Determining the placement of German verbs in English–to–German SMT

Anita Gojun **Alexander Fraser**
Institute for Natural Language Processing
University of Stuttgart, Germany
{gojunaa, fraser}@ims.uni-stuttgart.de

Abstract

When translating English to German, existing reordering models often cannot model the long-range reorderings needed to generate German translations with verbs in the correct position. We reorder English as a preprocessing step for English-to-German SMT. We use a sequence of hand-crafted reordering rules applied to English parse trees. The reordering rules place English verbal elements in the positions within the clause they will have in the German translation. This is a difficult problem, as German verbal elements can appear in different positions within a clause (in contrast with English verbal elements, whose positions do not vary as much). We obtain a significant improvement in translation performance.

1 Introduction

Phrase-based SMT (PSMT) systems translate word sequences (phrases) from a source language into a target language, performing reordering of target phrases in order to generate a fluent target language output. The reordering models, such as, for example, the models implemented in Moses (Koehn et al., 2007), are often limited to a certain reordering range since reordering beyond this distance cannot be performed accurately. This results in problems of fluency for language pairs with large differences in constituent order, such as English and German. When translating from English to German, verbs in the German output are often incorrectly left near their position in English, creating problems of fluency. Verbs are also often omitted since the distortion model cannot move verbs to positions which are licensed by the German language model, making the translations difficult to understand.

A common approach for handling the long-range reordering problem within PSMT is performing syntax-based or part-of-speech-based (POS-based) reordering of the input as a preprocessing step before translation (e.g., Collins et al. (2005), Gupta et al. (2007), Habash (2007), Xu et al. (2009), Niehues and Kolss (2009), Katz-Brown et al. (2011), Genzel (2010)).

We reorder English to improve the translation to German. The verb reordering process is implemented using deterministic reordering rules on English parse trees. The sequence of reorderings is derived from the clause type and the composition of a given verbal complex (a (possibly discontiguous) sequence of verbal elements in a single clause). Only one rule can be applied in a given context and for each word to be reordered, there is a unique reordered position. We train a standard PSMT system on the reordered English training and tuning data and use it to translate the reordered English test set into German.

This paper is structured as follows: in section 2, we outline related work. In section 3, English and German verb positioning is described. The reordering rules are given in section 4. In section 5, we show the relevance of the reordering, present the experiments and present an extensive error analysis. We discuss some problems observed in section 7 and conclude in section 8.

2 Related work

There have been a number of attempts to handle the long-range reordering problem within PSMT. Many of them are based on the reordering of a source language sentence as a preprocessing step

before translation. Our approach is related to the work of Collins et al. (2005). They reordered German sentences as a preprocessing step for German-to-English SMT. Hand-crafted reordering rules are applied on German parse trees in order to move the German verbs into the positions corresponding to the positions of the English verbs. Subsequently, the reordered German sentences are translated into English leading to better translation performance when compared with the translation of the original German sentences.

We apply this method on the opposite translation direction, thus having English as a source language and German as a target language. However, we cannot simply invert the reordering rules which are applied on German as a source language in order to reorder the English input. While the reordering of German implies movement of the German verbs into a single position, when reordering English, we need to split the English verbal complexes and, where required, move their parts into different positions. Therefore, we need to identify exactly which parts of a verbal complex must be moved and their possible positions in a German sentence.

Reordering rules can also be extracted automatically. For example, Niehues and Kolss (2009) automatically extracted discontiguous reordering rules (allowing gaps between POS tags which can include an arbitrary number of words) from a word-aligned parallel corpus with POS tagged source side. Since many different rules can be applied on a given sentence, a number of reordered sentence alternatives are created which are encoded as a word lattice (Dyer et al., 2008). They dealt with the translation directions German-to-English and English-to-German, but translation improvement was obtained only for the German-to-English direction. This may be due to missing information about clause boundaries since English verbs often have to be moved to the clause end. Our reordering has access to this kind of knowledge since we are working with a full syntactic parser of English.

