Improved Word Alignment with Statistics and Linguistic Heuristics

Ulf Hermjakob
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292, USA
ulf@isi.edu

Abstract

We present a method to align words in a bitext that combines elements of a traditional statistical approach with linguistic knowledge. We demonstrate this approach for Arabic-English, using an alignment lexicon produced by a statistical word aligner, as well as linguistic resources ranging from an English parser to heuristic alignment rules for function words. These linguistic heuristics have been generalized from a development corpus of 100 parallel sentences. Our aligner, UALIGN, outperforms both the commonly used GIZA++ aligner and the state-of-the-art LEAF aligner on F-measure and produces superior scores in end-to-end statistical machine translation, +1.3 BLEU points over GIZA++, and +0.7 over LEAF.

1 Introduction

Word alignment is a critical component in training statistical machine translation systems and has received a significant amount of research, for example, (Brown et al., 1993; Ittycheriah and Roukos, 2005; Fraser and Marcu, 2007), including work leveraging syntactic parse trees, e.g., (Cherry and Lin, 2006; DeNero and Klein, 2007; Fossum et al., 2008). Word alignment is also a required first step in other algorithms such as for learning sub-sentential phrase pairs (Lavie et al., 2008) or the generation of parallel treebanks (Zhechev and Way, 2002).

Yet word alignment precision remains surprisingly low, under 80% for state-of-the-art aligners on not closely related language pairs.

Consider the following Arabic/English sentence pair with alignments built by the statistical word aligner LEAF:

Bitext Arabic: وفاز التايلاندى بارادورن سريشافان على الاسترالى جايسون ستولتنبرغ 4 - 6 و 4 - 6 ,والتشيكى ييرى فانيك على الالمانى لارش بورغسمولر 4 - 6 و 4 - 6

Gloss: Won(1) Thai Paradorn Srichaphan(1) on/to(2) Australian Jason(2) Stoltenberg(3) 6(4) - 4(5) and(3) 6(4) - 4(5), and Czech Jiří(7) Vaněk(7) on/to German(6) Lars Burgsmüller 6(4) - 4 and(3) 6(4) - 4

Bitext English: Thailand 's(1) *Baradorn Srichfan*(1) beat(2) Australian *Gayson*(1) *Stultenberg*(3) 6(4) - *6*(4) 6(4) - 4(5) , Czech player(1) *Pierre*(1) *Vanic*(7) beat(6) Germany(6) 's Lars Burgsmuller 6 - 4 6 - 4

In the example above, words with the same index in the gloss for Arabic and the English are aligned to each other, alignment errors are underlined, translation errors are in *italics*. For example, the Arabic words for *won* and *Srichaphan* are aligned with the English words *'s*, *Srichfan*, *Gayson*, *player* and *Pierre*.

As reflected in the example above, typical alignment problems include

- words that change sentence position between languages, such as verbs, which in Arabic are often sentence-initial (e.g. *won/beat* in the example above)
- function words without a clear and explicit equivalent in the other language (e.g. the Arabic و/*and* in the example above)
- lack of robustness with respect to poor translations (e.g. *Gayson Stultenberg* instead of *Jason Stoltenberg*) or bad sentence alignment.

We believe we can overcome such problems with the increased use of linguistically based

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 229–237,
Singapore, 6-7 August 2009.

heuristics. We can model typical word order differences between English and Arabic using English parse trees and a few Arabic-specific phrase reordering heuristics. We can narrow the space of possible alignment candidates for function words using English parse trees and a few heuristics for each type of function word.

These heuristics have been developed using a development corpus of 100 parallel sentences. The heuristics are generalizations based on patterns of misaligned words, misaligned with respect to a Gold Standard alignment for that development corpus.

The following sections describe how our word aligner works, first how relatively reliable content words are aligned, and then how function words and any remaining content words are aligned, with a brief discussion of an interesting issue relating to the Gold Standard we used. Finally we present evaluations on word alignment accuracy as well as the impact on end-to-end machine translation quality.

2 Phase I: Content Words

We divide the alignment process into two phases: first, we align relatively reliable content words, which in phase II we then use as a skeleton to align function words and remaining content words.

