Bilingual phrase-to-phrase alignment for arbitrarily-small datasets

Kevin Flanagan 667644@swansea.ac.uk
Department of Languages, Translation and Communication, Swansea University, Swansea SA2 8PP, U.K.

Abstract

This paper presents a novel system for sub-sentential alignment of bilingual sentence pairs, however few, using readily-available machine-readable bilingual dictionaries. Performance is evaluated against an existing gold-standard parallel corpus where word alignments are annotated, showing results that are a considerable improvement on a comparable system and on GIZA++ performance for the same corpus. Since naïve application of the system for N languages would require N(N - 1) dictionaries, it is also evaluated using a pivot language, where only 2(N - 1) dictionaries would be required, with surprisingly similar performance. The system is proposed as an alternative to statistical methods, for use with very small corpora or for 'on-the-fly' alignment.

1. Introduction

The process of extracting phrase pairs from parallel corpora is relevant in several contexts, in particular when inducing a translation model with a Statistical Machine Translation (SMT) system, such as that described by Koehn, Och et al. (2003). Phrase pairs are derived from alignments between parts of a bilingual sentence pair, based on word-alignment probabilities. A related procedure is alignment of parts of parse trees for a bilingual sentence pair, to create models for syntax-based machine translation (MT) (Tinsley, Zhechev et al. 2007, Lavie, Parlikar et al. 2008). What these tasks have in common is a requirement for initial word-level alignment information to be established, typically using a tool such as GIZA++ (Och and Ney 2003). In turn, this requires the corpus to be large to achieve useful results, and above a certain minimum size to produce any results at all.

A different context where alignment below sentence level can be of relevance is when recalling content from Translation Memory (TM) systems, whether in order to propose translations for sub-sentential matches (Simard 2003, Planas 2005) or to identify where to edit inexact sentence-level matches (Kranias and Samiotou 2004, Esplà Gomis, Sánchez Martínez et al. 2011). While alignments of that kind can be achieved using statistical methods like those of GIZA++, the size requirements of those methods restrict their usefulness to cases where TMs are sufficiently large. Pre-training GIZA++ on a separate, large corpus does not give very good results, as shown by Esplà Gomis, Sánchez Martínez et al. (2012).

The system presented here aligns parts of bilingual sentence pairs in isolation, without any training step, so has no minimum size requirement. It is intended to be of particular relevance to TM applications, but may also be used to generate sub-sentential alignments in other contexts where they are required and where the data available is limited. It produces hierarchically-arranged pairs of word spans representing the alignments between sentence parts, including alignment of individual words where possible. Alignments between longer spans can be rendered as word alignments for evaluation or use in other contexts. The system is in

no way intended to be a wholesale replacement for an effective and established tool like GIZA++; rather, it provides an alternative in scenarios where GIZA++ use is not practical.

The remainder of this paper is organized as follows. Section 2 describes related work and Section 3 presents the alignment algorithm. Section 4 provides details on evaluation of the system against a parallel corpus with manually-annotated word alignments, while Sections 5 and 6 draw conclusions and discuss further research.

2. Related work

While a great deal of research has been carried out into word alignment of suitably-large parallel corpora – notably the influential 'IBM Models' described by Brown, Pietra et al. (1993) – fewer approaches suitable for smaller datasets have been described. A system for aligning regularized syntactic structures described by Grishman (1994) makes use of bilingual dictionaries to generate initial candidate alignments which are then used in tandem with the syntactic structures to extract subtree correspondences. Although the approach is applied to a small dataset (73 Spanish-English sentence pairs), it differs from the system described here in requiring syntactic structure information to guide the alignment process. Conversely, the system described by Planas (2005) applies flat rather than tree-structured sentence analysis, assigning part-of-speech (POS) categories to sentence sections, then using those categories to attempt alignment rather than any lexically-based translation resource. This only provides subsentential alignment "as long as the languages processed are parallel enough" (Planas 2005: 5), and requires an analyzer for each language concerned.

The word alignment technique described by Mandreoli, Martoglia et al. (2003) uses neither grammatical nor bilingual dictionary information, but instead applies a number of heuristic approaches to establishing potential points of alignment between sentence tokens – identical punctuation, numbers and proper nouns; similarity of lexical tokens based on Longest Common Subsequence (LCS) – each of which is scored to reflect other factors (e.g. longer LCS scores higher, distant relative token position within sentence scores lower), such that lower-scoring points are discarded before interpolating to produce a final alignment. English-Italian examples are given, where pairings such as 'electric/elettrico' and 'collect/collezionare' illustrate the LCS approach described. Nevertheless, given that it is not unusual for information sequencing to change in translation and that some language pairs exhibit little lexical similarity, this alignment technique seems likely to provide poor results for other languages or text types, and a relatively low accuracy level may be acknowledged where the authors note that "the goal of the word aligner is not to find the rigorous matching between each of the words, but to be able to determine, with good approximation, what target segment a given source segment corresponds to" (Mandreoli, Martoglia et al. 2003: 4).