Genzel (2010) proposed a language-independent method for learning reordering rules where the rules are extracted from parsed source language sentences. For each node, all possible reorderings (permutations) of a limited number of the child nodes are considered. The candidate reordering rules are applied on the dev set which is then translated and evaluated. Only those rule sequences are extracted which maximize the translation performance of the reordered dev set.

For the extraction of reordering rules, Genzel (2010) uses shallow constituent parse trees which are obtained from dependency parse trees. The trees are annotated using both Penn Treebank POS tags and using Stanford dependency types. However, the constraints on possible reorderings are too restrictive in order to model all word movements required for English-to-German translation. In particular, the reordering rules involve only the permutation of direct child nodes and do not allow changing of child-parent relationships (deleting of a child or attaching a node to a new father node). In our implementation, a verb can be moved to any position in a parse tree (according to the reordering rules): the reordering can be a simple permutation of child nodes, or attachment of these nodes to a new father node (cf. movement of *bought* and *read* in figure 1[1]).

Thus, in contrast to Genzel (2010), our approach does not have any constraints with respect to the position of nodes marking a verb within the tree. Only the syntactic structure of the sentence restricts the distance of the linguistically motivated verb movements.

3 Verb positions in English and German

3.1 Syntax of German sentences

Since in this work, we concentrate on verbs, we use the notion *verbal complex* for a sequence consisting of verbs, verbal particles and negation.

The verb positions in the German sentences depend on clause type and the tense as shown in table 1. Verbs can be placed in 1st, 2nd or clause-final position. Additionally, if a composed tense is given, the parts of a verbal complex can be interrupted by the *middle field* (MF) which contains arbitrary sentence constituents, e.g., subjects and objects (noun phrases), adjuncts (prepositional phrases), adverbs, etc. We assume that the German sentences are SVO (analogously to English); topicalization is beyond the scope of our work.

In this work, we consider two possible positions of the negation in German: (1) directly in

[1]The verb movements shown in figure 1 will be explained in detail in section 4.

	1st	2nd	MF	clause-final
decl	subject	**finV**	any	∅
	subject	**finV**	any	**mainV**
int/perif	**finV**	subject	any	∅
	finV	subject	any	**mainV**
sub/inf	relCon	subject	any	**finV**
	relCon	subject	any	**VC**

Table 1: Position of the German subjects and verbs in declarative clauses (*decl*), interrogative clauses and clauses with a peripheral clause (*int/perif*), subordinate/infinitival (*sub/inf*) clauses. *mainV* = main verb, *finV* = finite verb, *VC* = verbal complex, *any* = arbitrary words, *relCon* = relative pronoun or conjunction. We consider extraponed consituents in *perif*, as well as optional interrogatives in *int* to be in position 0.

front of the main verb, and (2) directly after the finite verb. The two negation positions are illustrated in the following examples:

(1) Ich behaupte, dass ich es **nicht** gesagt habe.
I claim that I it not say did.

(2) Ich denke **nicht**, dass er das gesagt hat.
I think not that he that said has.

It should, however, be noted that in German, the negative particle *nicht* can have several positions in a sentence depending on the context (verb arguments, emphasis). Thus, more analysis is ideally needed (e.g., discourse, etc.).

3.2 Comparison of verb positions

English and German verbal complexes differ both in their construction and their position. The German verbal complex can be discontiguous, i.e., its parts can be placed in different positions which implies that a (large) number of other words can be placed between the verbs (situated in the MF). In English, the verbal complex can only be interrupted by adverbials and subjects (in interrogative clauses). Furthermore, in German, the finite verb can sometimes be the last element of the verbal complex, while in English, the finite verb is always the first verb in the verbal complex.

In terms of positions, the verbs in English and German can differ significantly. As previously noted, the German verbal complex can be discontiguous, simultaneously occupying 1st/2nd and clause-final position (cf. rows *decl* and *int/perif* in table 1), which is not the case in English. While in English, the verbal complex is placed in the 2nd position in declarative, or in the 1st position in interrogative clauses, in German, the entire verbal complex can additionally be placed at the clause end in subordinate or infinitival clauses (cf. row *sub/inf* in table 1).

