Clause Restructuring for Statistical Machine Translation

Michael Collins
MIT CSAIL
mcollins@csail.mit.edu

Philipp Koehn
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk

Ivona Kučerová
MIT Linguistics Department
kucerova@mit.edu

Abstract

We describe a method for incorporating syntactic information in statistical machine translation systems. The first step of the method is to parse the source language string that is being translated. The second step is to apply a series of transformations to the parse tree, effectively reordering the surface string on the source language side of the translation system. The goal of this step is to recover an underlying word order that is closer to the target language word-order than the original string. The reordering approach is applied as a pre-processing step in both the training and decoding phases of a phrase-based statistical MT system. We describe experiments on translation from German to English, showing an improvement from 25.2% Bleu score for a baseline system to 26.8% Bleu score for the system with reordering, a statistically significant improvement.

1 Introduction

Recent research on statistical machine translation (SMT) has lead to the development of *phrase-based* systems (Och et al., 1999; Marcu and Wong, 2002; Koehn et al., 2003). These methods go beyond the original IBM machine translation models (Brown et al., 1993), by allowing multi-word units ("phrases") in one language to be translated directly into phrases in another language. A number of empirical evaluations have suggested that phrase-based systems currently represent the state–of–the–art in statistical machine translation.

In spite of their success, a key limitation of phrase-based systems is that they make little or no direct use of syntactic information. It appears likely that syntactic information will be crucial in accurately modeling many phenomena during translation, for example systematic differences between the word order of different languages. For this reason there is currently a great deal of interest in methods which incorporate syntactic information within statistical machine translation systems (e.g., see (Alshawi, 1996; Wu, 1997; Yamada and Knight, 2001; Gildea, 2003; Melamed, 2004; Graehl and Knight, 2004; Och et al., 2004; Xia and McCord, 2004)).

In this paper we describe an approach for the use of syntactic information within phrase-based SMT systems. The approach constitutes a simple, direct method for the incorporation of syntactic information in a phrase–based system, which we will show leads to significant improvements in translation accuracy. The first step of the method is to parse the source language string that is being translated. The second step is to apply a series of transformations to the resulting parse tree, effectively reordering the surface string on the source language side of the translation system. The goal of this step is to recover an underlying word order that is closer to the target language word-order than the original string. Finally, we apply a phrase-based system to the reordered string to give a translation into the target language.

We describe experiments involving machine translation from German to English. As an illustrative example of our method, consider the following German sentence, together with a "translation" into English that follows the original word order:

Original sentence: Ich werde Ihnen die entsprechenden Anmerkungen aushaendigen, damit Sie das eventuell bei der Abstimmung uebernehmen koennen.

English translation: *I will to you the corresponding comments pass on, so that you them perhaps in the vote adopt can.*

The German word order in this case is substantially different from the word order that would be seen in English. As we will show later in this paper, translations of sentences of this type pose difficulties for phrase-based systems. In our approach we reorder the constituents in a parse of the German sentence to give the following word order, which is much closer to the target English word order (words which have been "moved" are underlined):

Reordered sentence: Ich werde <u>aushaendigen</u> Ihnen die entsprechenden Anmerkungen, damit Sie <u>koennen</u> <u>uebernehmen</u> das eventuell bei der Abstimmung.

English translation: *I will <u>pass on</u> to you the corresponding comments, so that you <u>can</u> <u>adopt</u> them perhaps in the vote.*

Proceedings of the 43rd Annual Meeting of the ACL, pages 531–540,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

We applied our approach to translation from German to English in the Europarl corpus. Source language sentences are reordered in test data, and also in training data that is used by the underlying phrase-based system. Results using the method show an improvement from 25.2% Bleu score to 26.8% Bleu score (a statistically significant improvement), using a phrase-based system (Koehn et al., 2003) which has been shown in the past to be a highly competitive SMT system.

2 Background

2.1 Previous Work

2.1.1 Research on Phrase-Based SMT

The original work on statistical machine translation was carried out by researchers at IBM (Brown et al., 1993). More recently, phrase-based models (Och et al., 1999; Marcu and Wong, 2002; Koehn et al., 2003) have been proposed as a highly successful alternative to the IBM models. Phrase-based models generalize the original IBM models by allowing multiple words in one language to correspond to multiple words in another language. For example, we might have a translation entry specifying that *I will* in English is a likely translation for *Ich werde* in German.