Function words such as English *a, ah, all, am, an, and, any, are, as, at, ...* are common words that often do not have an explicit equivalent word or words in the other side of the bitext. In our system, we use a list of 96 English and 110 Arabic function words with those characteristics. For the purposes of our algorithm, a word is a function word if and only if it is on the function word list for its language. A content word then is defined as a word that is neither a function word nor punctuation.

The approach for aligning content words in phase I is as follows: First, we score each combination of an Arabic content word and English content word in an aligned sentence and align those pairs that pass a threshold, typically generating too many alignments. Second, we compute a more comprehensive score that also takes into consideration matching alignments in the context around each alignment. Third, we eliminate inferior alignments that are incompatible with higher-scoring alignments.

The score in the first step is *pointwise mutual information* (PMI). The key resource to compute this PMI is an alignment lexicon generated beforehand by a statistical word alignment system from a large bitext. An alignment lexicon is a list of triples, each consisting of an English word, an Arabic word, and how often they have been aligned for a given bitext. Additional counts on how often each English and Arabic word occurs allow us use this alignment lexicon to compute PMI(e,f) = $\log \frac{p(e,f)}{p(e) \cdot p(f)}$. We align those Arabic and English content words that have a PMI > 0 and a minimum alignment lexicon count (≥ 10 initially). Using the alignment lexicon generated by a statistical word aligner to compute PMIs is the principal statistical component in our system. We explored alternative metrics such as the dice-coefficient that was used by other researchers in earlier alignment work, but found PMI to work better for our system.

In a second step, we lay a window of size 5 around each aligned pair of Arabic and English words (counting only content words) and then add to the PMI score of the link itself the PMI scores of other links within that window, with a distance weight of $\frac{1}{distance+1}$. This yields a new score that takes into account whether a link is supported by context.

In the third step, we check for overgenerated links, comparing links that share an Arabic or an English word. If a word on one side of the bitext is linked to multiple *adjacent* words on the other, we leave them alone, as one word in one language often corresponds to multiple words in the other. However, if a word on one side is linked to non-adjacent words on the other side, this flags an incompatibility, and we remove those links that have inferior context-sensitive scores. This removal is done one link at a time, with the lowest relative scores first.

We boost the process we just described in a few ways. In the first alignment step, we also include as alignment candidates any content words that are string-identical on each side, such as ASCII numbers and ASCII words. We finally also include as alignment candidates those word pairs that are transliterations of each other to cover rare proper names (Hermjakob et al., 2008), which is important for language pairs that don't share the same alphabet such as Arabic and English.

2.1 Reordering Using an English Parser

We use a refined notion of context window that models word order differences between Arabic and English. Traversing a parse tree for English, we identify sub-trees for which the order in Arabic can be substantially different. In Arabic, for example, the verb is often sentence-initial. So for trees or subtrees identified by the parser as sentences, we generate an alternative reordering of its subtrees where the verb has been moved to the front. Similarly, in a noun phrase, we generate an alternative order where adjectives are moved to the right of the noun they modify.

For example, consider the sentence *John bought a new car* . We can reorder its parse tree both at the sentence level: *(bought) (John) (a new car) (.)* as well as at its object NP level: *(a) (car) (new)*. If fully enumerated, this would yield these four reordering alternatives:

1. John bought a new car .
2. John bought a car new .
3. bought John a new car .
4. bought John a car new .

We don't actually explicitly enumerate all variants but keep all reordering alternatives in a reordering forest, since the number of fully expanded reorderings grows exponentially with the number of phrases with reordering(s). At the beginning of Phase I, we compute from this reordering forest a minimum distance matrix, which, for specific instances of the words *John* and *car* would record a minimum distance of 1 (based on reordering 4, skipping the function word *a*).

For the example sentence at the beginning of the paper we would get reorderings including the following:

Engl. orig.: thailand 's baradorn srichfan beat ...
A reordering: beat thailand 's baradorn srichfan ...
Arabic (gloss): won thai paradorn srichaphan ...

In the above reordered English alternative, *beat* and *thailand* are next to each other, so their minimum distance is 1, which means that a link between English *thailand* and Arabic *thai* now strongly boosts the context-sensitive score between English *beat* and Arabic *won*.

2.2 Morphological Variation

Another challenge to content word alignment is morphological variation which can create data sparsity in the alignment lexicon. For example, in a given bitext sentence, the Arabic word *AlAwDAE* might be translated as *situational*, for which there might be no support in the alignment lexicon. However the PMI between *AlAwDAE* and *situation* might be sufficiently high. Additionally, there is another Arabic word, *AlHAlAt*, which often translates as both *situation* and *situational*.