A system requiring more resources is described by Macken (2010), where sentence pair tokens are first lemmatized and assigned POS tags, by external tools. The sentence pair is then sub-sententially aligned in a two-step process. The first step splits the sentences into 'chunks' using a rule-based chunker, while bilingual dictionaries are used to establish tentative lexical correspondences. Pairs of chunks having a high proportion of POS and lexical correspondence are designated as 'anchor' chunks. The second step then continues chunk alignment around these anchors, using similar methods combined with some additional heuristics. Performance is evaluated against a set of gold-standard alignments created by annotators for the project, using a variety of text types. Performance statistics are provided with several bilingual dictionaries, of which some are statistically induced – though not using the entirety of the parallel corpora to be aligned – but others are derived from existing bilingual dictionary data. Alignment performance against the gold-standard data indicates that the approach produces useful results. While this system could in principle be used with small datasets, it would re-

quire a lemmatizer, POS tagger and rule-based chunker for each language, as well as suitable dictionary data.

An interesting approach to aligning nodes in a bilingual phrase-structure parse tree pair is described by Tinsley, Zhechev et al. (2007). For a parallel corpus, word alignment probabilities are first automatically induced using the Moses toolkit (Koehn, Hoang et al. 2007). For each sentence pair, hypotheses are constructed, each aligning a node in one parse tree with a node in the other parse tree. The hypotheses are scored using the word alignment probabilities, and a 'greedy' algorithm used to select and retain the most probable, while excluding those that then contradict well-formedness criteria in relation to those retained. Under their *span1* strategy, scoring of hypotheses with a node dominating a single terminal is deferred until other hypotheses have been processed. This guides selection in cases such as "where source terminal x most likely translates to target terminal y but there is more than one occurrence of both x and y in a single sentence pair" (Tinsley, Zhechev et al. 2007: 4), and cases where there exist two different target terminals with induced word alignment probabilities for a source terminal x, and x is more correctly aligned to the less-probable of the two. Inducing word alignment probabilities from the corpus to be aligned is not a suitable approach for arbitrarily-small datasets, but a similar use of hypothesis evaluation may be, if another source of word alignment information is used.

In Esplà Gomis, Sánchez Martínez et al. (2012), the term *source of bilingual information* (SBI) is used to denote the different resources that can potentially be exploited by their word alignment method, such as bilingual dictionaries, translation memories, and in particular MT. Each sentence in a pair to be aligned is split into all possible sub-segments up to a given length *L* in words, and the available SBIs are queried for translations of each sub-segment. Where the translation of the sub-segment is found in the translated sentence, a tentative alignment is established between the sub-segment and the occurrence of the translation. Once all such alignments have been established, an alignment score for each word pair is calculated using a formula to measure the *alignment pressure* exerted by the translations found, and final alignments are selected that maximize these scores. Performance is evaluated using three different MT systems in combination as SBIs to align sentence pairs from an existing gold-standard word-aligned dataset of 400 sentence pairs. Results are measured against the gold-standard alignments and specifically compared with results using GIZA++ to align the same data, measured against the same gold-standard alignments, in two different cases, with a GIZA++ baseline trained only on the test dataset, and with GIZA++ pre-trained on a separate, much larger parallel corpus. Precision and recall figures show the system gives better results than pre-trained GIZA++, and results comparable with the GIZA++ baseline, though the authors show that GIZA++ performance deteriorates as the dataset size reduces, while the performance of their system is not in principle affected by having as little as a single sentence pair, to align 'on the fly'. Overall, performance of the system appears promising, though they note that "the weakness of our method is the recall, which may be improved by combining other SBIs" (Esplà Gomis, Sánchez Martínez et al. 2012: 98). Nevertheless, this system is specifically intended for use with arbitrarily small datasets and provides a useful comparison for the system presented in this paper.

