Estimating Word Alignment Quality for SMT Reordering Tasks

Sara Stymne  Jörg Tiedemann  Joakim Nivre
Uppsala University
Department of Linguistics and Philology
firstname.lastname@lingfil.uu.se

Abstract
Previous studies of the effect of word alignment on translation quality in SMT generally explore link level metrics only and mostly do not show any clear connections between alignment and SMT quality. In this paper, we specifically investigate the impact of word alignment on two pre-reordering tasks in translation, using a wider range of quality indicators than previously done. Experiments on German–English translation show that reordering may require alignment models different from those used by the core translation system. Sparse alignments with high precision on the link level, for translation units, and on the subset of crossing links, like intersected HMM models, are preferred. Unlike SMT performance the desired alignment characteristics are similar for small and large training data for the pre-reordering tasks. Moreover, we confirm previous research showing that the fuzzy reordering score is a useful and cheap proxy for performance on SMT reordering tasks.

1 Introduction
Word alignment is a key component in all state-of-the-art statistical machine translation (SMT) systems, and there has been some work exploring the connection between word alignment quality and translation quality (Och and Ney, 2003; Fraser and Marcu, 2007; Lambert et al., 2012). The standard way to evaluate word alignments in this context is by using metrics like alignment error rate (AER) and F-measure on the link level, and the general conclusion appears to be that translation quality benefits from alignments with high recall (rather than precision), at least for large training data. Although many other ways of measuring alignment quality have been proposed, such as working on translation units (Ahrenberg et al., 2000; Ayan and Dorr, 2006; Søgaard and Kuhn, 2009) or using link degree and related measures (Ahrenberg, 2010), these methods have not been used to study the relation between alignment and translation quality, with the exception of Lambert et al. (2012).

Word alignment is also used for many other tasks besides translation, including term bank creation (Merkel and Foo, 2007), cross-lingual annotation projection for part-of-speech tagging (Yarowsky et al., 2001), semantic roles (Pado and Lapata, 2005), pronoun anaphora (Postolache et al., 2006), and cross-lingual clustering (Täckström et al., 2012). Even within SMT itself, there are tasks such as reordering that often make crucial use of word alignments. For instance, source language reordering commonly relies on rules learnt automatically from word-aligned data (e.g., Xia and McCord (2004)). As far as we know, no one has studied the impact of alignment quality on these additional tasks, and it seems to be tacitly assumed that alignments that are good for translation are also good for other tasks.

In this paper we set out to explore the impact of alignment quality on two pre-reordering tasks for SMT. In doing so, we employ a wider range of quality indicators than is customary, and for reference these indicators are used also to assess overall translation quality. To allow an in-depth exploration of the connections between several aspects of word alignment and reordering, we limit our study to one language pair, German–English. We think this is a suitable language pair for studying reordering since it has both short range and long range reorderings. Our main focus is on using relatively large training data, 2M sentences, but we also report results with small training data, 170K sentences. The main conclusion of our study is that alignments that are optimal for translation are not necessarily optimal for reordering, where pre-
cision is of greater importance than recall. For SMT the best alignments are different depending on corpus size, but for the reordering tasks results are stable across training data size.

In section 2 we discuss previous work related to word alignment and SMT. In section 3, we introduce the word alignment quality indicators we use, and show experimental results for a number of alignment systems on an SMT task. In section 4, we turn to reordering for SMT and use the same quality indicators to study the impact of alignment quality on reordering quality. In section 5 we briefly describe results using small training data. In section 6, we conclude and suggest directions for future work.

2 Word Alignment and SMT

Word alignment is the task of relating words in one language to words in the translation in another language, see an example in Figure 1. Word alignment models can be learnt automatically from large corpora of sentence aligned data. Brown et al. (1993) proposed the so-called IBM models, which are still widely used. These five models estimate alignments from corpora using the expectation-maximization algorithm, and each model adds some complexity. Model 4 is commonly used in SMT systems. There have been many later suggestions of alternatives to these models. These are often alternatives to model 2, such as the HMM model (Vogel et al., 1996) and fast_align (Dyer et al., 2013).

All these generative models produce directional alignments where one word in the source can be linked to many target words (1–m links) but not vice versa. It is generally desirable to also allow n–1 and n–m links, and to achieve this it is common practice to perform word alignment in both directions and to symmetrize them using some heuristic. A number of common symmetrization strategies are described in Table 1 (Koehn et al., 2005). There are also other alternatives, such as the refined method (Och and Ney, 2003), or link deletion from the union (Fossum et al., 2008).

There is also a wide range of alternative approaches to word alignment. For example, various discriminative models have been proposed in the literature (Liu et al., 2005; Moore, 2005; Taskar et al., 2005). Their advantage is that they may integrate a wide range of features that may lead to improved alignment quality. However, most of these models require external tools (for creating linguistic features) and manually aligned training data, which we do not have for our data sets (besides the data we need for evaluation). Investigating these types of models are outside the scope of our current work.

Word alignments are used as an important knowledge source for training SMT systems. In word-based SMT, the parameters of the generative word alignment models are essentially the translation model of the system. In phrase-based SMT (PBSMT) (Koehn et al., 2003), which is among the state-of-the-art systems today, word alignments are used as a basis for extracting phrases and estimating phrase alignment probabilities. Similarly, word alignments are also used for estimating rule probabilities in various kinds of hierarchical and syntactic SMT (Chiang, 2007; Yamada and Knight, 2002; Galley et al., 2004).