Because of these differences, for nearly all types of English clauses, reordering is needed in order to place the English verbs in the positions which correspond to the correct verb positions in German. Only English declarative clauses with simple present and simple past tense have the same verb position as their German counterparts. We give statistics on clause types and their relevance for the verb reordering in section 5.1.

4 Reordering of the English input

The reordering is carried out on English parse trees. We first enrich the parse trees with clause type labels, as described below. Then, for each node marking a clause (*S* nodes), the corresponding sequence of reordering rules is carried out. The appropriate reordering is derived from the clause type label and the composition of the given verbal complex. The reordering rules are deterministic. Only one rule can be applied in a given context and for each verb to be reordered, there is a unique reordered position.

The reordering procedure is the same for the training and the testing data. It is carried out on English parse trees resulting in modified parse trees which are read out in order to generate the reordered English sentences. These are input for training a PSMT system or input to the decoder. The processing steps are shown in figure 1.

For the development of the reordering rules, we used a small sample of the training data. In particular, by observing the English parse trees extracted randomly from the training data, we developed a set of rules which transform the original trees in such a way that the English verbs are moved to the positions which correspond to the placement of verbs in German.

4.1 Labeling clauses with their type

As shown in section 3.1, the verb positions in German depend on the clause type. Since we use English parse trees produced by the generative parser of Charniak and Johnson (2005) which do not have any function labels, we implemented a simple rule-based clause type labeling script which

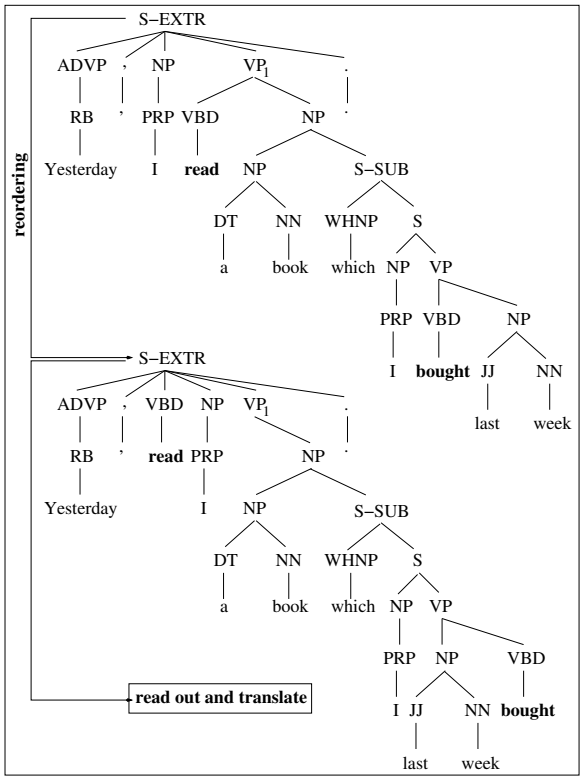

Figure 1: Processing steps: Clause type labeling annotates the given original tree with clause type labels (in figure, *S-EXTR* and *S-SUB*). Subsequently, the reordering is performed (cf. movement of the verbs *read* and *bought*). The reordered sentence is finally read out and given to the decoder.

enriches every clause starting node with the corresponding clause type label. The label depends on the context (father, child nodes) of a given clause node. If, for example, the first child node of a given *S* node is *WH** (wh-word) or *IN* (subordinating conjunction), then the clause type label is *SUB* (subordinate clause, cf. figure 1).

We defined five clause type labels which indicate main clauses (*MAIN*), main clauses with a peripheral clause in the prefield (*EXTR*), subordinate (*SUB*), infinitival (*XCOMP*) and interrogative clauses (*INT*).

4.2 Clause boundary identification

The German verbs are often placed at the clause end (cf. rows *decl*, *int/perif* and *sub/inf* in table 1), making it necessary to move their English counterparts into the corresponding positions within an English tree. For this reason, we identify the clause ends (the right boundaries). The search for the clause end is implemented as a breadth-first search for the next *S* node or sentence end. The starting node is the node which marks the verbal phrase in which the verbs are enclosed. When the next node marking a clause is identified, the search stops and returns the position in front of the identified clause marking node.