3. Alignment algorithm

The algorithm presented here produces hierarchically-arranged pairs of word spans, where each pairing represents an alignment between the source and target words spanned. For a given sentence, spans can enclose smaller spans, but may not partially overlap. Sentences are

first tokenized. In the processing that follows, punctuation tokens are first ignored[1], while all other tokens are considered 'words'. The algorithm then operates in four phases. Firstly, tentative *seed alignments* are established between words in the source and target sentences, using whatever bilingual dictionary resources are available. Secondly, each sentence is divided into all possible word spans of length two or more, and alignment scores are calculated for a subset of source and target span pairings. A 'greedy' selection process then records the highest-scoring pairing as aligned and eliminates pairings involving spans that overlap the recorded spans and seed alignments that contradict the recorded span alignment. (Remaining pairings affected by seed alignment removal are then rescored.) The selection process continues until no further pairings can be recorded. Thirdly, seed alignments whose source and target words occur within recorded span pairings are also recorded as aligned span pairings, as are matching punctuation tokens. Finally, further aligned span pairings are deduced using 'remainder' logic, then any redundant span pairings (aligned spans where the words in both spans are all contained within shorter aligned spans) are removed. The following sections describe these phases in more detail. The system as evaluated in section 4 implements each of these phases, using only the external resources described in section 4.2.

3.1. Generation of seed alignments

To establish tentative word correspondences for use as seed alignments, a variety of external data resources can be used, where each resource is queried using a given word in either sentence in order to retrieve any available translations for that word. Machine-readable bilingual dictionaries are an obvious example, as are lists of lexical probabilities generated with GIZA++ or similar from a separate, larger parallel corpus, while domain-specific terminology databases can also be queried to provide translations in this way, as can MT systems. Where tools are available, the lemmatized or stemmed forms of words can also be used for querying, which may increase the likelihood of retrieving translations.

For a given query word, any retrieved translations are compared with the words in the actual sentence translation. Where a match is found, a seed alignment is recorded between the query word and the translation word(s), and given a probability based on a number of factors, including provenance (fixed values for terminology databases or MT systems; retrieved values if found in lexical probability lists) and whether the query word and/or matching translation word(s) are lemmatized or stemmed forms. If two or more queries for a given word retrieve translations matching the same translation word(s) – such as when there is agreement between separate external resources, or query results for the lemmatized form match those for the original form – only the highest-probability seed alignment is retained. While queries consist of single words, retrieved translations may often consist of multiple words (e.g. translations of French 'compenser' into English may include 'make up for'). Seed alignment information retains the 1-to-n relationship between those words for use when calculating span alignment score, as described below. In addition to external resources, a heuristic is used to establish further seed alignments. Where a word in one sentence exactly matches a word in the translated sentence (possible proper noun or other non-translated item), a lower-probability seed alignment is recorded between them.

3.2. Span pairing selection

Calculating scores for all possible span pairings of all possible spans in two sentences *S* and *T* is a problem of polynomial complexity. Since 1-word spans are excluded, for a sentence *S* of

[1] Initial experimentation found that, with this approach, poorer results were achieved when punctuation tokens were used to generate seed alignments. Comparative results and example cases are omitted here for brevity.

length *m* words, there are *m*(*m* - 1)/2 spans to consider, so when aligning with a sentence *T* of *n* words, there are (*m*(*m* - 1))(*n*(*n* - 1))/4 pairings available. To reduce running time, scores are not calculated for pairings considered 'invalid' based on the relative span lengths and sentence lengths. For example, with *S* of length 15 words and *T* of length 17 words, an alignment between a 2-word span in *S* and a 12-word span in *T* will have very low score using the formulae below. The result of an intuition-based function of span and sentence lengths is used to build a subset of all possible pairings that excludes 'invalid' cases.

Each remaining case consists of a span *s* in *S* and *t* in *T*. In a similar way to Tinsley, Zhechev et al. (2007: 4), the following strings are computed:

$$s_l = s_i \dots s_{ix} \qquad \overline{s_l} = S_1 \dots s_{i-1} s_{ix+1} \dots S_m$$

(1)

$$t_l = t_j \dots t_{jy} \qquad \overline{t_l} = T_1 \dots t_{j-1} t_{jy+1} \dots T_n$$

where $s_i...s_{ix}$ and $t_j...t_{jy}$ denote the spans *s* and *t* respectively, and $S_1...S_m$ and $T_1...T_n$ denote the set of words in *S* and *T* respectively. The score γ for a given span pair (s, t) is computed according to (2).

(2)
$$\gamma(s,t) = \alpha(s_l, t_l) \cdot \alpha(\overline{s_l}, \overline{t_l})$$

Individual string-correspondence scores *α(x,y)* are computed using a selection of the seed alignments between *x* and *y* to create a set of seed alignments *A* as described below. Having established *A* for strings *x* and *y*, and defining the set of words in *x* having seed alignments in *A* as A_x, and the set of words in *y* having seed alignments in *A* as A_y, the score *α(x,y)* is calculated as given in (3).