Intrinsic evaluation of word alignment is generally based on a comparison to a gold standard of human alignments. Based on the gold standard, metrics like precision, recall and F-measure can be calculated for each alignment link, see Eqs. 1–2, where A are hypothesized alignment links and G are gold standard links. Another common metric is alignment error rate (AER) (Och and Ney, 2000), which is based on a distinction between sure, S, and possible, P, links in the gold standard. 1−AER is identical to balanced F-measure when the gold standard does not make a distinction between S and P.

$$\text{Precision}(A, G) = \frac{|G \cap A|}{|A|}$$  \hspace{1cm} (1)

$$\text{Recall}(A, G) = \frac{|G \cap A|}{|G|}$$  \hspace{1cm} (2)

$$\text{AER} = 1 - \frac{|P \cap A| + |S \cap A|}{|S| + |A|}$$  \hspace{1cm} (3)

<table>
<thead>
<tr>
<th>Symmetrization</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>int: intersection</td>
<td>$A_{TS} \cap A_{ST}$</td>
</tr>
<tr>
<td>uni: union</td>
<td>$A_{TS} \cup A_{ST}$</td>
</tr>
<tr>
<td>gd: grow-diag</td>
<td>intersection plus adjacent links from the union if both linked words are unaligned</td>
</tr>
<tr>
<td>gdf: grow-diag-final</td>
<td>gd with links from the union added in a final step if either linked word is unaligned</td>
</tr>
<tr>
<td>gdfa: grow-diag-final-and</td>
<td>gd with links from the union added in a final step if both linked words are unaligned</td>
</tr>
</tbody>
</table>

Table 1: Symmetrization strategies for word alignments $A_{TS}$ and $A_{ST}$ in two directions
The relation between word alignment quality and PBSMT has been studied by some researchers. Och and Ney (2000) looked at the impact of IBM and HMM models on the alignment template approach (Och et al., 1999) in terms of AER. They found that AER correlates with human evaluation of sentence level quality, but not with word error rate. Fraser and Marcu (2007) found that there is no correlation between AER and Bleu (Papineni et al., 2002), especially not when the P-set is large. They found that a balanced F-measure is a better indicator of Bleu, but that a weighted F-measure is even better (see Eq. 4) mostly with a higher weight for recall than for precision. This weight, however, needs to be optimized for each data set, language pair, and gold standard alignment separately.

Ayan and Dorr (2006) on the other hand found some evidence for the importance of precision over recall. However, they used much smaller training data than Fraser and Marcu (2007). They also suggested using a measure called consistent phrase error-rate (CPER), but found that it was hard to assess the impact of alignment on MT, both with AER and CPER. Lambert et al. (2012) performed a study where they investigated the effect of word alignment on MT using a large number of word alignment indicators. They found that there was a difference between large and small datasets in that alignment precision was more important with small data sets, and recall more important with large data sets. Overall they did not find any indicator that was significant over two language pairs and different corpus sizes. There were more significant indicators for large datasets, however.

Most researchers who propose new alignment models perform both a gold standard evaluation and an SMT evaluation (Liang et al., 2006; Ganchev et al., 2008; Junczys-Dowmunt and Szal, 2012; Dyer et al., 2013). The relation between the two types of evaluation is often quite weak. Several of these studies only show AER on their gold standard, despite its well-known shortcomings.

Even though many studies have shown some relation between translation quality and AER or weighted F-measure, it has rarely been investigated thoroughly in its own right, and, as far as we are aware, not for other tasks than SMT. Furthermore, most of these studies consider nothing else but link level agreement. In this paper we take a broader view on alignment quality and explore the effect of other types of quality indicators as well.

3 Word Alignment Quality Indicators

We investigate four groups of quality indicators. The first group is the classic group where metrics are calculated on the alignment link level, which has been used in several studies. In our experiments we use a gold standard that does not make use of distinctions between sure and possible links, as suggested by Fraser and Marcu (2007). With this, we can calculate the standard metrics P(recision) R(recall) and F(-measure). We will mainly use balanced F-measure, but occasionally also report weighted F-measure. As noted before, 1−AER is equivalent to balanced F when only sure links are used, and will thus not be reported separately.

Søgaard and Kuhn (2009) and Søgaard and Wu (2009) suggested working on the translation unit (TU) level, instead of the link level. A translation unit, or cept (Goutte et al., 2004), is defined as a maximally connected subgraph of an alignment. In Figure 1, the twelve links form nine translation units. Søgaard and Wu (2009) suggest the metric TUER, translation unit error rate, shown in Eq. 5, where \(A_U\) are hypothesized translation units, and \(G_U\) are gold standard translation units. They use TUER to establish lower bounds for the coverage of alignments from different formalisms, not to evaluate SMT. While they only use TUER, it is similar to CPER (Ayan and Dorr, 2006), which measures the error rate of extracted phrases. Due to how phrase extraction handle null links, there are differences, however.

3.1 Link Level Quality Indicators

The relation between word alignment quality and PBSMT has been studied by some researchers. Och and Ney (2000) looked at the impact of IBM and HMM models on the alignment template approach (Och et al., 1999) in terms of AER. They found that AER correlates with human evaluation of sentence level quality, but not with word error rate. Fraser and Marcu (2007) found that there is no correlation between AER and Bleu (Papineni et al., 2002), especially not when the P-set is large. They found that a balanced F-measure is a better indicator of Bleu, but that a weighted F-measure is even better (see Eq. 4) mostly with a higher weight for recall than for precision. This weight, however, needs to be optimized for each data set, language pair, and gold standard alignment separately.