In this paper we use the phrase-based system of (Koehn et al., 2003) as our underlying model. This approach first uses the original IBM models to derive word-to-word alignments in the corpus of example translations. Heuristics are then used to grow these alignments to encompass phrase-to-phrase pairs. The end result of the training process is a lexicon of phrase-to-phrase pairs, with associated costs or probabilities. In translation with the system, a beam search method with left-to-right search is used to find a high scoring translation for an input sentence. At each stage of the search, one or more English words are added to the hypothesized string, and one or more consecutive German words are "absorbed" (i.e., marked as having already been translated—note that each word is absorbed at most once). Each step of this kind has a number of costs: for example, the log probability of the phrase-to-phrase correspondance involved, the log probability from a language model, and some "distortion" score indicating how likely it is for the proposed words in the English string to be aligned to the corresponding position in the German string.

2.1.2 Research on Syntax-Based SMT

A number of researchers (Alshawi, 1996; Wu, 1997; Yamada and Knight, 2001; Gildea, 2003; Melamed, 2004; Graehl and Knight, 2004; Galley et al., 2004) have proposed models where the translation process involves syntactic representations of the source and/or target languages. One class of approaches make use of "bitext" grammars which simultaneously parse both the source and target languages. Another class of approaches make use of syntactic information in the target language alone, effectively transforming the translation problem into a parsing problem. Note that these models have radically different structures and parameterizations from phrase–based models for SMT. As yet, these systems have not shown significant gains in accuracy in comparison to phrase-based systems.

Reranking methods have also been proposed as a method for using syntactic information (Koehn and Knight, 2003; Och et al., 2004; Shen et al., 2004). In these approaches a baseline system is used to generate N-best output. Syntactic features are then used in a second model that reranks the N-best lists, in an attempt to improve over the baseline approach. (Koehn and Knight, 2003) apply a reranking approach to the sub-task of noun-phrase translation. (Och et al., 2004; Shen et al., 2004) describe the use of syntactic features in reranking the output of a full translation system, but the syntactic features give very small gains: for example the majority of the gain in performance in the experiments in (Och et al., 2004) was due to the addition of IBM Model 1 translation probabilities, a non-syntactic feature.

An alternative use of syntactic information is to employ an existing statistical parsing model as a language model within an SMT system. See (Charniak et al., 2003) for an approach of this form, which shows improvements in accuracy over a baseline system.

2.1.3 Research on Preprocessing Approaches

Our approach involves a preprocessing step, where sentences in the language being translated are modified before being passed to an existing phrase-based translation system. A number of other researchers (Berger et al., 1996; Niessen and Ney, 2004; Xia and McCord, 2004) have described previous work on preprocessing methods. (Berger et al., 1996) describe an approach that targets translation of French phrases of the form *NOUN de NOUN* (e.g., *conflit d'intérêt*). This was a relatively limited study, concentrating on this one syntactic phenomenon which involves relatively local transformations (a parser was not required in this study). (Niessen and Ney, 2004) describe a method that combines morphologically–split verbs in German, and also reorders questions in English and German. Our method goes beyond this approach in several respects, for example considering phenomena such as declarative (non-question) clauses, subordinate clauses, negation, and so on.

(Xia and McCord, 2004) describe an approach for translation from French to English, where reordering rules are acquired automatically. The reordering rules in their approach operate at the level of context-free rules in the parse tree. Our method differs from that of (Xia and McCord, 2004) in a couple of important respects. First, we are considering German, which arguably has more challenging word order phenonema than French. German has relatively free word order, in contrast to both English and French: for example, there is considerable flexibility in terms of which phrases can appear in the first position in a clause. Second, Xia et. al's (2004) use of reordering rules stated at the context-free level differs from ours. As one example, in our approach we use a single transformation that moves an infinitival verb to the first position in a verb phrase. Xia et. al's approach would require learning of a different rule transformation for every production of the form `VP => ...`. In practice the German parser that we are using creates relatively "flat" structures at the VP and clause levels, leading to a huge number of context-free rules (the flatness is one consequence of the relatively free word order seen within VP's and clauses in German). There are clearly some advantages to learning reordering rules automatically, as in Xia et. al's approach. However, we note that our approach involves a handful of linguistically–motivated transformations and achieves comparable improvements (albeit on a different language pair) to Xia et. al's method, which in contrast involves over 56,000 transformations.

```
S PPER-SB  Ich
  VAFIN-HD  werde
  VP PPER-DA  Ihnen
     NP-OA ART   die
           ADJA  entsprechenden
           NN    Anmerkungen
     VVINF-HD    aushaendigen
     ,  ,
     S KOUS     damit
       PPER-SB  Sie
       VP PDS-OA  das
          ADJD eventuell
          PP APPR bei
             ART  der
             NN   Abstimmung
          VVINF-HD  uebernehmen
       VMFIN-HD  koennen
```

Figure 1: An example parse tree. Key to non-terminals: `PPER` = personal pronoun; `VAFIN` = finite verb; `VVINF` = infinitival verb; `KOUS` = complementizer; `APPR` = preposition; `ART` = article; `ADJA` = adjective; `ADJD` = adverb; `-SB` = subject; `-HD` = head of a phrase; `-DA` = dative object; `-OA` = accusative object.