(3)
$$\alpha(x,y) = \frac{2 \cdot \sum_{k=1}^{n} P(A_k)}{2 \cdot |A| + |\{x_i : A_x \not\ni x_i\}| + |\{y_j : A_y \not\ni y_j\}|}$$

The process of selecting the seed alignments between *x* and *y* for *A* merits some explanation.

Consider the following sentence pair:

(4)
EN: He made good use of the afternoon to make up for lost time by drawing a map.
FR: Il a profité de l'après-midi pour rattraper le temps perdu en faisant un plan.

Suppose the seed alignments shown in Table 1 have been generated (lemmatized forms in parentheses), and for ease of illustration, all have a probability value of 1.0. The score to attribute to a given string pair should be a function of the number of seed alignments between the pair and the number of words concerned. So, the string pair ("lost time", "temps perdu") should score more highly than ("lost time by", "temps perdu"), since neither of the seeds for 'by' (rows 12 and 13 in Table 1) match a word in "temps perdu". A simple calculation method would be to attribute the seed alignment probability to each word in the strings covered by a seed alignment, sum those probabilities, then divide by the total number of words, in this simplified example giving 1.0 for ("lost time", "temps perdu") and 0.8 for ("lost time by", "temps perdu"). However, in this example it is desirable for ("make up for lost time by", "rattraper le temps perdu en") to score higher than ("make up for lost time", "rattraper le temps perdu en faisant"). With that simple calculation method, however, those pairs would score $0.\overline{72}$ and $0.8\overline{1}$, since 'make' has a seed alignment with both 'rattraper' (row 18 in Table 1) and 'faisant' (row 2 in Table 1). To avoid this distortion, when selecting seed alignments to be used for computing string-correspondence scores, a seed alignment *a* may only be added to the set if no seed alignment in the set aligns any words also aligned by *a*. This is referred to herein as the *uniqueness requirement* when selecting seed alignments with which to score a string correspondence.

	EN	FR	Query
1	He	Il	he => il
2	made (make)	faisant (faire)	make => faire
3	of	de	of => de
4	of	en	of => en
5	the	l'	l' => the
6	the	le	le => the
7	afternoon	après-midi	afternoon => après-midi
8	make	faisant (faire)	make => faire
9	for	pour	for => pour
10	lost (lose)	perdu (perdre)	lose => perdre
11	time	temps	time => temps
12	by	de	by => de
13	by	en	by => en
14	a	a	(heuristic)
15	a	un	a => un
16	map	plan	map => plan
17	made (make), use, of, make	profité (profiter)	profiter => make use of
18	made (make), make, up, for	rattraper	rattraper => make up for

Table 1 - example seed alignments for a sentence pair

A further consideration is applied for 1-to-n seed alignments. In this example, it is desirable for ("He made good use", "Il a profité") to score higher than ("the afternoon to make up for", "profité de l'après-midi pour"). With that simple calculation method, those pairs would score $0.\overline{714285}$ and $0.\overline{72}$ respectively, as there is a seed alignment between 'make' and 'profiter', and no weight is given to the component words of the translation of 'profiter' being found together. To address this, the seed alignment probability attributed to each translated word matching a 1-to-n query result is divided by the number of words in that result, then per (2), the divisor is reduced when scoring spans containing multiple words matching the same 1-to-n seed alignment. This is referred to herein as the *grouping adjustment* when selecting seed alignments with which to score a string correspondence.

The set *A* of seed alignments selected to score string pair *(x,y)* is then assembled by gathering all the seed alignments that exist between *x* and *y*, applying the grouping adjustment then selecting the highest-probability seed alignments available that meet the uniqueness requirement.

Once scores for the span pairs have been calculated, zero-scoring pairs are discarded, then the 'greedy' selection procedure continues per **Algorithm 1** *selection*.

