The accusative phrase *in das internationale Rampenlicht* does occur, however in this case there is no portmanteau, but a one-to-one mapping between *in the* and *in das*. For a given context, only one of accusative or dative case is valid, and a strongly disfluent sentence results from the incorrect choice. In our system, these two cases are handled in the same way (*def-article international Rampenlicht*). This allows us to generalize from the accusative example with no portmanteau and take advantage of longer phrase pairs, even when translating to something that will be inflected as dative and should be realized as a portmanteau. The baseline does not have this capability. It should be noted that the portmanteau merging method described in Section 3 remerges all occurrences of APPR and ART that can technically form a portmanteau. There are a few cases where merging, despite being grammatical, does not lead to a good result. Such exceptions require semantic interpretation and are difficult to capture with a fixed set of rules.

8 Adding Compounds to the System

Compounds are highly productive in German and lead to data sparsity. We split the German compounds in the training data, so that our stem translation system can now work with the individual words in the compounds. After we have translated to a split/stemmed representation, we determine whether to merge words together to form a compound. Then we merge them to create stems in the same representation as before and we perform inflection and portmanteau merging exactly as previously discussed.

8.1 Details of Splitting Process

We prepare the training data by splitting compounds in two steps, following the technique of Fritzinger and Fraser (2010). First, possible split points are extracted using SMOR, and second, the best split points are selected using the geometric mean of word part frequencies.

compound	word parts	gloss
Inflationsrate	Inflation Rate	inflation rate
auszubrechen	aus zu brechen	out to break (to break out)

Training data is then stemmed as described in Section 2.3. The formerly modifying words of the compound (in our example the words to the left of the rightmost word) do not have a stem markup assigned, except for two cases: i) they are nouns themselves or ii) they are particles separated from a verb. In these cases, former modifiers are represented identically to their individual occurring counterparts, which helps generalization.

8.2 Model for Compound Merging

After translation, compound parts have to be resynthesized into compounds before inflection. Two decisions have to be taken: i) where to merge and ii) how to merge. Following the work of Stymne and Cancedda (2011), we implement a linear-chain CRF merging system using the following features: stemmed (separated) surface form, part-of-speech[14] and frequencies from the training corpus for bigrams/merging of *word* and *word+1*, *word* as true prefix, *word+1* as true suffix, plus frequency comparisons of these. The CRF is trained on the split monolingual data. It only proposes merging decisions, merging itself uses a list extracted from the monolingual data (Popovic et al., 2006).

8.3 Experiments

We evaluated the end-to-end inflection system with the addition of compounds.[15] As in the inflection experiments described in Section 5, we use a 5-gram surface LM and a 7-gram POS LM, but for this experiment, they are trained on stemmed, split data. The POS LM helps compound parts and heads appear in correct order. The results are in Table 7. The BLEU score of the CRF on test is 14.04, which is low. However the system produces 19 compound types which are in the reference but not in the parallel data, and therefore not accessible to other systems. We also observe many more compounds in general. The 100-best inflection rescoring technique previously discussed reached 14.07 on the test set. Blind test results with CRF prediction are much better, 14.08, which is a statistically significant improvement over the baseline (13.68) and approaches the result we obtained without compounds (14.18). Correctly generated compounds are single words which usually carry the same information as multiple words in English, and are hence likely underweighted by BLEU. We again see many interesting generalizations. For instance, take the case of translating English *miniature cameras* to the German compound *Miniaturkameras*. *miniature camera* or *miniature cameras* does not occur in the training data, and so there is no appropriate phrase pair in any system (baseline, inflection, or inflection&compound-splitting). However, our system with compound splitting has learned from split composita that English *miniature* can be translated as German *Miniatur-* and gets the correct output.

1	1 JSM morphological features	13.94
2	4 CRFs morphological features, lexical information	14.04

Table 7: Results with Compounds on the test set

9 Related Work

There has been a large amount of work on translating from a morphologically rich language to English, we omit a literature review here due to space considerations. Our work is in the opposite direction, which primarily involves problems of generation, rather than problems of analysis.

The idea of translating to stems and then inflecting is not novel. We adapted the work of Toutanova et al. (2008), which is effective but limited by the conflation of two separate issues: word formation and inflection.

Given a stem such as *brother*, Toutanova et. al's system might generate the "stem and inflection" corresponding to *and his brother*. Viewing *and* and *his* as inflection is problematic since a mapping from the English phrase *and his brother* to the Arabic stem for *brother* is required. The situation is worse if there are English words (e.g., adjectives) separating *his* and *brother*. This required mapping is a significant problem for generalization. We view this issue as a different sort of problem entirely, one of word-formation (rather than inflection). We apply a "split in preprocessing and resynthesize in postprocessing" approach to these phenomena, combined with inflection prediction that is similar to that of Toutanova et. al. The only work that we are aware of which deals with both issues is the work of de Gispert and Mariño (2008), which deals with verbal morphology and attached pronouns. There has been other work on solving inflection. Koehn and Hoang (2007) introduced factored SMT. We use more complex context features. Fraser (2009) tried to solve the inflection prediction problem by simply building an SMT system for translating from stems to inflected forms. Bojar and Kos (2010) improved on this by marking prepositions with the case they mark (one of the most important markups in our system). Both efforts were ineffective on large data sets. Williams and Koehn (2011) used unification in an SMT system to model some of the

[14] Compound modifiers get assigned a special tag based on the POS of their former heads, e.g., *Inflation* in the example is marked as a non-head of a noun.

[15] We found it most effective to merge word parts during MERT (so MERT uses the same stem references as before).

agreement phenomena that we model. Our CRF framework allows us to use more complex context features.