Ayan and Dorr (2006) on the other hand found some evidence for the importance of precision over recall. However, they used much smaller training data than Fraser and Marcu (2007). They also suggested using a measure called consistent phrase error-rate (CPER), but found that it was hard to assess the impact of alignment on MT, both with AER and CPER. Lambert et al. (2012) performed a study where they investigated the effect of word alignment on MT using a large number of word alignment indicators. They found that there was a difference between large and small datasets in that alignment precision was more important with small data sets, and recall more important with large data sets. Overall they did not find any indicator that was significant over two language pairs and different corpus sizes. There were more significant indicators for large datasets, however.

Most researchers who propose new alignment models perform both a gold standard evaluation and an SMT evaluation (Liang et al., 2006; Ganchev et al., 2008; Junczys-Dowmunt and Szal, 2012; Dyer et al., 2013). The relation between the two types of evaluation is often quite weak. Several of these studies only show AER on their gold standard, despite its well-known shortcomings.

Even though many studies have shown some relation between translation quality and AER or weighted F-measure, it has rarely been investigated thoroughly in its own right, and, as far as we are aware, not for other tasks than SMT. Furthermore, most of these studies consider nothing else but link level agreement. In this paper we take a broader view on alignment quality and explore the effect of other types of quality indicators as well.

3 Word Alignment Quality Indicators

We investigate four groups of quality indicators. The first group is the classic group where metrics are calculated on the alignment link level, which has been used in several studies. In our experiments we use a gold standard that does not make use of distinctions between sure and possible links, as suggested by Fraser and Marcu (2007). With this, we can calculate the standard metrics P(recision) R(recall) and F(-measure). We will mainly use balanced F-measure, but occasionally also report weighted F-measure. As noted before, 1−AER is equivalent to balanced F when only sure links are used, and will thus not be reported separately.

Søgaard and Kuhn (2009) and Søgaard and Wu (2009) suggested working on the translation unit (TU) level, instead of the link level. A translation unit, or cept (Goutte et al., 2004), is defined as a maximally connected subgraph of an alignment. In Figure 1, the twelve links form nine translation units. Søgaard and Wu (2009) suggest the metric TUER, translation unit error rate, shown in Eq. 5, where \(A_U\) are hypothesized translation units, and \(G_U\) are gold standard translation units. They use TUER to establish lower bounds for the coverage of alignments from different formalisms, not to evaluate SMT. While they only use TUER, it is similar to CPER (Ayan and Dorr, 2006), which measures the error rate of extracted phrases. Due to how phrase extraction handle null links, there are differences, however.
is also possible to define Precision, Recall and F-measure over translation units in the same way as for alignment links. We will use these three measures to get a broader picture of TUs in alignment evaluation. Also in this case, 1−TUER is equivalent to F-measure.

\[
\text{TUER}(A,G) = 1 - \frac{2|A_U \cap G_U|}{|A_U| + |G_U|} \tag{5}
\]

The TU metrics are quite strict, since they require exact matching of TUs. Tiedemann (2005) suggested the MWU metrics for word alignment evaluation, which also consider partial matches of annotated multi-word units, which is a similar concept to TUs. In those metrics, precision and recall grow proportionally to the number of correctly aligned words within translation units. Proposed links are in this way scored according to their overlap with translation units in the gold standard. Precision and recall are defined in Eqs. 6–7, where \( \text{overlap}(X_U, Y) \) is the number of source and target words in \( X_U \) that overlap with translation units in \( Y \) normalized by the size of \( X_U \) (in terms of source and target words). Note, that TUs need to overlap in source and target. Otherwise, their overlap will be counted as zero.

\[
P_{MWU} = \sum_{A_U \in A} \frac{\text{overlap}(A_U, G)}{|A|} \tag{6}
\]

\[
R_{MWU} = \sum_{G_U \in G} \frac{\text{overlap}(G_U, A)}{|G|} \tag{7}
\]

There have also been attempts at classifying alignments in other ways, not related to a gold standard. Ahrenberg (2010) proposed several ways to categorize human alignments, including link degree, reordering of links, and structural correspondence. He used these indicators to profile hand-aligned corpora from different domains. We will not use structural correspondence, which requires a dependency parser, and which we believe is error prone when performed automatically. We will use what we call link degree, i.e., how many alignment links each word obtains. Ahrenberg (2010) used a fine-grained scheme of the percentage for different degrees, including isomorphism 1–1, deletion 0–1, reduction m–1, and paraphrase m–n. Similar link degree classes were used by Lambert et al. (2012). In this work we will reduce these classes into three: 1–1 links, null links, which combine the 0–1 and 1–0 cases, and multi links where there are many words on at least one side.