When, for example, searching for the clause boundary of *S-EXTR* in figure 1, we search recursively for the first clause marking node within VP_1, which is *S-SUB*. The position in front of *S-SUB* is marked as clause-final position of *S-EXTR*.

4.3 Basic verb reordering rules

The reordering procedure takes into account the following word categories: verbs, verb particles, the infinitival particle *to* and the negative particle *not*, as well as its abbreviated form *'t*. The reordering rules are based on POS labels in the parse tree.

The reordering procedure is a sequence of applications of the reordering rules. For each element of an English verbal complex, its properties are derived (tense, main verb/auxiliary, finiteness). The reordering is then carried out corresponding to the clause type and verbal properties of a verb to be processed.

In the following, the reordering rules are presented. Examples of reordered sentences are given in table 2, and are discussed further here.

Main clause (S-MAIN)

(i) simple tense: no reordering required (cf. $appears_{finV}$ in input 1);

(ii) composed tense: the main verb is moved to the clause end. If a negative particle exists, it is moved in front of the reordered main verb, while the optional verb particle is moved after the reordered main verb (cf. $[has]_{finV}$ $[been\ developing]_{mainV}$ in input 2).

Main clause with peripheral clause (S-EXTR)

(i) simple tense: the finite verb is moved together with an optional particle to the 1st position (i.e. in front of the subject);

(ii) composed tense: the main verb, as well as optional negative and verb particles are moved to the clause end. The finite verb is moved in the 1st position, i.e. in front of the subject (cf. $have_{finV}$ $[gone\ up]_{mainV}$ in input 3).

Subordinate clause (S-SUB)

(i) simple tense: the finite verb is moved to the clause end (cf. *boasts*$_{finV}$ in input 3);

(ii) composed tense: the main verb, as well as optional negative and verb particles are moved to the clause end, the finite verb is placed after the reordered main verb (cf. *have*$_{finV}$ *[been executed]*$_{mainV}$ in input 5).

Infinitival clause (S-XCOMP)

The entire English verbal complex is moved from the 2nd position to the clause-final position (cf. *[to discuss]*$_{VC}$ in input 4).

Interrogative clause (S-INT)

(i) simple tense: no reordering required;

(ii) composed tense: the main verb, as well as optional negative and verb particles are moved to the clause end (cf. *[did]*$_{finV}$ *know*$_{mainV}$ in input 5).

4.4 Reordering rules for other phenomena

4.4.1 Multiple auxiliaries in English

Some English tenses require a sequence of auxiliaries, not all of which have a German counterpart. In the reordering process, non-finite auxiliaries are considered to be a part of the main verb complex and are moved together with the main verb (cf. movement of *has*$_{finV}$ *[been developing]*$_{mainV}$ in input 2).

4.4.2 Simple vs. composed tenses

In English, there are some tenses composed of an auxiliary and a main verb which correspond to a German tense composed of only one verb, e.g., *am reading* ⇔ *lese* and *does John read?* ⇔ *liest John?* Splitting such English verbal complexes and only moving the main verbs would lead to constructions which do not exist in German. Therefore, in the reordering process, the English verbal complex in present continuous, as well as interrogative phrases composed of *do* and a main verb, are not split. They are handled as *one* main verb complex and reordered as a single unit using the rules for main verbs (e.g. *[because I am reading a book]*$_{SUB}$ ⇒ *because I a book am reading* ⇔ *weil ich ein Buch lese*.[2]

[2]We only consider present continuous and verbs in combination with *do* for this kind of reordering. There are also other tenses which could (or should) be treated in the same way (cf. *has been developing* on input 2, table 2). We do not do this to keep the reordering rules simple and general.

4.4.3 Flexible position of German verbs

We stated that the English verbs are never moved outside the subclause they were originally in. In German there are, however, some constructions (infinitival and relative clauses), in which the main verb can be placed *after* a subsequent clause. Consider two German translations of the English sentence *He has promised to come*:

(3a) Er hat [zu kommen]$_S$ versprochen.
he has to come promised.

(3b) Er hat versprochen, [zu kommen]$_S$.
he has promised, to come.

In (3a), the German main verb *versprochen* is placed after the infinitival clause *zu kommen* (*to come*), while in (3b), the same verb is placed in front of it. Both alternatives are grammatically correct.