2.2 German Clause Structure

In this section we give a brief description of the syntactic structure of German clauses. The characteristics we describe motivate the reordering rules described later in the paper.

Figure 1 gives an example parse tree for a German sentence. This sentence contains two clauses:

Clause 1: Ich/*I* werde/*will* Ihnen/*to_you* die/*the* entsprechenden/*corresponding* Anmerkungen/*comments* aushaendigen/*pass_on*

Clause 2: damit/*so_that* Sie/*you* das/*them* eventuell/*perhaps* bei/*in* der/*the* Abstimmung/*vote* uebernehmen/*adopt* koennen/*can*

These two clauses illustrate a number of syntactic phenomena in German which lead to quite different word order from English:

Position of finite verbs. In Clause 1, which is a matrix clause, the finite verb *werde* is in the second position in the clause. Finite verbs appear rigidly in 2nd position in matrix clauses. In contrast, in subordinate clauses, such as Clause 2, the finite verb comes last in the clause. For example, note that *koennen* is a finite verb which is the final element of Clause 2.

Position of infinitival verbs. In German, infinitival verbs are final within their associated verb phrase. For example, returning to Figure 1, notice that *aushaendigen* is the last element in its verb phrase, and that *uebernehmen* is the final element of its verb phrase in the figure.

Relatively flexible word ordering. German has substantially freer word order than English. In particular, note that while the verb comes second in matrix clauses, essentially any element can be in the first position. For example, in Clause 1, while the subject *Ich* is seen in the first position, potentially any of the other constituents (e.g., *Ihnen*) could also appear in this position. Note that this often leads to the subject following the finite verb, something which happens very rarely in English.

There are many other phenomena which lead to differing word order between German and English. Two others that we focus on in this paper are negation (the differing placement of items such as *not* in English and *nicht* in German), and also verb-particle constructions. We describe our treatment of these phenomena later in this paper.

2.3 Reordering with Phrase-Based SMT

We have seen in the last section that German syntax has several characteristics that lead to significantly different word order from that of English. We now describe how these characteristics can lead to difficulties for phrase–based translation systems when applied to German to English translation.

Typically, reordering models in phrase-based systems are based solely on movement distance. In particular, at each point in decoding a "cost" is associated with skipping over 1 or more German words. For example, assume that in translating

> Ich werde Ihnen die entsprechenden Anmerkungen aushaendigen.

we have reached a state where "Ich" and "werde" have been translated into "I will" in English. A potential decoding decision at this point is to add the phrase "pass on" to the English hypothesis, at the same time absorbing "aushaendigen" from the German string. The cost of this decoding step will involve a number of factors, including a cost of skipping over a phrase of length 4 (i.e., *Ihnen die entsprechenden Anmerkungen*) in the German string.

The ability to penalise "skips" of this type, and the potential to model multi-word phrases, are essentially the main strategies that the phrase-based system is able to employ when modeling differing word-order across different languages. In practice, when training the parameters of an SMT system, for example using the discriminative methods of (Och, 2003), the cost for skips of this kind is typically set to a very high value. In experiments with the system of (Koehn et al., 2003) we have found that in practice a large number of complete translations are completely monotonic (i.e., have 0 skips), suggesting that the system has difficulty learning exactly what points in the translation should allow reordering. In summary, phrase-based systems have relatively limited potential to model word-order differences between different languages.

The reordering stage described in this paper attempts to modify the source language (e.g., German) in such a way that its word order is very similar to that seen in the target language (e.g., English). In an ideal approach, the resulting translation problem that is passed on to the phrase-based system will be solvable using a completely monotonic translation, without any skips, and without requiring extremely long phrases to be translated (for example a phrasal translation corresponding to *Ihnen die entsprechenden Anmerkungen aushaendigen*).

Note than an additional benefit of the reordering phase is that it may bring together groups of words in German which have a natural correspondance to phrases in English, but were unseen or rare in the original German text. For example, in the previous example, we might derive a correspondance between *werde aushaendigen* and *will pass on* that was not possible before reordering. Another example concerns verb-particle constructions, for example in

> Wir machen die Tuer auf

machen and *auf* form a verb-particle construction. The reordering stage moves *auf* to precede *machen*, allowing a phrasal entry that "auf machen" is translated to *to open* in English. Without the reordering, the particle can be arbitrarily far from the verb that it modifies, and there is a danger in this example of translating *machen* as *to make*, the natural translation when no particle is present.