We have directly addressed the question as to whether inflection should be predicted using surface forms as the target of the prediction, or whether linguistic features should be predicted, along with the use of a subsequent generation step. The direct prediction of surface forms is limited to those forms observed in the training data, which is a significant limitation. However, it is reasonable to expect that the use of features (and morphological generation) could also be problematic as this requires the use of morphologically-aware syntactic parsers to annotate the training data with such features, and additionally depends on the coverage of morphological analysis and generation. Despite this, our research clearly shows that the feature-based approach is superior for English-to-German SMT. This is a striking result considering state-of-the-art performance of German parsing is poor compared with the best performance on English parsing. As parsing performance improves, the performance of linguistic-feature-based approaches will increase.

Virpioja et al. (2007), Badr et al. (2008), Luong et al. (2010), Clifton and Sarkar (2011), and others are primarily concerned with using morpheme segmentation in SMT, which is a useful approach for dealing with issues of word-formation. However, this does not deal directly with linguistic features marked by inflection. In German these linguistic features are marked very irregularly and there is widespread syncretism, making it difficult to split off morphemes specifying these features. So it is questionable as to whether morpheme segmentation techniques are sufficient to solve the inflectional problem we are addressing.

Much previous work looks at the impact of using source side information (i.e., feature functions on the aligned English), such as those of Avramidis and Koehn (2008), Yeniterzi and Oflazer (2010) and others. Toutanova et. al.'s work showed that it is most important to model target side coherence and our stem markup also allows us to access source side information. Using additional source side information beyond the markup did not produce a gain in performance.

For compound splitting, we follow Fritzinger and Fraser (2010), using linguistic knowledge encoded in a rule-based morphological analyser and then selecting the best analysis based on the geometric mean of word part frequencies. Other approaches use less deep linguistic resources (e.g., POS-tags Stymne (2008)) or are (almost) knowledge-free (e.g., Koehn and Knight (2003)). Compound merging is less well studied. Popovic et al. (2006) used a simple, list-based merging approach, merging all consecutive words included in a merging list. This approach resulted in too many compounds. We follow Stymne and Cancedda (2011), for compound merging. We trained a CRF using (nearly all) of the features they used and found their approach to be effective (when combined with inflection and portmanteau merging) on one of our two test sets.

10 Conclusion

We have shown that both the prediction of surface forms and the prediction of linguistic features are of interest for improving SMT. We have obtained the advantages of both in our CRF framework, and also integrated handling of compounds, and an inflection-dependent word formation phenomenon, portmanteaus. We validated our work on a well-studied large corpora translation task.

Acknowledgments

The authors wish to thank the anonymous reviewers for their comments. Aoife Cahill was partly supported by Deutsche Forschungsgemeinschaft grant SFB 732. Alexander Fraser, Marion Weller and Fabienne Cap were funded by Deutsche Forschungsgemeinschaft grant Models of Morphosyntax for Statistical Machine Translation. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement Nr. 248005. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views. We thank Thomas Lavergne and Helmut Schmid.

References

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching Morphologically Poor Languages for Statistical Machine Translation. In *Proceedings of ACL-*

08: HLT, pages 763–770, Columbus, Ohio, June. Association for Computational Linguistics.

Ibrahim Badr, Rabih Zbib, and James Glass. 2008. Segmentation for English-to-Arabic statistical machine translation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 153–156, Columbus, Ohio, June. Association for Computational Linguistics.

Ondřej Bojar and Kamil Kos. 2010. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden, July. Association for Computational Linguistics.

Ann Clifton and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 32–42, Portland, Oregon, USA, June. Association for Computational Linguistics.

Adrià de Gispert and José B. Mariño. 2008. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*, 50(11-12):1034–1046.

Alexander Fraser. 2009. Experiments in Morphosyntactic Processing for Translating to and from German. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 115–119, Athens, Greece, March. Association for Computational Linguistics.

Fabienne Fritzinger and Alexander Fraser. 2010. How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing. In *Proceedings of the Fifth Workshop on Statistical Machine Translation*, pages 224–234. Association for Computational Linguistics.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *EACL '03: Proceedings of the 10th conference of the European chapter of the Association for Computational Linguistics*, pages 187–193, Morristown, NJ, USA. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

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.

Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A Hybrid Morpheme-Word Representation for Machine Translation of Morphologically Rich Languages. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 148–157, Cambridge, MA, October. Association for Computational Linguistics.

Maja Popovic, Daniel Stein, and Hermann Ney. 2006. Statistical Machine Translation of German Compound Words. In *Proceedings of FINTAL-06*, pages 616–624, Turku, Finland. Springer Verlag, LNCS.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *4th International Conference on Language Resources and Evaluation*.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of Coling 2004*, pages 162–168, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *International Conference on Spoken Language Processing*.

Sara Stymne and Nicola Cancedda. 2011. Productive Generation of Compound Words in Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 250–260, Edinburgh, Scotland UK, July. Association for Computational Linguistics.

Sara Stymne. 2008. German Compounds in Factored Statistical Machine Translation. In *Proceedings of GOTAL-08*, pages 464–475, Gothenburg, Sweden. Springer Verlag, LNCS/LNAI.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio, June. Association for Computational Linguistics.

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *PROC. OF MT SUMMIT XI*, pages 491–498.

Philip Williams and Philipp Koehn. 2011. Agreement constraints for statistical machine translation into German. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 217–226, Edinburgh, Scotland, July. Association for Computational Linguistics.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-Morphology Mapping in Factored Phrase-Based

Statistical Machine Translation from English to Turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 454–464, Uppsala, Sweden, July. Association for Computational Linguistics.