Algorithm 1 *selection*

```
while span pairs remain in the list
        if there is a single non-pending span pair with the highest score then
                confirm the span pair
        else if there are tied non-pending highest-scoring span pairs whose spans do not overlap then
                confirm those span pairs
        else if there are tied non-pending highest-scoring span pairs whose overlaps
                                                meet intersection criteria then
                confirm the intersection span pair(s)
        else
                flag the tied highest-scoring span pairs as pending
                flag all other span pairs involving any highest-scoring spans as pending
                if all span pairs are flagged as pending then
                        remove all span pairs
                end if
        end if
end while
```

Algorithm 2 *confirm*

for each span pair provided
    record the span pair and remove from list
    remove all overlapping span pairs
    discard all seed alignments contradicting it
    rescore affected span pairs
    remove all zero-scoring span pairs
end for
reset all pending flags

In a similar way to Tinsley, Zhechev et al. (2007: 3), where there are tied highest-scoring span pairs, they are left 'pending' while lower-scoring span pairs are examined. However, tied highest-scoring span pairs that have no overlaps with other span pairs having the same score are recorded immediately, since it is desirable to record the highest-scoring pairs wherever possible. Furthermore, *intersection criteria* are applied when there are tied highest-scoring span pairs. This is much more likely to happen when using external resources where – unlike lexical probabilities generated by GIZA++ – there is no specific probability value associated with query results, and so seed alignment probabilities are assigned based on the provenance of the results, and therefore have relatively uniform values. For the sentence pair at (4), if the seed alignments generated from whatever resources cause the span pairs ("made good use", "profité de l'après-midi") and ("to make up for lost", "l'après-midi pour rattraper") to have the same score, nothing can be inferred from those two pairs, which in any event are not good-quality alignments. However, if the pairs ("He made good use of the afternoon", "Il a profité de l'après-midi") and ("the afternoon to make up for", "l'après-midi pour rattraper") have the same score, it is undesirable to flag these two good-quality alignments as pending in order to examine lower-scoring alignments which are *a priori* less likely to be good-quality. In this case, unlike the preceding low-quality span pairs, there is a level of agreement between the two pairings, in that both English and French spans share intersecting words. These tied span pairs are then considered to meet the intersection criteria, and the intersection span pair ("l'après-midi", "the afternoon") is recorded immediately, on the basis that words appearing in both highest-scoring span pairs are most likely to be aligned. While not the same procedure, this technique recalls the *similarity template learning* heuristic applied in Cicekli and Güvenir (2001) to translation examples with common sequences.

3.3. Seed alignment confirmation

When no span pairs remain to be recorded as alignments, seed alignments are examined. In increasing order of combined word length (since the shorter the spans, the less chance of ambiguous seed alignments), recorded span pairs are compared with seed alignments. Where a span pair contains a query word that generated only one seed alignment within that span pair, a further span pair is added, aligning the query word that generated the seed alignment with the resulting translation word(s).

3.4. Span pair deduction/reduction

Following span pair alignment as above, it may be possible to deduce further alignments. For example, if the span pair ("lost time by drawing a map", "temps perdu en faisant un plan") has been aligned, and the hierarchy contains child span pairs ("lost time", "temps perdu") and ("a map", "un plan"), then a new span pair ("en faisant", "by drawing") is recorded as aligned. This process is repeated until no further deductions can be made. Thereafter, redundant alignments in the hierarchy are removed. In this case, the aforementioned parent span pair

("lost time by drawing a map", "temps perdu en faisant un plan") is removed, since it is completely expressed by contiguous child span pairs.

4. Evaluation

The aligner is evaluated here by comparing the alignments produced by the algorithm using a given set of resources against a manually-aligned gold standard, first using bilingual dictionaries providing direct translations between source and target languages, then using dictionaries that provide those translations via a pivot language. In each case, GIZA++ is also used to align the corpus, and results are again compared with the gold standard.

4.1. Gold-standard data

The main gold-standard data used for evaluation with direct-translation dictionaries was the English-Spanish word-aligned data from the 'tagged EPPS corpus' distributed for TC-STAR 2006 evaluation (Lambert, De Gispert et al. 2005), consisting of 400 sentence pairs drawn from the Europarl corpus (Koehn 2005). In that data, only alignments between individual tokens can be recorded, and what might be considered alignments between spans of words are represented by recording a word alignment between every word in one of the spans and every word in the other. The start of this sentence pair provides an example:

EN: Here in Parliament, we have [...]
ES: En esta Asamblea hemos […]

Between the words 'Here in Parliament' and 'En esta Asamblea', there are a total of nine alignments annotated, linking each of the three words with each of the three translation words. The data also distinguishes between word alignments that are 'sure' and those that are 'possible'. Of those nine alignments, there are two that are annotated as 'sure', aligning ('in', 'En') and ('Parliament', 'Asamblea'). The system under evaluation also generates alignments between spans of words, but records them as distinct spans rather than multiply-linked word alignments. In order to produce alignment data from the system that could be compared more readily with the gold-standard data, the alignment data produced was first subject to an automated conversion. For each aligned span pair, each word not subject to a child span alignment was aligned to each such word in the other span.