Ahrenberg (2010) also proposed to measure reorderings. He does this by calculating the percentage of links with crossings of different lengths. To define this he only considers adjacent links in the source using the distance between corresponding target words, which means that his metric becomes a directional measure. Reorderings of alignments was also used by Genzel (2010), who used crossing score, the number of crossing links, to rank reordering rules. This is non-directional and simpler to calculate than Ahrenberg (2010)'s metrics, and implicitly covers length since a long distance reordering leads to a higher number of pairwise crossing links. Birch and Osborne (2011) suggest using squared Kendall \( \tau \) distance (SKTD), see Eq. 8, where \( n \) is the number of links, as a basis of LR-score, an MT metric that takes reordering into account. They found that squaring \( \tau \) better explained reordering, than using only \( \tau \). In this study we will use both, crossing score and SKTD. Figure 1 shows these scores for an example sentence. These two measures only tell us how much reordering there is. To quantify this relative to the gold standard we also report the absolute difference between the number of gold standard crossings and system crossings, which we call Crossdiff. To account for the quality of crossings, to some extent, we will also report precision, recall, and F-measure for the subset of translation units that are involved in a crossing.

\[
\text{SKTD} = \sqrt{\frac{|\text{crossing link pairs}|}{(n^2 - n)/2}} \tag{8}
\]

### 3.1 Alignment Experiments

We perform all our experiments for German–English. The alignment indicators are calculated on a corpus of 987 hand aligned sentences (Pado and Lapata, 2005). The gold standard contains explicit null links, which the symmetrized automatic alignments do not. To allow a straightforward comparison we consistently remove all null links when comparing system alignments to the gold standard.

For creating the automatic alignments we used GIZA++ (Och and Ney, 2003) to compute directional alignments for model 2–4 and the HMM model, and fast_align (fa) (Dyer et al., 2013) as newer alternatives to model 2. These models require large amounts of data to be estimated reliably. To achieve this we concatenated the gold standard with the large SMT training data (see
<table>
<thead>
<tr>
<th>Model</th>
<th>Alignment links</th>
<th>Translation units</th>
<th>MWU</th>
<th>Link degree</th>
<th>Link crossings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total P R F</td>
<td>Total P R F</td>
<td>P R F</td>
<td>1-1 null multi</td>
<td>Total SKTD P R F Crossdiff</td>
</tr>
<tr>
<td>gold</td>
<td>22629 .850 .577 .687</td>
<td>17068 .701 .631 .664</td>
<td>.849 .712 .774</td>
<td>.542 .328 .130</td>
<td>30163 .292 – – – 0</td>
</tr>
<tr>
<td>2-int</td>
<td>15362 .850 .577 .687</td>
<td>15362 .701 .631 .664</td>
<td>.849 .712 .774</td>
<td>.542 .328 .130</td>
<td>10064 .267 .551 .463 .503 20099</td>
</tr>
<tr>
<td>3-int</td>
<td>16573 .850 .577 .687</td>
<td>16573 .701 .631 .664</td>
<td>.849 .712 .774</td>
<td>.542 .328 .130</td>
<td>12682 .274 .553 .521 .537 17481</td>
</tr>
<tr>
<td>4-int</td>
<td>16529 .850 .577 .687</td>
<td>16529 .701 .631 .664</td>
<td>.849 .712 .774</td>
<td>.542 .328 .130</td>
<td>11229 .251 .663 .522 .584 18934</td>
</tr>
<tr>
<td>HMM-int</td>
<td>14871 .850 .577 .687</td>
<td>14871 .701 .631 .664</td>
<td>.849 .712 .774</td>
<td>.542 .328 .130</td>
<td>8077 .221 .663 .522 .584 22086</td>
</tr>
<tr>
<td>fa-int</td>
<td>15997 .850 .577 .687</td>
<td>15997 .701 .631 .664</td>
<td>.849 .712 .774</td>
<td>.542 .328 .130</td>
<td>9724 .246 .568 .471 .515 20439</td>
</tr>
<tr>
<td>2-gd</td>
<td>22882 .702 .710 .706</td>
<td>16511 .599 .579 .589</td>
<td>.806 .827 .816</td>
<td>.524 .389 .186</td>
<td>21823 .270 .446 .444 .445 8340</td>
</tr>
<tr>
<td>3-gd</td>
<td>21961 .757 .734 .745</td>
<td>17644 .650 .672 .661</td>
<td>.817 .855 .836</td>
<td>.500 .500 .000</td>
<td>21886 .278 .492 .523 .507 8277</td>
</tr>
<tr>
<td>4-gd</td>
<td>22754 .768 .772 .770</td>
<td>17611 .670 .692 .681</td>
<td>.839 .866 .862</td>
<td>.500 .500 .000</td>
<td>21966 .259 .583 .517 .548 8197</td>
</tr>
<tr>
<td>HMM-gd</td>
<td>19430 .812 .698 .751</td>
<td>18732 .621 .681 .650</td>
<td>.817 .855 .836</td>
<td>.500 .500 .000</td>
<td>17399 .246 .584 .475 .524 15829</td>
</tr>
<tr>
<td>fa-gd</td>
<td>23148 .702 .719 .710</td>
<td>17043 .589 .588 .588</td>
<td>.802 .839 .820</td>
<td>.548 .258 .194</td>
<td>18578 .242 .454 .447 .450 11585</td>
</tr>
<tr>
<td>2-gdfa</td>
<td>23840 .687 .724 .705</td>
<td>17449 .575 .588 .582</td>
<td>.780 .841 .809</td>
<td>.590 .216 .194</td>
<td>25616 .279 .419 .473 .444 6718</td>
</tr>
<tr>
<td>3-gdfa</td>
<td>23049 .734 .749 .742</td>
<td>18732 .621 .681 .650</td>
<td>.780 .841 .809</td>
<td>.590 .216 .194</td>
<td>27119 .294 .451 .561 .500 4547</td>
</tr>
<tr>
<td>4-gdfa</td>
<td>22704 .751 .778 .769</td>
<td>18561 .645 .701 .672</td>
<td>.813 .901 .855</td>
<td>.673 .172 .154</td>
<td>26977 .275 .529 .562 .545 3044</td>
</tr>
<tr>
<td>HMM-gdfa</td>
<td>20554 .799 .726 .761</td>
<td>16955 .685 .681 .683</td>
<td>.857 .851 .854</td>
<td>.655 .337 .098</td>
<td>17399 .246 .584 .475 .524 12764</td>
</tr>
<tr>
<td>fa-gdfa</td>
<td>23148 .702 .719 .710</td>
<td>17043 .589 .588 .588</td>
<td>.802 .839 .820</td>
<td>.548 .258 .194</td>
<td>18578 .242 .454 .447 .450 11585</td>
</tr>
<tr>
<td>3-gdf</td>
<td>26575 .660 .775 .713</td>
<td>18354 .588 .632 .609</td>
<td>.778 .891 .831</td>
<td>.712 .064 .225</td>
<td>50834 .344 .387 .552 .455 20671</td>
</tr>
<tr>
<td>4-gdf</td>
<td>26529 .693 .812 .748</td>
<td>18269 .628 .673 .650</td>
<td>.810 .922 .862</td>
<td>.706 .070 .223</td>
<td>47216 .322 .459 .585 .514 17053</td>
</tr>
<tr>
<td>HMM-gdf</td>
<td>23886 .725 .756 .744</td>
<td>16660 .651 .635 .643</td>
<td>.851 .887 .869</td>
<td>.579 .251 .169</td>
<td>36881 .309 .473 .499 .486 6718</td>
</tr>
<tr>
<td>2-uni</td>
<td>30712 .566 .769 .652</td>
<td>15864 .503 .468 .485</td>
<td>.774 .869 .818</td>
<td>.584 .002 .413</td>
<td>71223 .349 .305 .396 .345 41060</td>
</tr>
<tr>
<td>3-uni</td>
<td>28093 .636 .789 .704</td>
<td>17391 .592 .603 .597</td>
<td>.791 .889 .837</td>
<td>.684 .067 .249</td>
<td>61823 .355 .381 .523 .441 31660</td>
</tr>
<tr>
<td>4-uni</td>
<td>27920 .670 .827 .740</td>
<td>17411 .636 .649 .642</td>
<td>.826 .921 .871</td>
<td>.682 .074 .244</td>
<td>57408 .333 .456 .564 .504 27245</td>
</tr>
<tr>
<td>fa-uni</td>
<td>27951 .612 .756 .676</td>
<td>16385 .512 .491 .504</td>
<td>.781 .867 .822</td>
<td>.548 .111 .346</td>
<td>38285 .396 .336 .407 .368 8122</td>
</tr>
</tbody>
</table>