To take advantage of such constellations, we built morphological variation lists for both Arabic and English, lists that for a given head word such as *situational* lists variants such as *situation*, and *situations*.

We built these lists in a one-time process by identifying superficially similar words, i.e. those that vary only with respect to an ending or a prefix, and then semantically validating such candidates using a pivot word in the other language such as *AlHAlAt* that has sufficiently strong alignment lexicon co-alignment counts with both *situation* and *situational*. The alignment lexicon co-alignment count of an Arabic word w_{ar} and an English word w_{en} is considered strong enough, if it is at least 2.0 and at least 0.001 times as high as the highest co-alignment count of w_{ar} with any English word; words shorter than four letters are excluded from consideration. So because *situation* and *situational* are superficially similar **and** they are both have a strong alignment count with *AlHAlAt* in the alignment lexicon, *situation* is added to the English morphological variation list as a variant of *situational* and vice versa.

Exploring whether we can align *situational* and *AlAwDAE* in the bitext, we find that *situational* is a morphological variant of *situation* (based on our morphological variation list for English); next we find that based on the alignment lexicon, there is a positive PMI between *situation* and *AlAwDAE*, which completes the chain between *situational* and *AlAwDAE*, so we include them as an alignment candidate after all. The PMI of such a morphological-variation-based candidate is weighted by a 'penalty' factor of 0.5 when compared with the PMI of any competing alignment candidate without such morphological-variation step.

Similarly, the English pivot word *situations* can be used to semantically validate the similarity between Arabic *AlAwDAE* and *AwDAE* for our Arabic morphological variation list. The resulting Arabic morphological variation list has entries for 193,263 Arabic words with an average of 4.2 variants each; our English morphological variation list has 57,846 entries with 2.8 variants

each.

At the end of phase I, most content words will be aligned with relatively high precision. Since function words often do not have an explicit equivalent word or words in the other side of a bitext, they can not be aligned as reliably as content words based on bilingual PMI.[1] Note that due to data sparsity, some content words will remained unaligned in phase I and will subsequently be aligned in phase II as explained in section 3.3.

3 Phase II: Function Words

In Phase II, we align function words, punctuation, and some remaining content words. Function words can be classified into three categories: monovalent, divalent and independent. Monovalent function words modify one head; they include articles (which modify nouns), possessive pronouns, demonstrative adjectives and auxiliary verbs. Divalent function words connect two words or phrases; they include conjunctions and prepositions. Independent function words include non-possessive pronouns and copula (e.g. *is* as a main verb). Each of these types of function words is aligned according to its own heuristics.

In this section we present three representative examples, one for articles (monovalent), one for prepositions (divalent), as well as a structural heuristic.

3.1 Example: Articles

Monovalent function words have the simplest heuristics. Recall that Arabic does not have articles (only a definite prefix *Al-* added to one or more words in a definite noun phrase), so there is usually no explicit equivalent of the English article on the Arabic side.

For an English article, our system identifies the English head word that it modifies based on the English parse tree, and then aligns it with the same Arabic word(s) which that head word is aligned with.

3.2 Example: Prepositions

Divalent function words are much more interesting. In many cases, an English preposition corresponds to an explicit Arabic preposition in basically the same position. Alignment in that case is straightforward. However, some Arabic prepositions and even more English prepositions do not have an explicit counterpart on the other side. We call such prepositions *orphan prepositions*. The English preposition *of* is almost always orphaned in this way.

The decision how to align such an orphan preposition is not trivial. Consider the bitext *island of Basilan/jzyrp bAsylAn*, a typical (NP1 (P NP2)) construction on the English side. Should we co-align the preposition *of* with the head of NP1 or the head of NP2? In English syntax, the preposition is grouped with NP2, but a preposition is often better "motivated" by NP1. We therefore decided to use the English parse tree to identify the heads of both NP1 and NP2, identify the Arabic words aligned to these heads as candidates, and then align the preposition to the Arabic candidate word with which it has the highest bilingual PMI. It turns out that in most cases this will be the candidate on the "left". For the example at the top of this paragraph, *of* will be aligned with *jzyrp* ("island"), which is actually desirable for MT, as it facilitates subsequent rule extraction of type "island of X/jzyrp X". We refer to this orphan preposition alignment style as *MT-style*.