For the purposes of evaluating aligner performance for a different language pair with direct-translation dictionaries, 20 English sentences were taken from the tagged EPPS corpus and paired with their German translations in the Europarl corpus from which the English sentences were originally drawn. These English-German sentences pairs were then manually word-aligned by a single annotator using the same principles as applied for the English-Spanish tagged EPPS corpus, to create a small English-German manually-annotated gold-standard corpus against which to test the system.

In order to have gold-standard data against which to evaluate the system operating with pivot-language dictionaries, 20 Spanish sentences were also taken from the tagged EPPS corpus, and paired with their French translations in the Europarl corpus from which the Spanish sentences were originally drawn. These French-Spanish sentence pairs were also then manually word-aligned by a single annotator using the same principles as applied for the English-Spanish tagged EPPS corpus, to create a small French-Spanish manually-annotated gold-standard corpus against which to test the system.

4.2. External resources

For this evaluation, the external resources used were lemmatizers for English, Spanish and French, and a number of machine-readable bilingual dictionaries: French-English (20,403 entries), English-French (21,498 entries), German-English (127,879 entries), English-German (121,380), Spanish-English (17,243 entries) and English-Spanish (20,820 entries). Lemmatizers used were: for English, Morfologik 1.6[2]; for Spanish, Freeling 3.0 (Carreras, Chao et al. 2004); for French, a purpose-built lemmatizer using the data from the Morphalou project (Romary, Salmon-Alt et al. 2004); for German, a purpose-built lemmatizer using the data distributed with the Morphy analysis tool (Lezius, 2000). Dictionaries used were all from the XML Dictionary Exchange Format project[3]. (The files exhibited some corruption and omission, repaired manually.)

4.3. Metrics

Precision and recall were computed versus the gold-standard corpora for the alignments produced both by the system presented here and by GIZA++. These were then combined to obtain the F-measure. These three metrics were computed in two ways, for only the 'sure' alignments in the gold standard, and for all alignments in the gold standard.

4.4. GIZA++

Alignments for the test corpora were also produced using GIZA++, running it in both directions (source to target and target to source) then combining both sets of alignments using the *grow-diag-final-and* heuristic (Koehn, Och et al. 2003).

4.5. Results

Table 2 shows the results obtained aligning the English-Spanish gold-standard corpus of 400 sentence pairs using the system described above with direct-translation dictionaries, compared with the results reported for the same corpus using the system described by Esplà Gomis, Sánchez Martínez et al. (2012) and with alignment of the same corpus using GIZA++.

Alignment type	**Dictionary-based aligner**			**GIZA++**			**Esplà Gomis et al**		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
'sure' only	76.2%	61.3%	66.6%	58.5%	65.1%	60.8%	68.5%	57.6%	62.6%
all	81.4%	46.0%	57.6%	66.7%	52.4%	57.8%	75.7%	43.9%	55.6%

Table 2: Precision (*P*), recall (*R*) and F-measure (*F*) produced for 'sure' alignments, and separately for all alignments, when aligning the EN-ES gold-standard corpus of 400 sentence pairs.

The results show that the system described in this paper produces significantly higher precision than GIZA++, with slightly lower recall for 'sure' alignments and more noticeably lower recall for 'sure' and 'possible' alignments taken together. This raises interesting questions about which metric and which alignment type is of more importance for a given application. For use in TM as described by Simard (2003) and Planas (2005), alignment quality will have a direct bearing on translation suggestions recalled from the TM. In that context, high precision is arguably of some importance, since the lower the precision, the more 'noise' there is likely to be in the results, undesirably distracting the translator. The same consideration applies to the similarity coefficient threshold used to recall TM 'fuzzy matches', where "users

[2] http://sourceforge.net/projects/morfologik/files/morfologik/
[3] http://sourceforge.net/projects/xdxf/

are generally advised not to set the similarity coefficient too low, to avoid being swamped by dissimilar and irrelevant examples" (Macklovitch and Russell 2000: 4). For similar reasons, matches from 'sure' alignments are more likely to be of immediate use than 'possible' alignments. The results also show that the system described here produces higher recall and significantly higher precision than that achieved by Esplà Gomis, Sánchez Martínez et al. (2012) using the same corpus.