Table 2: Values for alignment quality indicators for the different alignments, where 2–4, HMM, and fa are alignment models, and symmetrization strategies refer to Table 1.
Section 3.2) of 2M sentences during alignment. For symmetrization we used all methods in Table 1, as implemented in the Moses toolkit (Koehn et al., 2007) and in fast_align (Dyer et al., 2013).

Based on the automatically aligned gold standard, we calculated all alignment indicators for all settings. The complete results can be found in Table 2, where we have ordered the symmetrization methods with the most sparse, intersection, on top. Overall we can see that while several of the alignment methods create a much higher number of alignment links than the gold standard, they do not produce many more translation units. This is very interesting and indicates why link level statistics may not be accurate enough to predict the performance of certain downstream applications. As expected, the metric scores for translation units are lower than for link level metrics. This is partly due to the fact that these measures do not count any partially correct links; the MWU metric which considers partial matches often have higher scores than link level metrics. Another finding is that the number of crossings vary a lot with more than twice as many as the reference for model2+union, and less than three times as many for HMM+intersection. The HMM and fa models have fewer reorderings than the IBM models.

We are now interested in the relation between alignment evaluation on the link level and on the translation unit level, which has not been thoroughly investigated before. Table 3 shows the correlations between the various metrics. Both precision and F-measure at the link level have significant correlations to all TU metrics. Link level recall, on the other hand, is significantly negatively correlated with TU precision, but not significantly correlated to any other TU metric, not even TU recall. Link level precision is thus highly important for matching translation units. We can also note here that while there is a trade-off between precision and recall on link level, this is not the case for translation units, which can have both high precision and high recall. The same is not true for MWU, that allows partial matching, where we also see at least some precision/recall trade-off.

### 3.2 SMT Experiments

For reference, we first study the impact of alignment on SMT performance. Our SMT system is a standard PBSMT system trained on WMT13 data. We trained a German–English system on 2M sentences from Europarl and News Commentary. We used the target side of the parallel corpus and the SRILM toolkit (Stolcke, 2002) to train a 5-gram language model. For training the translation model and for decoding we used the Moses toolkit (Koehn et al., 2007). We applied a standard feature set consisting of a language model feature, four translation model features, word penalty, phrase penalty, and distortion cost. For tuning we used minimum error-rate training (Och, 2003). In order to minimize the risk of tuning influencing the results, we used a fixed set of weights for each experiment, tuned on a model 44+gdfa alignment. For tuning we used newstest2009 with 2525 sentences, and for testing we used newstest2013 with 3000 sentences. Evaluation was performed using the Bleu metric (Papineni et al., 2002). The same system setup was used for the SMT systems with reordering.