If a German verb should come after an embedded clause as in example (3a) or precede it (cf. example (3b)), depends not only on syntactic but also on stylistic factors. Regarding the verb reordering problem, we would therefore have to examine the given sentence in order to derive the *correct* (or *more probable*) new verb position which is beyond the scope of this work. Therefore, we allow only for reorderings which do not cross clause boundaries as shown in example (3b).

5 Experiments

In order to evaluate the translation of the reordered English sentences, we built two SMT systems with Moses (Koehn et al., 2007). As training data, we used the Europarl corpus which consists of 1,204,062 English/German sentence pairs. The baseline system was trained on the original English training data while the contrastive system was trained on the reordered English training data. In both systems, the same original German sentences were used. We used WMT 2009 dev and test sets to tune and test the systems. The baseline system was tuned and tested on the original data while for the contrastive system, we used the reordered English side of the dev and test sets. The German 5-gram language model used in both systems was trained on the WMT 2009 German language modeling data, a large German newspaper corpus consisting of 10,193,376 sentences.

Input 1	The programme **appears to be** successful for published data **shows** that MRSA **is** on the decline in the UK.
Reordered	The programme **appears** successful **to be** for published data **shows** that MRSA on the decline in the UK **is**.
Input 2	The real estate market in Bulgaria **has been developing** at an unbelievable rate - all of Europe **has** its eyes on this heretofore rarely heard-of Balkan nation.
Reordered	The real estate market in Bulgaria **has** at an unbelievable rate **been developing** - all of Europe **has** its eyes on this heretofore rarely heard-of Balkan nation.
Input 3	While Bulgaria **boasts** the European Union's lowest real estate prices, they **have** still **gone up** by 21 percent in the past five years.
Reordered	While Bulgaria the European Union's lowest real estate prices **boasts**, **have** they still by 21 percent in the past five years **gone up**.
Input 4	Professionals and politicians from 192 countries **are slated to discuss** the Bali Roadmap that **focuses** on efforts **to cut** greenhouse gas emissions after 2012, when the Kyoto Protocol **expires**.
Reordered	Professionals and politicians from 192 countries **are slated** the Bali Roadmap **to discuss** that on efforts **focuses** greenhouse gas emissions after 2012 **to cut**, when the Kyoto Protocol **expires**.
Input 5	**Did** you **know** that in that same country, since 1976, 34 mentally-retarded offenders **have been executed**?
Reordered	**Did** you **know** that in that same country, since 1976, 34 mentally-retarded offenders **been executed have**?

Table 2: Examples of reordered English sentences

5.1 Applied rules

In order to see how many English clauses are relevant for reordering, we derived statistics about clause types and the number of reordering rules applied on the training data.

In table 3, the number of the English clauses with all considered clause type/tense combination are shown. The bold numbers indicate combinations which are relevant to the reordering. Overall, 62% of all EN clauses from our training data (2,706,117 clauses) are relevant for the verb reordering. Note that there is an additional category *rest* which indicates incorrect clause type/tense combinations and might thus not be correctly reordered. These are mostly due to parsing and/or tagging errors.

The performance of the systems was measured by BLEU (Papineni et al., 2002). The evaluation results are shown in table 4. The contrastive system outperforms the baseline. Its BLEU score is 13.63 which is a gain of 0.61 BLEU points over the baseline. This is a statistically significant improvement at $p<0.05$ (computed with Gimpel's implementation of the pairwise bootstrap resampling method (Koehn, 2004)).

Manual examination of the translations produced by both systems confirms the result of the automatic evaluation. Many translations produced by the contrastive system now have verbs in the correct positions. If we compare the generated translations for input sentence 1 in table 5, we see that the contrastive system generates a translation in which all verbs are placed correctly. In the baseline translation, only the translation of the finite verb *was*, namely *war*, is placed correctly, while the translation of the main verb (*diagnosed* $\rightarrow$ *festgestellt*) should be placed at the clause end as in the translation produced by our system.

tense	MAIN	EXTR	SUB	INT	XCOMP
simple	675,095	**170,806**	**449,631**	8,739	-
composed	**343,178**	**116,729**	**277,733**	**8,817**	**314,573**
rest	98,464	5,158	90,139	306	146,746

Table 3: Counts of English clause types and used tenses. Bold numbers indicate clause type/tense combinations where reordering is required.