Original sentence: Ich werde Ihnen die entsprechenden Anmerkungen aushaendigen, damit Sie das eventuell bei der Abstimmung uebernehmen koennen. (*I will to you the corresponding comments pass on, so that you them perhaps in the vote adopt can.*)

Reordered sentence: Ich werde aushaendigen Ihnen die entsprechenden Anmerkungen, damit Sie koennen uebernehmen das eventuell bei der Abstimmung.
(*I will pass on to you the corresponding comments, so that you can adopt them perhaps in the vote.*)

Figure 2: An example of the reordering process, showing the original German sentence and the sentence after reordering.

3 Clause Restructuring

We now describe the method we use for reordering German sentences. As a first step in the reordering process, we parse the sentence using the parser described in (Dubey and Keller, 2003). The second step is to apply a sequence of rules that reorder the German sentence depending on the parse tree structure. See Figure 2 for an example German sentence before and after the reordering step.

In the reordering phase, each of the following six restructuring steps were applied to a German parse tree, in sequence (see table 1 also, for examples of the reordering steps):

[1] Verb initial In any verb phrase (i.e., phrase with label `VP-...`) find the head of the phrase (i.e., the child with label `-HD`) and move it into the initial position within the verb phrase. For example, in the parse tree in Figure 1, *aushaendigen* would be moved to precede *Ihnen* in the first verb phrase (`VP-OC`), and *uebernehmen* would be moved to precede *das* in the second `VP-OC`. The subordinate clause would have the following structure after this transformation:

```
S-MO KOUS-CP  damit
     PPER-SB  Sie
     VP-OC VVINF-HD  uebernehmen
           PDS-OA  das
           ADJD-MO  eventuell
           PP-MO APPR-DA  bei
                 ART-DA  der
                 NN-NK  Abstimmung
     VMFIN-HD  koennen
```

[2] Verb 2nd In any subordinate clause labelled `S-...`, with a complementizer `KOUS`, `PREL`, `PWS` or `PWAV`, find the head of the clause, and move it to directly follow the complementizer.

For example, in the subordinate clause in Figure 1, the head of the clause *koennen* would be moved to follow the complementizer *damit*, giving the following structure:

```
S-MO KOUS-CP  damit
     VMFIN-HD  koennen
     PPER-SB  Sie
     VP-OC VVINF-HD  uebernehmen
           PDS-OA  das
           ADJD-MO  eventuell
           PP-MO APPR-DA  bei
                 ART-DA  der
                 NN-NK  Abstimmung
```

[3] Move Subject For any clause (i.e., phrase with label `S...`), move the subject to directly precede the head. We define the subject to be the left-most child of the clause with label `...-SB` or `PPER-EP`, and the head to be the leftmost child with label `...-HD`.

For example, in the subordinate clause in Figure 1, the subject *Sie* would be moved to precede *koennen*, giving the following structure:

```
S-MO KOUS-CP  damit
     PPER-SB  Sie
     VMFIN-HD  koennen
     VP-OC VVINF-HD  uebernehmen
           PDS-OA  das
           ADJD-MO  eventuell
           PP-MO APPR-DA  bei
                 ART-DA  der
                 NN-NK  Abstimmung
```

[4] Particles In verb particle constructions, move the particle to immediately precede the verb. More specifically, if a finite verb (i.e., verb tagged as `VVFIN`) and a particle (i.e., word tagged as `PTKVZ`) are found in the same clause, move the particle to precede the verb.

As one example, the following clause contains both a verb (*forden*) as well as a particle (*auf*):

```
S PPER-SB   Wir
  VVFIN-HD  fordern
  NP-OA ART das
        NN  Praesidium
  PTKVZ-SVP auf
```

After the transformation, the clause is altered to:

```
S PPER-SB   Wir
  PTKVZ-SVP auf
  VVFIN-HD  fordern
  NP-OA ART das
        NN  Praesidium
```