Nevertheless, an alignment system using bilingual dictionaries may be of less use in a TM or other translation context if for use with N languages, N(N - 1) bilingual dictionaries are required. This could be reduced to 2(N - 1) dictionaries if the system can be used with a pivot language. Table 3 shows the results from using the system to align the small French-Spanish corpus described above in this way, specifically, by 'chaining' translations from the French-English dictionary to the English-Spanish dictionary to act as a French-Spanish dictionary, and similarly combining the Spanish-English and English-French dictionaries to act as a Spanish-French dictionary.

Alignment type	**Dictionary-based aligner**			**GIZA++**		
	P	*R*	*F*	*P*	*R*	*F*
'sure' only	72.0%	67.2%	68.8%	34.1%	42.7%	37.7%
All	81.0%	58.2%	66.4%	37.7%	36.1%	36.5%

Table 3: Precision (*P*), recall (*R*) and F-measure (*F*) produced for 'sure' alignments, and separately for all alignments, when aligning the small FR-ES gold-standard corpus.

It may appear surprising that recall is significantly higher than with the English-Spanish corpus. Close reading of the results suggest this is because the French and Spanish sentences are often lexically more similar to each other than either is to the corresponding English sentence, and have more closely-corresponding word order, as with the following example:

(5)
FR : Il nous faut deux mois au minimum pour faire ce travail, avec le maximum de célérité et de sérieux requis.
ES : Necesitamos dos meses como mínimo para hacer ese trabajo, con la máxima celeridad y seriedad requerida.
EN: We needed at least two months to do this work with the required care, even at maximum speed.

As a result, the dictionary-based seed alignments that fuel the alignment process are more likely to be confirmed for French-Spanish that for English-Spanish. (Closely-corresponding word order also results in fewer 'possible' alignments in the gold standard.) However, overall precision is reduced when using a pivot language, typically for sentence pairs with less lexical correspondence, since the 'chained' translation suggestions for a query word can be more numerous and more distant semantically from the query word. For example, querying a French-English dictionary for a French word may result in three English translations, then querying an English-Spanish dictionary for each of those English words may result in nine Spanish translations in total, making spurious seed alignments more likely. Even so, precision and recall are both considerably higher than results achieved with GIZA++ for the larger gold-standard English-Spanish corpus shown in Table 2, although results for that language combination are not directly comparable with those for the small French-Spanish corpus, and naturally much higher than the GIZA++ results on this much smaller corpus, shown alongside in Table 3.

Results from using the system to align the small English-German corpus are shown in Table 4. Although direct-translation dictionaries were used, as for the larger English-Spanish corpus, precision is noticeably lower, while recall is noticeably higher. Performance of the aligner is conditioned by how many seed alignments are created in the first phase of the process,

that is, the *seed density* for the sentence pair to align. Where seed density is low, alignments produced are less fine-grained (there are more long spans where individual word correspondences can not be identified), producing more multiply-linked word alignments when converted as described above. Seed density statistics for the corpora used are show in Table 5, expressed as the percentage of total words in both languages for which a seed alignment was established by dictionary translation or heuristic.

Alignment type	Dictionary-based aligner			GIZA++		
	P	*R*	*F*	*P*	*R*	*F*
'sure' only	72.1%	63.1%	66.4%	25.0%	30.8%	27.0%
All	78.2%	53.3%	61.4%	30.2%	26.1%	27.1%

Table 4: Precision (P), recall (R) and F-measure (F) produced for 'sure' alignments, and separately for all alignments, when aligning the small EN-DE gold-standard corpus.

Although the German-English and English-German dictionaries are far larger than the other dictionaries used, manual inspection of aligned sentences indicates that they contain relatively few synonyms. In a German sentence containing 'Unterfangen', the only translation retrieved is 'undertaking', while the corresponding English word is 'enterprise', for which the only translation retrieved is 'Unternehmung', and therefore no seed alignment is established between those two words. Seed density would be improved by using synonyms during seed alignment generation, even if only for one of the languages concerned. Where span alignments are produced that have low seed density, they have a corresponding low alignment score. For the TM-related applications considered above, this would allow these less-reliable alignments to be rejected when recalling translation suggestions.

Corpus	EN-ES	FR-ES	EN-DE
Avg. seed density	77.3%	81.2%	68.4%

Table 5: Corpora seed densities

5. Conclusions

This paper presented a system producing phrase-to-phrase alignment for arbitrarily-small datasets, whose output can also be expressed as word alignments. The system has the advantages of being able to make use of readily-available machine-readable bilingual dictionaries, requiring no training step, and allowing domain-specific resources such as terminology databases to be easily exploited to assist in alignment accuracy for specialized text types. Evaluation of the system against a gold standard showed precision and recall were considerably better than achieved using the state-of-the-art GIZA++ word-alignment tool when aligning a relatively small dataset (400 sentence pairs). Results from alignment in a pivot-language scenario – albeit on a small set of sentence pairs – indicated that, for N languages, it would be feasible to require only 2(N - 1) dictionaries rather than N(N - 1).