	Baseline	**Reordered**
BLEU	13.02	13.63

Table 4: Scores of baseline and contrastive systems

5.2 Evaluation

Often, the English verbal complex is translated only partially by the baseline system. For example, the English verbal complexes in sentence 2 in table 5 *will climb* and *will drop* are only partially translated (*will climb* $\rightarrow$ *wird (will)*, *will drop* $\rightarrow$ *fallen (fall)*). Moreover, the generated verbs are placed incorrectly. In our translation, all verbs are translated and placed correctly.

Another problem which was often observed in the baseline is the omission of the verbs in the German translations. The baseline translation of the example sentence 3 in table 5 illustrates such

a case. There is no translation of the English infinitival verbal complex *to have*. In the translation generated by the contrastive system, the verbal complex does get translated (*zu haben*) and is also placed correctly. We think this is because the reordering model is not able to identify the position for the verb which is licensed by the language model, causing a hypothesis with no verb to be scored higher than the hypotheses with incorrectly placed verbs.

6 Error analysis

6.1 Erroneous reordering in our system

In some cases, the reordering of the English parse trees fails. Most erroneous reorderings are due to a number of different parsing and tagging errors.

Coordinated verbs are also problematic due to their complexity. Their composition can vary, and thus it would require a large number of different reordering rules to fully capture this. In our reordering script, the movement of complex structures such as verbal phrases consisting of a sequence of child nodes is not implemented (only nodes with one child, namely the verb, verbal particle or negative particle are moved).

6.2 Splitting of the English verbal complex

Since in many cases, the German verbal complex is discontiguous, we need to split the English verbal complex and move its parts into different positions. This ensures the correct placement of German verbs. However, this does not ensure that the German verb forms are correct because of highly ambiguous English verbs. In some cases, we can lose contextual information which would be useful for disambiguating ambiguous verbs and generating the appropriate German verb forms.

6.2.1 Subject–verb agreement

Let us consider the English clause in (4a) and its reordered version in (4b):
(4a) ... because they **have said** it to me yesterday.
(4b) ... because they it to me yesterday **said have**.
In (4b), the English verbs *said have* are separated from the subject *they*. The English *said have* can be translated in several ways into German. Without any information about the subject (the distance between the verbs and the subject can be very large), it is relatively likely that an erroneous German translation is generated.

On the other hand, in the baseline SMT system, the subject *they* is likely to be a part of a translation phrase with the correct German equivalent (*they have said* $\rightarrow$ *sie haben gesagt*). *They* is then used as a disambiguating context which is missing in the reordered sentence (but the order is wrong).

6.2.2 Verb dependency

A similar problem occurs in a verbal complex:
(5a) They **have said** it to me yesterday.
(5b) They **have** it to me yesterday **said**.

In sentence (5a), the English consecutive verbs *have said* are a sequence consisting of a finite auxiliary *have* and the past participle *said*. They should be translated into the corresponding German verbal complex *haben gesagt*. But, if the verbs are split, we will probably get translations which are completely independent. Even if the German auxiliary is correctly inflected, it is hard to predict how *said* is going to be translated. If the distance between the auxiliary *habe* and the hypothesized translation of *said* is large, the language model will not be able to help select the correct translation. Here, the baseline SMT system again has an advantage as the verbs are consecutive. It is likely they will be found in the training data and extracted with the correct German phrase (but the German order is again incorrect).

6.3 Collocations

Collocations (verb–object pairs) are another case which can lead to a problem:
(6a) I think that the discussion **would take place** later this evening.
(6b) I think that the discussion **place** later this evening **take would**.
The English collocation in (6a) consisting of the verb *take* and the object *place* corresponds to the German verb *stattfinden*. Without this specific object, the verb *take* is likely to be translated literally. In the reordered sentence, the verbal complex *take would* is indeed separated from the object *place* which would probably lead to the literal translation of both parts of the mentioned collocation. So, as already described in the preceding paragraphs, an important source of contextual information is lost which could ensure the correct translation of the given phrase.