According to the gold standard alignment guidelines used for the LDC Gold Standard however, an orphan preposition should always be aligned to the "right", to *bAsylAn* in the example above. We therefore implemented an alternative *GS-style* (for "Gold Standard") to be able to later evaluate the impact of these alternatives alignment styles.

The question whether GIZA or LEAF alignments will indeed give meaningful scores to support the *MT-style* attachments will be answered by the MT experiments described in section 4.3.

Here is a more complex example with Arabic (A), its gloss (G) and English (E):

Arabic: الاحد اغارت الطائرات الاميركية على منطقة جوار
Gloss: sunday attacked aircraft american on/to area jiwar
Engl.: on sunday american aircraft attacked the area of jiwar

For the Arabic orphan preposition على/*ElY* ("on/to"), our system identifies two candidates based on the English parse tree: *attacked* and *area*. Based on a higher mutual information, our system then aligns Arabic *ElY* ("on/to") with English *attacked*, which results in the English word *attacked* now being aligned to both Arabic *attacked* and the

[1] It is this lack of reliability that is the defining characteristic of our function words, differentiating them from the concept of marker words used in EBMT chunking (Way and Gough, 2002).

Arabic *on/to*, even though they are not adjacent. In the Gold Standard, Arabic *on/to* is aligned with English *area*, and LEAF aligns it with English *on* (yes, the one preceding Sunday). This is apparently very tempting as Arabic *on/to* is often translated as English *on*, but here it is incorrect, and our system avoids this tempting alignment because it is ruled out linguistically.

Note that in some cases, such as sentence-initial prepositional phrases, there is only one candidate; occasionally, when relevant content words remain unaligned, no candidate can be identified, in which case the orphan preposition remains unaligned as well.

3.3 Example: Adjectives

It is not uncommon that content words that we would like to be aligned are not supported by the alignment lexicon, due to general data sparsity or maybe a somewhat unorthodox translation. In those cases we can use structure and word order knowledge to make reasonable alignments anyway.

Consider an English noun phase ADJ-E NOUN-E and the corresponding Arabic NOUN-A ADJ-A. If the nouns are already aligned, but the adjectives are not yet aligned, we can use the English parse tree to identify ADJ-E as a modifier to NOUN-E, and, aware that adjectives in Arabic post-modify their nouns, identify the corresponding Arabic word based on structure and word order alone. This can be done the other way around as well (link nouns based on already aligned adjectives) and other elements of other phrases as well.

As more and more function words and remaining content words get aligned, heuristics that weren't applicable before may now apply to the remaining unaligned words, so we perform four passes through a sentence pair to align unaligned words using heuristics. We found that an additional fifth pass did not yield any further improvements.

4 Experiments

We evaluated our word aligner in terms of both alignment accuracy and its impact on an end-to-end machine translation system.

4.1 Alignment Experiments

We evaluated our word aligner against a Gold Standard distributed by LDC. The human alignments of the sentences in this Gold Standard are based on the 2006 GALE Guidelines for Arabic Word Alignment Annotation.

Both the 100-sentence development set and the separate 837-sentence test set are Arabic newswire sentences from LDC2006E86. The test set includes only sentences for which our English parser (Soricut and Marcu, 2003) could produce a parse tree, which effectively excluded a few very long sentences.

In the first set of experiments, we compare two settings of our UALIGN system with other aligners, GIZA++ (Union) (Och and Ney, 2003) and LEAF (with 2 iterations) (Fraser and Marcu, 2007). The GIZA++ aligner is based on IBM Model 4 (Brown et al., 1993). We chose GIZA Union for our comparison, because it led to a higher BLEU score for our overall MT system than other GIZA variants such as GIZA Intersect and Grow-Diag. The two settings of our system vary in the style on how to align orphan prepositions. Besides precision, recall and (balanced) F-measure, we also include an F-measure variant strongly biased towards recall (α=0.1), which (Fraser and Marcu, 2007) found to be best to tune their LEAF aligner for maximum MT accuracy. GIZA++ and LEAF alignments are based on a parallel training corpus of 6.6 million sentence pairs, incl. the LDC2006E86 set mentioned above.