6. Further work

Application of the alignment system to corpora consisting of other language pairs (German-French, German-Spanish, Welsh-English) is currently underway, both using direct-translation dictionaries and pivot-language dictionaries, for further intrinsic evaluation of results against gold-standard data. The algorithm is also to be integrated into a TM system to provide sub-sentential recall, allowing for extrinsic evaluation of performance. This system will optionally word-align TM data (when of sufficient size) using GIZA++, allowing results using the two aligners to be compared, and for GIZA++ to be the default aligner for large TMs. Testing is

also planned to measure the effect of using terminology databases as an external resource when aligning specialized texts with the system described. In that regard, an enhancement to the seed-alignment-generation process is to be developed, to allow for tentative alignments to be established by querying not only with single words, but with short spans of words.

References

Brown, P. F., V. J. D. Pietra, S. A. D. Pietra and R. L. Mercer (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* **19**(2): 263-311.

Carreras, X., I. Chao, L. Padró and M. Padró (2004). FreeLing: An Open-Source Suite of Language Analyzers. *Proc. 4th Intern. Conf. on Language Resources and Evaluation (LREC-04)*, Portugal.

Cicekli, I. and H. A. Güvenir (2001). Learning translation templates from bilingual translation examples. *Applied Intelligence* **15**(1): 57-76.

Esplà Gomis, M., F. Sánchez Martínez and M. L. Forcada Zubizarreta (2011). Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited. *Proceedings of the 15th Conference of the European Association for Machine Translation:* 81-88, Leuven, Belgium.

Esplà Gomis, M., F. Sánchez Martínez and M. L. Forcada (2012). A simple approach to use bilingual information sources for word alignment. *Procesamiento del Lenguaje Natural* **49**: 93-100.

Grishman, R. (1994). Iterative alignment of syntactic structures for a bilingual corpus. *Second Annual Workshop on Very Large Corpora (WVLC2)*, Kyoto, Japan, 57-68.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of the MT Summit 2005*, Phuket, Thailand.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran and R. Zens (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*: 177-180, Prague, Czech Republic.

Koehn, P., F. J. Och and D. Marcu (2003). Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*: 127-133, Edmonton, Canada.

Kranias, L. and A. Samiotou (2004). Automatic Translation Memory Fuzzy Match Post-Editing: A Step Beyond Traditional TM/MT Integration. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC '04)*: 331–334. Lisbon, Portugal.

Lambert, P., A. De Gispert, R. Banchs and J. B. Marino (2005). Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation* **39**(4): 267-285.

Lavie, A., A. Parlikar and V. Ambati (2008). Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, Columbus, OH.

Lezius, W. (2000). Morphy-German morphology, part-of-speech tagging and applications. *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart, Germany.

Macken, L. (2010). Sub-sentential alignment of translational correspondences. Ph.D. thesis, University of Antwerp.

Macklovitch, E. and G. Russell (2000). What's been forgotten in translation memory. *Envisioning Machine Translation in the Information Future: Proceedings of Fourth Conference of the Association for Machine Translation in the Americas (AMTA-2000)*: 137–146, Cuernavaca, Mexico.

Mandreoli, F., R. Martoglia and P. Tiberio (2003). Exploiting multi-lingual text potentialities in EBMT systems. *Proc. of the 13th IEEE International Workshop on Research Issues in Data Engineering: Multi Lingual Information Management (RIDE-MLIM)*: 9–15.

Och, F. J. and H. Ney (2003). A systematic comparison of various statistical alignment models. *Computational linguistics* **29**(1): 19-51.

Planas, E. (2005). SIMILIS Second-generation translation memory software. *27th International Conference on Translating and the Computer (TC27)*, London, United Kingdom.

Romary, L., S. Salmon-Alt and G. Francopoulo (2004). Standards going concrete: from LMF to Morphalou. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

Simard, M. (2003). Mémoires de traduction sous-phrastiques, Ph.D. thesis, Université de Montréal, Quebec, Canada.

Tinsley, J., V. Zhechev, M. Hearne and A. Way (2007). Robust language pair-independent sub-tree alignment. *Proceedings of Machine Translation Summit XI*: 467-474, Copenhagen, Denmark.