Transformation	Example
Verb Initial	Before: Ich werde Ihnen die entsprechenden Anmerkungen **aushaendigen**, . . . After: Ich werde **aushaendigen** Ihnen die entsprechenden Anmerkungen, . . . I shall be passing on to you some comments, . . .
Verb 2nd	Before: . . . damit Sie uebernehmen das eventuell bei der Abstimmung **koennen**. After: . . . damit **koennen** Sie uebernehmen das eventuell bei der Abstimmung . . . . so that could you adopt this perhaps in the voting.
Move Subject	Before: . . . damit koennen **Sie** uebernehmen das eventuell bei der Abstimmung. After: . . . damit **Sie** koennen uebernehmen das eventuell bei der Abstimmung . . . . so that you could adopt this perhaps in the voting.
Particles	Before: Wir fordern das Praesidium **auf**, . . . After: Wir **auf** fordern das Praesidium, . . . We ask the Bureau, . . .
Infinitives	Before: Ich werde der Sache **nachgehen** dann, . . . After: Ich werde **nachgehen** der Sache dann, . . . I will look into the matter then, . . .
Negation	Before: Wir konnten einreichen es **nicht** mehr rechtzeitig, . . . After: Wir konnten **nicht** einreichen es mehr rechtzeitig, . . . We could not hand it in in time, . . .

Table 1: Examples for each of the reordering steps. In each case the item that is moved is underlined.

[5] Infinitives In some cases, infinitival verbs are still not in the correct position after transformations [1]–[4]. For this reason we add a second step that involves infinitives. First, we remove all internal `VP` nodes within the parse tree. Second, for any clause (i.e., phrase labeled `S...`), if the clause dominates both a finite and infinitival verb, and there is an argument (i.e., a subject, or an object) between the two verbs, then the infinitive is moved to directly follow the finite verb.

As an example, the following clause contains an infinitival (*einreichen*) that is separated from a finite verb *konnten* by the direct object *es*:

```
S PPER-SB  Wir
  VMFIN-HD  konnten
  PPER-OA  es
  PTKNEG-NG  nicht
  VP-OC VVINF-HD  einreichen
        AP-MO ADV-MO  mehr
              ADJD-HD  rechtzeitig
```

The transformation removes the `VP-OC`, and moves the infinitive, giving:

```
S PPER-SB  Wir
  VMFIN-HD  konnten
  VVINF-HD  einreichen
  PPER-OA  es
  PTKNEG-NG  nicht
  AP-MO ADV-MO  mehr
        ADJD-HD  rechtzeitig
```

[6] Negation As a final step, we move negative particles. If a clause dominates both a finite and infinitival verb, as well as a negative particle (i.e., a word tagged as `PTKNEG`), then the negative particle is moved to directly follow the finite verb.

As an example, the previous example now has the negative particle *nicht* moved, to give the following clause structure:

```
S PPER-SB  Wir
  VMFIN-HD  konnten
  PTKNEG-NG  nicht
  VVINF-HD  einreichen
  PPER-OA  es
  AP-MO ADV-MO  mehr
        ADJD-HD  rechtzeitig
```

4 Experiments

This section describes experiments with the reordering approach. Our baseline is the phrase-based MT system of (Koehn et al., 2003). We trained this system on the Europarl corpus, which consists of 751,088 sentence pairs with 15,256,792 German words and 16,052,269 English words. Translation performance is measured on a 2000 sentence test set from a different part of the Europarl corpus, with average sentence length of 28 words.

We use BLEU scores (Papineni et al., 2002) to measure translation accuracy. We applied our re-

	Annotator 2		
Annotator 1	R	B	E
R	33	2	5
B	2	13	5
E	9	4	27

Table 2: Table showing the level of agreement between two annotators on 100 translation judgements. **R** gives counts corresponding to translations where an annotator preferred the reordered system; **B** signifies that the annotator preferred the baseline system; **E** means an annotator judged the two systems to give equal quality translations.

ordering method to both the training and test data, and retrained the system on the reordered training data. The BLEU score for the new system was 26.8%, an improvement from 25.2% BLEU for the baseline system.

4.1 Human Translation Judgements

We also used human judgements of translation quality to evaluate the effectiveness of the reordering rules. We randomly selected 100 sentences from the test corpus where the English reference translation was between 10 and 20 words in length.[1] For each of these 100 translations, we presented the two annotators with three translations: the reference (human) translation, the output from the baseline system, and the output from the system with reordering. No indication was given as to which system was the baseline system, and the ordering in which the baseline and reordered translations were presented was chosen at random on each example, to prevent ordering effects in the annotators' judgements. For each example, we asked each of the annotators to make one of two choices: 1) an indication that one translation was an improvement over the other; or 2) an indication that the translations were of equal quality.

Annotator 1 judged 40 translations to be improved by the reordered model; 40 translations to be of equal quality; and 20 translations to be worse under the reordered model. Annotator 2 judged 44 translations to be improved by the reordered model; 37 translations to be of equal quality; and 19 translations to be worse under the reordered model. Table 2 gives figures indicating agreement rates between the annotators. Note that if we only consider preferences where both annotators were in agreement (and consider all disagreements to fall into the "equal" category), then 33 translations improved under the reordering system, and 13 translations became worse. Figure 3 shows a random selection of the translations where annotator 1 judged the reordered model to give an improvement; Figure 4 shows examples where the baseline system was preferred by annotator 1. We include these examples to give a qualitative impression of the differences between the baseline and reordered system. Our (no doubt subjective) impression is that the cases in figure 3 are more clear cut instances of translation improvements, but we leave the reader to make his/her own judgement on this point.