Aligner	Prec.	Recall	F-0.5	F-0.1
GIZA	26.9	84.3	40.8	69.5
LEAF	73.3	79.7	76.4	79.0
UALIGN MT-style	82.5	80.0	81.2	80.2
UALIGN GS-style	84.0	82.9	83.5	83.0

Table 1: Alignment precision, recall, F-measure (α=0.5), F-measure(α=0.1) for different aligners; with UALIGN using LEAF alignment lexicon.

Our aligner outperforms both GIZA and LEAF on all metrics. Not surprisingly, the GS-style alignments, which align "orphan" prepositions according to Gold Standard guidelines, yield higher scores than MT-style alignments. And interestingly by a remarkably high margin.

In a second set of experiments, we measure the impact of using different input alignment lexicon used by our aligner on alignment accuracy. In one case UALIGN uses as input the alignment lexicon produced by LEAF, in the other the alignment lexicon produced by GIZA. All experiments in table 2

are for UALIGN.

Style	A-Lexicon	Prec.	Recall	F-0.5	F-0.1
MT	from LEAF	82.5	80.0	81.2	80.2
MT	from GIZA	80.8	79.2	80.0	79.4
GS	from LEAF	84.0	82.9	83.5	83.0
GS	from GIZA	82.1	81.8	82.0	81.9

Table 2: Alignment precision, recall, F-measure (α=0.5), F-measure(α=0.1), all of UALIGN, for different alignment styles, different input alignment lexicons.

As LEAF clearly outperforms GIZA on F-0.1 (79.0 vs. 69.5, see table 1), the alignment lexicon based on LEAF is better, so it is not surprising that when we use an alignment lexicon based on GIZA, all metrics degrade, and consistently so for both alignment styles. However the drop in F-0.1 of about 1 point (80.2 $\rightarrow$ 79.4 and 83.0 $\rightarrow$ 81.9) is much smaller than the differences between the underlying aligners themselves. Our aligner therefore degrades quite gracefully for a worse alignment lexicon.

Aligner	Arabic aligned	Engl. aligned
GIZA Union	100%	100%
LEAF	99.99%	97.25%
UALIGN	92.10%	91.55%
Gold Standard	95.37%	95.86%

Table 3: Percentages of Arabic and English words aligned

Table 3 shows how much LEAF and UALIGN differ in the percentage of Arabic and English words aligned (correctly or incorrectly). LEAF is much more aggressive in making alignments, aligning almost every Arabic word. Our aligner still leaves some 8% of all words in a sentence unaligned (an opportunity for further improvements). For comparison, in the Gold Standard, 4-5% of all words in our test corpus are left unaligned.

4.2 Impact of Sub-Components

To better understand the impact of several alignment system sub-components, we ran a number of experiments disabling individual sub-components and then comparing the resulting alignment scores with those of the full system. We also measured alignment scores running Phase II with 0 to 5 passes. The test set was the same as in section 4.1.

System	Prec.	Recall	F-0.1
Full system (FS)	84.0	82.9	83.0
FS w/o morph.variation	84.0	82.4	82.5
FS w/o Engl. tree reord.	83.8	82.7	82.8
FS w/o string identity	84.0	82.8	82.9
FS w/o name translit.	84.0	82.8	82.9
System after Phase I	90.6	44.5	46.8
+ Phase II w/ 1 pass	87.6	77.1	78.0
+ Phase II w/ 2 passes	85.8	80.3	80.8
+ Phase II w/ 3 passes	84.2	82.7	82.8
+ Phase II w/ 4 passes	84.0	82.9	83.0
+ Phase II w/ 5 passes	84.0	82.9	83.0

Table 4: Impact of sub-components on alignment precision, recall, F-measure, with GS-style attachments, based on the LEAF alignment lexicon.

Special sub-components of Phase I include adding link candidates for ASCII-string-identical words and transliterated names (see last paragraph before section 2.1), reordering using an English parser (section 2.1) and morphological variation (section 2.2). Each of these sub-components provides a small boost to F-0.1, ranging from +0.1 to +0.5. The second part of the table shows alignment scores before and after each pass of Phase II. Our full system includes 4 passes; an additional 5th pass did not yield any further improvements. Note that during Phase II, precision drops. This is a reflection of (1) our strategy to first align relatively reliable content words in Phase I, followed by less reliable function words and remaining content words, and (2) the challenges of building reliable Gold Standard alignments for function words and non-literal translations.