Table 4 shows the results on the SMT task. Model 3 and 4 with gd/gdfa symmetrization yield the highest scores. There is a larger difference between systems with different symmetrization than between systems with different alignment models. The sparse intersection symmetrization gives the poorest results. The top row in Table 5 shows correlations between Bleu and all word alignment quality indicators. There are significant correlations with link level recall. A weighted link level F-measure with $\alpha = 0.3$ gives a significant correlation of .72, which confirms the results of Fraser and Marcu (2007). There are no significant correlations with the TU metrics but a positive correlation with the number of TUs. For the MWU metrics the correlations are similar to the link level, significant.

<table>
<thead>
<tr>
<th>Link level</th>
<th>Translation unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.95</td>
</tr>
<tr>
<td>R</td>
<td>-0.57</td>
</tr>
<tr>
<td>F</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 3: Pearson correlations between gold standard word alignment evaluation on the link level and on translation unit level. Significant correlations are marked with bold (< 0.01).

---

2 http://www.statmt.org/wmt13/translation-task.html

3 This could have disfavored the other alignments, so we also performed control experiments where we ran separate tunings for each alignment. While the absolute results varied somewhat, the correlations with alignment indicators were stable.
suggesting that they measure similar things. Intuitively it seems important for SMT to match full translation units, but it might be the case that the phrase extraction strategy is robust as long as there are partial matches. There are no significant correlations with link degree or link crossings, except a negative correlation with Crossdiff, which means that it is good to have a similar number of crossings as the baseline. These results confirm results from previous studies that link level measures, especially recall and weighted F-measure show some correlation with SMT quality whereas precision does not.

### Table 4: Baseline Bleu scores for different symmetrization heuristics

<table>
<thead>
<tr>
<th></th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
<th>HMM</th>
<th>fa</th>
</tr>
</thead>
<tbody>
<tr>
<td>inter</td>
<td>18.1</td>
<td>19.1</td>
<td>19.3</td>
<td>18.8</td>
<td>18.9</td>
</tr>
<tr>
<td>gd</td>
<td>20.4</td>
<td>20.9</td>
<td>20.9</td>
<td>20.5</td>
<td>20.6</td>
</tr>
<tr>
<td>gdfa</td>
<td>20.4</td>
<td>20.7</td>
<td>20.8</td>
<td>20.5</td>
<td>20.5</td>
</tr>
<tr>
<td>gdf</td>
<td>19.4</td>
<td>19.7</td>
<td>20.1</td>
<td>19.9</td>
<td>20.0</td>
</tr>
<tr>
<td>union</td>
<td>19.2</td>
<td>19.6</td>
<td>19.8</td>
<td>19.7</td>
<td>20.0</td>
</tr>
</tbody>
</table>

In the reordering studies cited above it is often not even stated which alignment model was used. A few authors mention the alignment tool that has been applied but no comparison between different alignment models is performed in any of the papers we are aware of. Li et al. (2007), for example, simply state that they used GIZA++ and gd symmetrization and that they removed less probable multi links. Lerner and Petrov (2013) use the intersection of HMM alignments and claims that model 4 did not add much value. Genzel (2010) did mention that using a standard model 4 was not successful for his rule learning approach. Instead he used filtered model-1-alignments, which he claims was more successful. However, there are no further analyses or comparisons between the alignments reported in any of these papers.

### 4 Reordering Tasks for SMT

Reordering is an important part of any SMT system. One way to address it is to add reordering models to standard PBSMT systems, for instance lexicalized reordering models (Koehn et al., 2005), or to directly model reordering in hierarchical (Chiang, 2007) or syntactic translation models (Yamada and Knight, 2002). Another type of approach is preordering, where the source side is reordered to mimic the target side before translation. There have also been approaches where reordering is modeled as part of the evaluation of MT systems (Birch and Osborne, 2011).

We can distinguish two main types of approaches to preordering in SMT, either by using hand-written rules, which often operate on syntactic trees (Collins et al., 2005), or by reordering rules that are learnt automatically based on a word aligned corpus (Xia and McCord, 2004). The latter approach is of interest to us, since it is based on word alignments.

There has been much work on automatic learning of reordering rules, which can be based on different levels of annotation, such as part-of-speech tags (Rottmann and Vogel, 2007; Niehues and Kolss, 2009; Genzel, 2010), chunks (Zhang et al., 2007) or parse trees (Xia and McCord, 2004). In general, all these approaches lead to improvements of translation quality. The reordering is always applied on the translation input. It can also be applied on the source side of the training corpora, which sometimes improves the results (Rottmann and Vogel, 2007), but sometimes does not make a difference (Stymne, 2012). When preordering is performed on the translation input, it can be presented to the decoder as a 1-best reordering (Xia and McCord, 2004), as an n-best list (Li et al., 2007), or as a lattice of possible reorderings (Rottmann and Vogel, 2007; Zhang et al., 2007).