4.2 Statistical Significance

We now describe statistical significance tests for our results. We believe that applying significance tests to *Bleu* scores is a subtle issue, for this reason we go into some detail in this section.

We used the sign test (e.g., see page 166 of (Lehmann, 1986)) to test the statistical significance of our results. For a source sentence X, the sign test requires a function $f(X)$ that is defined as follows:

$$f(X) = \begin{cases} + & \text{If reordered system produces a better} \\ & \text{translation for } X \text{ than the baseline} \\ - & \text{If baseline produces a better translation} \\ & \text{for } X \text{ than the reordered system.} \\ = & \text{If the two systems produce equal} \\ & \text{quality translations on } X \end{cases}$$

We assume that sentences X are drawn from some underlying distribution $P(X)$, and that the test set consists of independently, identically distributed (IID) sentences from this distribution. We can define the following probabilities:

$$p_+ = \text{Probability}(f(X) = +) \quad (1)$$

$$p_- = \text{Probability}(f(X) = -) \quad (2)$$

where the probability is taken with respect to the distribution $P(X)$. The sign test has the null hypothesis $H_0 = \{p_+ \leq p_-\}$ and the alternative hypothesis $H_1 = \{p_+ > p_-\}$. Given a sample of n test points $\{X_1, \ldots, X_n\}$, the sign test depends on calculation of the following counts: $c_+ = |\{i : f(X_i) = +\}|$, $c_- = |\{i : f(X_i) = -\}|$,

[1] We chose these shorter sentences for human evaluation because in general they include a single clause, which makes human judgements relatively straightforward.

and $c_0 = |\{i : f(X_i) = 0\}|$, where $|\mathcal{S}|$ is the cardinality of the set $\mathcal{S}$.

We now come to the definition of $f(X)$ — how should we judge whether a translation from one system is better or worse than the translation from another system? A critical problem with *Bleu* scores is that they are a function of *an entire test corpus* and do not give translation scores for single sentences. Ideally we would have some measure $f_R(X) \in \mathbb{R}$ of the quality of the translation of sentence X under the reordered system, and a corresponding function $f_B(X)$ that measures the quality of the baseline translation. We could then define $f(X)$ as follows:

$$\begin{array}{ll} f(X) = + & \text{If } f_R(X) > f_B(X) \\ f(X) = - & \text{If } f_R(X) < f_B(X) \\ f(X) = 0 & \text{If } f_R(X) = f_B(X) \end{array}$$

Unfortunately *Bleu* scores do not give per-sentence measures $f_R(X)$ and $f_B(X)$, and thus do not allow a definition of $f(X)$ in this way. In general the lack of per-sentence scores makes it challenging to apply significance tests to *Bleu* scores.[2]

To get around this problem, we make the following approximation. For any test sentence X_i, we calculate $f(X_i)$ as follows. First, we define s to be the *Bleu* score for the test corpus when translated by the baseline model. Next, we define s_i to be the *Bleu* score when all sentences other than X_i are translated by the baseline model, and where X_i itself is translated by the *reordered* model. We then define

$$\begin{array}{ll} f(X_i) = + & \text{If } s_i > s \\ f(X_i) = - & \text{If } s_i < s \\ f(X_i) = 0 & \text{If } s_i = s \end{array}$$

Note that strictly speaking, this definition of $f(X_i)$ is not valid, as it depends on the entire set of sample points $X_1 \ldots X_n$ rather than X_i alone. However, we believe it is a reasonable approximation to an ideal function $f(X)$ that indicates whether the translations have improved or not under the reordered system. Given this definition of $f(X)$, we found that $c_+ = 1057$, $c_- = 728$, and $c_0 = 215$. (Thus 52.85% of all test sentences had improved translations under the baseline system, 36.4% of all sentences had worse translations, and 10.75% of all sentences had the same quality as before.) If our definition of $f(X)$ was correct, these values for c_+ and c_- would be significant at the level $p \leq 0.01$.

We can also calculate confidence intervals for the results. Define P to be the probability that the reordered system improves on the baseline system, given that the two systems do not have equal performance. The relative frequency estimate of P is $\hat{P} = 1057/(1057 + 728) = 59.2\%$. Using a normal approximation (e.g., see Example 6.17 from (Wasserman, 2004)) a 95% confidence interval for a sample size of 1785 is $\hat{P} \pm 2.3\%$, giving a 95% confidence interval of $[56.9\%, 61.5\%]$ for P.