This problem is not specific to English–to–German. For instance, the same problem occurs when translating German into English. If, for ex-

Input 1	An MRSA - an antibiotic resistant staphylococcus - infection **was** recently **diagnosed** in the traumatology ward of János hospital.
Reordered input	An MRSA - an antibiotic resistant staphylococcus - infection **was** recently in the traumatology ward of János hospital **diagnosed**.
Baseline translation	Ein MRSA - ein Antibiotikum resistenter Staphylococcus - **war** vor kurzem in der **festgestellt** A MRSA - an antibiotic resistant Staphylococcus - was before recent in the diagnosed traumatology Ward von János Krankenhaus. traumatology ward of János hospital.
Reordered translation	Ein MRSA - ein Antibiotikum resistenter Staphylococcus - Infektion **wurde** vor kurzem in den A MRSA - an antibiotic resistant Staphylococcus - infection was before recent in the traumatology Station der János Krankenhaus **diagnostiziert**. traumatology ward of János hospital diagnosed.
Input 2	The ECB **predicts** that 2008 inflation **will climb** to 2.5 percent from the earlier 2.1, but **will drop** back to 1.9 percent in 2009.
Reordered input	The ECB **predicts** that 2008 inflation to 2.5 percent from the earlier 2.1 **will climb**, but back to 1.9 percent in 2009 **will drop**.
Baseline translation	Die EZB **sagt**, dass 2008 die Inflationsrate **wird** auf 2,5 Prozent aus der früheren 2,1, sondern The ECB says, that 2008 the inflation rate will to 2.5 percent from the earlier 2.1, but **fallen** zurück auf 1,9 Prozent im Jahr 2009. fall back to 1.9 percent in the year 2009.
Reordered translation	Die EZB **prophezeit**, dass 2008 die Inflation zu 2,5 Prozent aus der früheren 2,1 **ansteigen** The ECB predicts, that 2008 the inflation rate to 2.5 percent from the earlier 2.1 climb **wird**, aber auf 1,9 Prozent in 2009 **sinken wird**. will, but to 1.9 percent in 2009 fall will.
Input 3	Labour Minister Mónika Lamperth **appears not to have** a sensitive side.
R. input	Labour Minister Mónika Lamperth **appears** a sensitive side **not to have** .
Baseline translation	Arbeitsminister Mónika Lamperth **scheint nicht** eine sensible Seite. Labour Minister Mónika Lamperth appears not a sensitive side.
Reordered translation	Arbeitsminister Mónika Lamperth **scheint** eine sensible Seite **nicht zu haben**. Labour Minister Mónika Lamperth appears a sensitive side not to have.

Table 5: Example translations, the baseline has problems with verbal elements, reordered is correct

ample, the object *Kauf (buying)* of the collocation *nehmen + in Kauf (accept)* is separated from the verb *nehmen (take)*, they are very likely to be translated literally (rather than as the idiom meaning "to accept"), thus leading to an erroneous English translation.

6.4 Error statistics

We manually checked 100 randomly chosen English sentences to see how often the problems described in the previous sections occur. From a total of 276 clauses, 29 were not reordered correctly. 20 errors were caused by incorrect parsing and/or POS tags, while the remaining 9 are mostly due to different kinds of coordination. Table 6 shows correctly reordered clauses which might pose a problem for translation (see sections 6.2–6.3). Although the positions of the verbs in the translations are now correct, the distance between subjects and verbs, or between verbs in a single VP might lead to the generation of erroneously inflected verbs. The separate generation of German verbal morphology is an interesting area of future work, see (de Gispert and Mariño, 2008). We also found 2 problematic collocations but note that this only gives a rough idea of the problem, further study is needed.

6.5 POS-based disambiguation of the English verbs

With respect to the problems described in 6.2.1 and 6.2.2, we carried out an experiment in which

	total	d $\geq$ 5 tokens
subject–verb	40	19
verb dependency	32	14
collocations	8	2

Table 6: *total* is the number of clauses found for the respective phenomenon. *d $\geq$ 5 tokens* is the number of clauses where the distance between relevant tokens is at least 5, which is problematic.