Below we present the two tasks that we study in this paper: part-of-speech-based reordering for creating input lattices for SMT and alignment-based reordering for improving phrase-tables. We evaluate the performance of these tasks in relation to the use of different alignment models and symmetrization heuristics. For these tasks we are mainly interested in the full translation task, for which we report Bleu scores. In addition we also show fuzzy reordering score (FRS), which focuses...
only on the reordering component (Talbot et al., 2011). It compares a system reordering to a reference reordering, by measuring how many chunks that have to be moved to get an identical word order, see Eq. 9, where \( C \) is the number of continguously aligned chunks, and \( M \) the number of words. To find the reference ordering we apply the method of Holmqvist et al. (2009), described in Section 4.2, to the gold standard alignment.

\[
FRS = 1 - \frac{C - 1}{M - 1} \tag{9}
\]

4.1 Part-of-Speech-Based Reordering

Our first reordering task is a part-of-speech-based preordering method described by Rottmann and Vogel (2007) and Niehues and Kolss (2009), which was successfully used for German–English translation. Rules are learnt from a word aligned POS-tagged corpus. Based on the alignments, tag patterns are identified that give rise to specific reorderings. These patterns are then scored based on relative frequency.\(^4\) The rules are then applied to the translation input to create a reordering lattice, with normalized edge scores based on rule scores. In our experiments we only use rules with a score higher than 0.2, to limit the size of the lattices. For calculating FRS, we pick the highest scoring 1-best word order from the lattices.

We learn rules from our entire SMT training corpus varying alignment models and symmetrization. To investigate only the effect of word alignment for creating reordering rules, we do not change the SMT system, which is trained based on model 4+gdfa alignments. The only thing that varies for the translation task is thus the input lattice given to this SMT system.

The results are shown in Tables 6 and 7. Most Bleu scores are better than using the same SMT system without preordering, with a Bleu score of 20.8. The results on FRS and Bleu are highly correlated at .94, despite the fact that we use a lattice as SMT input, and the 1-best order for FRS. For both metrics sparse symmetrization like intersection and gd performs best. Model 4 and HMM perform best with similar Bleu scores, but FRS is better for the HMM model.

Table 5 shows the correlations with the word alignment indicators, in the rows labeled SMT, Bleu. There are strong correlations with all TU metrics, contrary to the SMT task. There are also significant correlations with link level precision and bal-
anced F-measure. The correlation with weighted link level F-measure is even higher, .91 for $\alpha = 0.6$. This is an indication that this algorithm is more sensitive to precision than the SMT task. As for the SMT task, the correlation patterns are similar for the MWU metrics as for link level. For link degree, null alignments are correlated, but there is a negative correlation for multi links. The correlations with the number of crossings and SKTD are negative, which means that it is better to have a low number of crossings. This may seem counter-intuitive, but note in Table 1 that many alignments have a much higher number of crossings than the baseline. The precision of the crossing links is highly correlated with performance on this task, while the recall is not. This tells us that it is important that the crossings we find in the alignment are good, but that it is less important that we find all crossings. This makes sense since the rule learner can then learn at least a subset of all existing crossings well.

### 4.2 Reordering for Alignment

In our second reordering task we investigate alignment-based reordering for improving phrase-tables (Holmqvist et al., 2009; Holmqvist et al., 2012). This strategy first performs a word alignment, based on which the source text is reordered to remove all crossings. A second alignment is trained on the reordered data, which is then restored to the original order before training the full SMT system. In Holmqvist et al. (2012) it was shown that this strategy leads to improvements in link level recall and F-measure as well as small translation improvements for English–Swedish. It also led to small improvements for German–English translation.

Similar to the previous experiments, we now vary alignment models and symmetrization that are used for reordering during the first step. The second step is kept the same using model 4+gdfa in order to focus on the reordering step in our comparisons. Tables 8 and 9 show the results of these experiments. In this case the reordering strategy was not successful, always producing lower Bleu scores than the baseline of 20.8. However, there are some interesting differences in these outcomes. On this task as well, FRS and Bleu scores are highly correlated at .89, which was expected, since this method directly uses the reordered data to train phrase tables. For the best systems, the

<table>
<thead>
<tr>
<th></th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
<th>HMM</th>
<th>fa</th>
</tr>
</thead>
<tbody>
<tr>
<td>inter</td>
<td>.583</td>
<td>.604</td>
<td>.669</td>
<td>.654</td>
<td>.598</td>
</tr>
<tr>
<td>gd</td>
<td>.548</td>
<td>.583</td>
<td>.646</td>
<td>.642</td>
<td>.561</td>
</tr>
<tr>
<td>gdfa</td>
<td>.532</td>
<td>.564</td>
<td>.633</td>
<td>.645</td>
<td>.553</td>
</tr>
<tr>
<td>gdf</td>
<td>.422</td>
<td>.482</td>
<td>.571</td>
<td>.574</td>
<td>.474</td>
</tr>
<tr>
<td>union</td>
<td>.395</td>
<td>.455</td>
<td>.552</td>
<td>.545</td>
<td>.452</td>
</tr>
</tbody>
</table>

Table 8: Fuzzy reordering scores for alignment-based reordering for different alignments