5 Conclusions

We have demonstrated that adding knowledge about syntactic structure can significantly improve the performance of an existing state-of-the-art statistical machine translation system. Our approach makes use of syntactic knowledge to overcome a weakness of tradition SMT systems, namely long-distance reordering. We pose clause restructuring as a problem for machine translation. Our current approach is based on hand-crafted rules, which are based on our linguistic knowledge of how German and English syntax differs. In the future we may investigate data-driven approaches, in an effort to learn reordering models automatically. While our experiments are on German, other languages have word orders that are very different from English, so we believe our methods will be generally applicable.

Acknowledgements

We would like to thank Amit Dubey for providing the German parser used in our experiments. Thanks to Brooke Cowan and Luke Zettlemoyer for providing the human judgements of translation performance. Thanks also to Regina Barzilay for many helpful comments on an earlier draft of this paper. Any remaining errors are of course our own. Philipp Koehn was supported by a grant from NTT, Agmt. dtd. 6/21/1998. Michael Collins was supported by NSF grants IIS-0347631 and IIS-0415030.

[2] The lack of per-sentence scores means that it is not possible to apply standard statistical tests such as the sign test or the t-test (which would test the hypothesis $E[f_R(X)] > E[f_B(X)]$, where $E[.]$ is the expected value under X). Note that previous work (Koehn, 2004; Zhang and Vogel, 2004) has suggested the use of bootstrap tests (Efron and Tibshirani, 1993) for the calculation of confidence intervals for Bleu scores. (Koehn, 2004) gives empirical evidence that these give accurate estimates for Bleu statistics. However, correctness of the bootstrap method relies on some technical properties of the statistic (e.g., Bleu scores) being used (e.g., see (Wasserman, 2004) theorem 8.3); (Koehn, 2004; Zhang and Vogel, 2004) do not discuss whether Bleu scores meet any such criteria, which makes us uncertain of their correctness when applied to Bleu scores.

R:	the current difficulties should encourage us to redouble our efforts to promote cooperation in the euro-mediterranean framework.
C:	the current problems should spur us to intensify our efforts to promote cooperation within the framework of the europa-mittelmeerprozesses.
B:	the current problems should spur us, our efforts to promote cooperation within the framework of the europa-mittelmeerprozesses to be intensified.
R:	propaganda of any sort will not get us anywhere.
C:	with any propaganda to lead to nothing.
B:	with any of the propaganda is nothing to do here.
R:	yet we would point out again that it is absolutely vital to guarantee independent financial control.
C:	however, we would like once again refer to the absolute need for the independence of the financial control.
B:	however, we would like to once again to the absolute need for the independence of the financial control out.
R:	i cannot go along with the aims mr brok hopes to achieve via his report.
C:	i cannot agree with the intentions of mr brok in his report persecuted.
B:	i can intentions, mr brok in his report is not agree with.
R:	on method, i think the nice perspectives, from that point of view, are very interesting.
C:	what the method is concerned, i believe that the prospects of nice are on this point very interesting.
B:	what the method, i believe that the prospects of nice in this very interesting point.
R:	secondly, without these guarantees, the fall in consumption will impact negatively upon the entire industry.
C:	and, secondly, the collapse of consumption without these guarantees will have a negative impact on the whole sector.
B:	and secondly, the collapse of the consumption of these guarantees without a negative impact on the whole sector.
R:	awarding a diploma in this way does not contravene uk legislation and can thus be deemed legal.
C:	since the award of a diploms is not in this form contrary to the legislation of the united kingdom, it can be recognised as legitimate.
B:	since the award of a diploms in this form not contrary to the legislation of the united kingdom is, it can be recognised as legitimate.
R:	i should like to comment briefly on the directive concerning undesirable substances in products and animal nutrition.
C:	i would now like to comment briefly on the directive on undesirable substances and products of animal feed.
B:	i would now like to briefly to the directive on undesirable substances and products in the nutrition of them.
R:	it was then clearly shown that we can in fact tackle enlargement successfully within the eu 's budget.
C:	at that time was clear that we can cope with enlargement, in fact, within the framework drawn by the eu budget.
B:	at that time was clear that we actually enlargement within the framework able to cope with the eu budget, the drawn.

Figure 3: Examples where annotator 1 judged the reordered system to give an improved translation when compared to the baseline system. Recall that annotator 1 judged 40 out of 100 translations to fall into this category. These examples were chosen at random from these 40 examples, and are presented in random order. **R** is the human (reference) translation; **C** is the translation from the system with reordering; **B** is the output from the baseline system.