	Baseline + POS	Reordered + POS
BLEU	13.11	13.68

Table 7: BLEU scores of the baseline and the contrastive SMT system using verbal POS tags

we used POS tags in order to disambiguate the English verbs. For example, the English verb *said* corresponds to the German participle *gesagt*, as well as to the finite verb in simple past, e.g. *sagte*. We attached the POS tags to the English verbs in order to simulate a disambiguating suffix of a verb (e.g. *said* $\Rightarrow$ *said_VBN, said_VBD*). The idea behind this was to extract the correct verbal translation phrases and score them with appropriate translation probabilities (e.g. *p(said_VBN, gesagt) > p(said_VBN, sagte)*.

We built and tested two PSMT systems using the data enriched with verbal POS tags. The first system is trained and tested on the original English sentences, while the contrastive one was trained and tested on the reordered English sentences. Evaluation results are shown in table 7.

The baseline obtains a gain of 0.09 and the contrastive system of 0.05 BLEU points over the corresponding PSMT system without POS tags. Although there are verbs which are now generated correctly, the overall translation improvement lies under our expectation. We will directly model the inflection of German verbs in future work.

7 Discussion and future work

We implemented reordering rules for English verbal complexes because their placement differs significantly from German placement. The implementation required dealing with three important problems: (i) definition of the clause boundaries, (ii) identification of the new verb positions and (iii) correct splitting of the verbal complexes.

We showed some phenomena for which a stochastic reordering would be more appropriate. For example, since in German, the auxiliary and the main verb of a verbal complex can occupy different positions in a clause, we had to define the English counterparts of the two components of the German verbal complex. We defined non-finite English verbal elements as a part of the main verb complex which are then moved together with the main verb. This rigid definition could be relaxed by considering multiple different splittings and movements of the English verbs.

Furthermore, the reordering rules are applied on a clause not allowing for movements across the clause boundaries. However, we also showed that in some cases, the main verbs may be moved after the succeeding subclause. Stochastic rules could allow for both placements or carry out the more probable reordering given a specific context. We will address these issues in future work.

Unfortunately, some important contextual information is lost when splitting and moving English verbs. When English verbs are highly ambiguous, erroneous German verbs can be generated. The experiment described in section 6.5 shows that more effort should be made in order to overcome this problem. The incorporation of separate morphological generation of inflected German verbs would improve translation.

8 Conclusion

We presented a method for reordering English as a preprocessing step for English–to–German SMT. To our knowledge, this is one of the first papers which reports on experiments regarding the reordering problem for English–to–German SMT. We showed that the reordering rules specified in this work lead to improved translation quality. We observed that verbs are placed correctly more often than in the baseline, and that verbs which were omitted in the baseline are now often generated. We carried out a thorough analysis of the rules applied and discussed problems which are related to highly ambiguous English verbs. Finally we presented ideas for future work.

Acknowledgments

This work was funded by Deutsche Forschungsgemeinschaft grant Models of Morphosyntax for Statistical Machine Translation.

References

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *ACL*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *ACL*.

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

Chris Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *ACL-HLT*.

Dmitriy Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *COLING*.

Deepa Gupta, Mauro Cettolo, and Marcello Federico. 2007. POS-based reordering models for statistical machine translation. In *Proceedings of the Machine Translation Summit (MT-Summit)*.

Nizar Habash. 2007. Syntactic preprocessing for statistical machine translation. In *Proceedings of the Machine Translation Summit (MT-Summit)*.

Jason Katz-Brown, Slav Petrov, Ryan McDonald, Franz Och, David Talbot, Hiroshi Ichikawa, Masakazu Seno, and Hideto Kazawa. 2011. Training a parser for machine translation reordering. In *EMNLP*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, Demonstration Program*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*.

Jan Niehues and Muntsin Kolss. 2009. A POS-based model for long-range reorderings in SMT. In *EACL Workshop on Statistical Machine Translation*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Peng Xu, Jaecho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve SMT for subject-object-verb languages. In *NAACL*.