4.3 MT Experiments

The ultimate test for a word aligner is to measure its impact on an end-to-end machine translation system. For this we aligned 170,863 pairs of Arabic/English newswire sentences from LDC, trained a state-of-the-art syntax-based statistical machine translation system (Galley et al., 2006) on these sentences and alignments, and measured BLEU scores (Papineni et al., 2002) on a separate set of 1298 newswire test sentences. Besides swapping in a new set of alignments for the same set of training sentences, and automatically retuning the parameters of the translation system for each set of alignments, no other changes or adjustments were made to the existing MT system.

In the first set of experiments, we compare two settings of our UALIGN system with other aligners, again GIZA++ (Union) and LEAF (with 2 iterations). The two settings vary in the alignment lexicon that the UALIGN aligner uses as input.

Aligner	BLEU
GIZA	47.4
LEAF	48.0
UALIGN using GIZA alignment-lexicon	48.4
UALIGN using LEAF alignment-lexicon	48.7

Table 5: BLEU scores in end-to-end statistical MT system based on different aligners. Both UALIGN variants use MT-style alignments.

With a BLEU score of 48.7, UALIGN using a LEAF alignment-lexicon is significantly better than both GIZA (+1.3) and LEAF (+0.7). This and other significance assertions in this paper are based on paired bootstrap resampling tests with 95% confidence. UALIGN using a GIZA alignment-lexicon significantly outperforms GIZA itself (+1.0).

In a second experiment, we measured the impact of the two alignment styles on BLEU. Recall that for GS-style alignments, orphan prepositions are always co-aligned to the right, following Gold Standard annotation guidelines, whereas for MT-style alignments, mutual information is used to decide whether to align orphan prepositions to the left or to the right.

Aligner	BLEU
LEAF	48.0
UALIGN with GS-style alignments	48.0
UALIGN with MT-style alignments	48.7

Table 6: BLEU scores in end-to-end statistical MT system based on different alignment styles for orphan prepositions. Both UALIGN variants use a LEAF alignment lexicon.

While the GS-style alignments yielded a 2.8 point higher F-0.1 score (83.0 vs. 80.2), the MT-style alignments result in a significantly better BLEU score (48.7 vs. 48.0). This shows that (1) a seemingly small difference in alignment styles can have a remarkably high impact on both BLEU scores and alignment accuracy as measured against a Gold Standard, and that (2) optimizing alignment accuracy against an alignment Gold Standard does **not** necessarily optimize BLEU in end-to-end MT. The latter has been observed by other researchers before, but these results additionally suggest that the gold-standard annotation style might itself have to shoulder part of the blame.

4.4 Corpus Noise Robustness

In a small random "sanity check" sample from the 170,863 training sentences for the MT experiment, we found cases where the sentence in one language contained much more material than the sentence in the other language. Consider, for example the following sentence pair (with spurious material underlined):

Arabic:
لكن ايضا هناك بند اخر ينص على انه اذا لم ينشا الفندق ،

Gloss: but also there-is clause another stipulates on/to that if not established the-hotel ,

English: but , also there is another clause that stipulates that if the hotel is not established , then the government shall be compensated .

Both LEAF and UALIGN correctly align the English "*but , also ... not established ,*" with the Arabic side. LEAF further aligns all words in the spurious English "*then the government shall be compensated .*" with seemingly random material on the Arabic side, whereas UALIGN leaves these spurious words completely unaligned. It would be reasonable to speculate that this behavior, observed in several cases, may be contributing to the good BLEU scores.

5 Discussion

Building on existing statistical aligners, our new word aligner significantly outperforms the best word aligner to date in both alignment error rate and BLEU score.

We have developed an approach to word alignment that combines a statistical component with linguistic heuristics. It is novel in that it goes beyond generic resources such as parsers, adding heuristics to explicitly model word order differences and function word alignment.

The approach has numerous benefits. Our system produces superior results both on alignment accuracy and end-to-end machine translation quality. Alignments have a high precision. The system is fast (about 0.7 seconds per sentence), and sentences are aligned individually so that a large corpus can easily be aligned on several computers in

parallel. All alignment links are tagged with additional information, such as which phase and/or heuristic created them, yielding extensive explanatory power to the developer for easy understanding on how the system arrived at a given alignment. Our approach needs and uses a parser for only one side (English) and not for the other (Arabic).