References

Alshawi, H. (1996). Head automata and bilingual tiling: Translation with minimal representations (invited talk). In *Proceedings of ACL 1996*.

Berger, A. L., Pietra, S. A. D., and Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–69.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–313.

Charniak, E., Knight, K., and Yamada, K. (2003). Syntax-based language models for statistical machine translation. In *Proceedings of the MT Summit IX*.

Dubey, A. and Keller, F. (2003). Parsing german with sister-head dependencies. In *Proceedings of ACL 2003*.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Springer-Verlag.

Galley, M., Hopkins, M., Knight, K., and Marcu, D. (2004). What's in a translation rule? In *Proceedings of HLT-NAACL 2004*.

Gildea, D. (2003). Loosely tree-based alignment for machine translation. In *Proceedings of ACL 2003*.

Graehl, J. and Knight, K. (2004). Training tree transducers. In *Proceedings of HLT-NAACL 2004*.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*.

Koehn, P. and Knight, K. (2003). Feature-rich statistical translation of noun phrases. In Hinrichs, E. and Roth, D., editors, *Proceedings of ACL 2003*, pages 311–318.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase based translation. In *Proceedings of HLT-NAACL 2003*.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses (Second Edition)*. Springer-Verlag.

R:	on the other hand non-british hauliers pay nothing when travelling in britain.
C:	on the other hand, foreign kraftverkehrsunternehmen figures anything if their lorries travelling through the united kingdom.
B:	on the other hand, figures foreign kraftverkehrsunternehmen nothing if their lorries travel by the united kingdom.
R:	i think some of the observations made by the consumer organisations are included in the commission ’s proposal.
C:	i think some of these considerations, the social organisations will be addressed in the commission proposal.
B:	i think some of these considerations, the social organisations will be taken up in the commission ’s proposal.
R:	during the nineties the commission produced several recommendations on the issue but no practical solutions were found.
C:	in the nineties, there were a number of recommendations to the commission on this subject to achieve without, however, concrete results.
B:	in the 1990s, there were a number of recommendations to the commission on this subject without, however, to achieve concrete results.
R:	now, in a panic, you resign yourselves to action.
C:	in the current paniksituation they must react necessity.
B:	in the current paniksituation they must of necessity react.
R:	the human aspect of the whole issue is extremely important.
C:	the whole problem is also a not inconsiderable human side.
B:	the whole problem also has a not inconsiderable human side.
R:	in this area we can indeed talk of a european public prosecutor.
C:	and we are talking here, in fact, a european public prosecutor.
B:	and here we can, in fact speak of a european public prosecutor.
R:	we have to make decisions in nice to avoid endangering enlargement, which is our main priority.
C:	we must take decisions in nice, enlargement to jeopardise our main priority.
B:	we must take decisions in nice, about enlargement be our priority, not to jeopardise.
R:	we will therefore vote for the amendments facilitating its use.
C:	in this sense, we will vote in favour of the amendments which, in order to increase the use of.
B:	in this sense we vote in favour of the amendments which seek to increase the use of.
R:	the fvo mission report mentioned refers specifically to transporters whose journeys originated in ireland.
C:	the quoted report of the food and veterinary office is here in particular to hauliers, whose rushed into shipments of ireland.
B:	the quoted report of the food and veterinary office relates in particular, to hauliers, the transport of rushed from ireland.

Figure 4: Examples where annotator 1 judged the reordered system to give a worse translation than the baseline system. Recall that annotator 1 judged 20 out of 100 translations to fall into this category. These examples were chosen at random from these 20 examples, and are presented in random order. **R** is the human (reference) translation; **C** is the translation from the system with reordering; **B** is the output from the baseline system.

Marcu, D. and Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP 2002*.

Melamed, I. D. (2004). Statistical machine translation by parsing. In *Proceedings of ACL 2004*.

Niessen, S. and Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of ACL 2003*.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., and Radev, D. (2004). A smorgasbord of features for statistical machine translation. In *Proceedings of HLT-NAACL 2004*.

Och, F. J., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proceedings of EMNLP 1999*, pages 20–28.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

Shen, L., Sarkar, A., and Och, F. J. (2004). Discriminative reranking for machine translation. In *Proceedings of HLT-NAACL 2004*.

Wasserman, L. (2004). *All of Statistics*. Springer-Verlag.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3).

Xia, F. and McCord, M. (2004). Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of Coling 2004*.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of ACL 2001*.

Zhang, Y. and Vogel, S. (2004). Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.