<table>
<thead>
<tr>
<th></th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
<th>HMM</th>
<th>fa</th>
</tr>
</thead>
<tbody>
<tr>
<td>inter</td>
<td>19.5</td>
<td>19.5</td>
<td>19.9</td>
<td>20.2</td>
<td>19.4</td>
</tr>
<tr>
<td>gd</td>
<td>19.3</td>
<td>19.5</td>
<td>19.8</td>
<td>20.2</td>
<td>19.3</td>
</tr>
<tr>
<td>gdfa</td>
<td>19.1</td>
<td>19.2</td>
<td>19.6</td>
<td>20.0</td>
<td>19.2</td>
</tr>
<tr>
<td>gdf</td>
<td>18.3</td>
<td>18.2</td>
<td>18.6</td>
<td>19.0</td>
<td>18.9</td>
</tr>
<tr>
<td>union</td>
<td>17.4</td>
<td>17.8</td>
<td>18.4</td>
<td>18.8</td>
<td>18.8</td>
</tr>
</tbody>
</table>

Table 9: Bleu scores for alignment-based reordering for different alignments

FRS scores are higher than for the previous task, see Table 6, which shows that reordering directly based on alignments is easier than learning and applying rules based on them, given suitable alignments. On this task, again, the sparser alignments are the most successful on both tasks. Here, however, the HMM model gives the best Bleu scores, and similar FRS scores to model 4.

Table 5 shows the correlations with the word alignment indicators, in the rows labeled Align-Reo. The correlation patterns are very similar to the previous task. A few more indicators are significantly negatively correlated with alignment-based reordering than with the other reordering tasks and metrics. The performance on our two reordering tasks are significantly correlated at .76. Again alignments with good scores on TU metrics, link level precision and crossing link precision are preferable. For this task, the best correlation with weighted link level F-measure is .86 for $\alpha = 0.8$. Again, we thus see that sparse alignments with high precision on all measures including the crossing subset, are important.

### 5 Small Training Data

Since previous work has suggested that training data size influences the relation between alignment and SMT quality for small and large training data (Lambert et al., 2012), we investigated this issue also for our reordering tasks. We repeated all our experiments on a small dataset, only the News Commentary data from WMT13, with 170K sentences. Due to space constraints we cannot show all results in the paper, but the main findings are
summarized in this section.

To acquire alignment results we realigned the gold standard concatenated with the smaller data, to reflect the actual quality of alignment with a small dataset. As expected the quality scores tend to be lower with less data. Overall the same systems tend to perform good on each metric with the small and large data, even though there is some variation in the ranking between systems. On the SMT task as well, the Bleu scores are lower, as expected. In this case fast_align is doing best followed by model 4 and 3. The best symmetrization is again gd and gdfa. There are also some differences in the correlation profile. Link recall and number of translation units are no longer significantly correlated, whereas the number of crossings and SKTD are. The highest correlation for link level F-measure is .60 for balanced F-measure, showing that precision is equally important to recall with less data.

For the reordering tasks the scores are again lower. The POS-based reorderings again help over the baseline SMT, whereas the alignment-based reordering leads to slightly lower scores. The correlation profile look exactly the same for Bleu for POS-based reordering. FRS for both tasks and Bleu for alignment-based reordering have the same correlation profiles as Bleu for alignment-based reordering on large data. There are thus very small differences in the word alignment quality indicators that are relevant with large and small training data, while there are some differences on the SMT task. For weighted link level F-measure, the highest correlations are found with $\alpha = 0.6-0.7$ on the different metrics, again showing that precision is more important than recall. For FRS on both tasks and Bleu for alignment-based reordering, model4 and HMM with intersection and gd still perform best. For Bleu for POS-based reordering, gdfa and model 3 also give good results.

6 Conclusion and Future Work

We have shown that the best combination of alignment and symmetrization models for SMT are not the best models for reordering tasks in our experimental setting. For SMT, high recall is more important than precision with large training data, while precision and recall are of equal importance with small training data. This finding supports previous research (Fraser and Marcu, 2007; Lambert et al., 2012). Translation unit metrics are not predictive of SMT performance. For the large data condition model 3 and 4 with gd and gdfa symmetrization gave the best results, whereas fast_align with gd and gdfa was best with small training data.

For the two preordering tasks we investigated, however, link level weighted F-measure that gave more weight to precision was important, as well as all TU metrics. It was also important to have high precision for the crossing subset of TUs. Hence, it is more important to reliably find some crossings than to find all crossings. This make sense since the extracted rules or performed reorderings are likely good in such cases, even if we are not able to find all possible reorderings. In conclusion, based on this study, we recommend intersection symmetrization with model 4 and HMM for SMT reordering tasks.

We have studied two relatively different reordering tasks with two training data sizes, but found that they to a large extent prefer the same types of alignments. Moreover, the results on these two reordering tasks correlates strongly with FRS, which is much cheaper to calculate than SMT metrics that may even require retraining of full SMT systems. This is consistent with Talbot et al. (2011) who suggested FRS for preordering tasks. We thus would encourage developers of alignment methods to not only give results for SMT, but also for FRS, as a proxy for reordering tasks. Furthermore, it is also useful to give results on TU metrics in addition to link level metrics to complement the evaluation.

In this paper, we have looked at existing generative alignment and symmetrization models. In future work, we would also like to investigate other models, including the removal of low-confidence links, which has previously been proposed for pre-reordering (Li et al., 2007; Genzel, 2010). Given the results, it also seems motivated to develop or adapt the existing models in general, to better fit the properties of specific auxiliary tasks. Furthermore, we need to validate our findings on other language pairs, especially for non-related languages with even more diverse word order.

Acknowledgments

This work was supported by the Swedish strategic research programme eSSENCE.
References