On the other hand, some of the components of this aligner are language-specific, such as word order heuristics, the list of specific function words, and morphological variation lists. While these parts of the system need to be adapted for new languages, the overall architecture and types of heuristics and function words are language-independent. Chinese for example has different specific types of function words such as aspect markers and measure words. But these fall into the existing category of monovalent function words and will be treated according the same principles as other monovalent function words (section 3.1). Similarly, Japanese postpositions would be treated like other divalent function words (such as Arabic or English prepositions). The author and developer has a basic knowledge of Arabic in general, and an intermediate knowledge of Arabic grammar, which means that no intimate knowledge of Arabic was required to develop the language-specific components. This same author and developer recently started to adapt UALIGN to Chinese-English word alignment.

The alignment rate is still somewhat low. We plan to increase it by enlarging our development set beyond 100 sentences and adding further heuristics, as well as generalizing the output word alignment structure to allow alignments of words to larger constituents in a tree, and to explicitly assert that some words are not covered by the other side of a bitext to model poor translations and poor sentence alignments.

Acknowledgment

This research was supported under DARPA Contract No. HR0011-06-C-0022. The author would like to thank Kevin Knight and the anonymous reviewers for their helpful suggestions, and Steve DeNeefe for running the end-to-end MT evaluations.

References

Peter E. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. In *Computational Linguistics* Vol. 19(2), pages 263–311.

Colin Cherry and Dekang Lin. 2006. Soft Syntactic Constraints for Word Alignment Through Discriminative Training. In *Proceedings of the 44th Annual Meeting on Association for Computational Linguistics*, Sydney, Australia, pages 105–112.

John DeNero and Dan Klein. 2007. Tailoring Word Alignments to Syntactic Machine Translation. In *Proceedings of the 45th Annual Meeting on Association for Computational Linguistics*, Prague, Czech Republic, pages 17–24.

Victoria Fossum, Kevin Knight and Steven Abney. 2008. Using Syntax to Improve Word Alignment Precision for Syntax-Based Machine Translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, Columbus, Ohio, pages 44–52.

Alexander Fraser and Daniel Marcu. 2007. Getting the Structure Right for Word Alignment: LEAF. In *Proceedings of Conference for Empirical Methods in Natural Language Processing (EMNLP)*, Prague, Czech Republic, pages 51–60.

Alexander Fraser and Daniel Marcu. 2007. Measuring Word Alignment Quality for Statistical Machine Translation. In *Computational Linguistics* Vol. 33(3), pages 293–303.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer 2006. Scalable Inference and Training of Context-Rich Syntactic Translation Models. In *Proceedings of the 44th Annual Meeting on Association for Computational Linguistics*, Sydney, Australia, pages 961–968.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III 2008. Name Translation in Statistical Machine Translation: Learning When to Transliterate. In *Proceedings of the 46th Annual Meeting on Association for Computational Linguistics*, Columbus, Ohio, pages 389–397.

Abraham Ittycheriah and Salim Roukos. 2005. A Maximum Entropy Word Aligner for Arabic-English Machine Translation. In *Proceedings of Joint Conference of Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, British Columbia, Canada, pages 89–96.

Alon Lavie, Alok Parlikar and Vamshi Ambati. 2008. Syntax-Driven Learning of Sub-Sentential Translation Equivalents and Translation Rules from Parsed Parallel Corpora. In *Proceedings of the ACL/HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, Columbus, Ohio, pages 87–95.

Dan Melamed. 2000. Models of translational equivalence among words. In *Computational Linguistics* Vol. 26(2), pages 221–249.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics* Vol. 29(1), pages 19–51.

Franz Josef Och and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. In *Computational Linguistics* Vol. 30(4), pages 417–449.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, PA, pages 311–318.

Radu Soricut and Daniel Marcu. 2003. Sentence Level Discourse Parsing Using Syntactic and Lexical Information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, Canada, pages 149–156.

Andy Way, Nano Gough. 2003. wEBMT: developing and validating an example-based machine translation system using the world wide web In *Computational Linguistics* Vol. 29(3), pages 421–457.

Ventsislav Zhechev, Andy Way. 2008. Automatic Generation of Parallel Treebanks. In *Proceedings of 22nd International Conference on Computational Linguistics (COLING)*, Manchester, UK, pages 1105–1112.