RC24024 (W0608-051) August 14, 2006
Computer Science

IBM Research Report

The IBM Arabic-English Word Alignment Corpus

Abraham Ittycheriah, Yaser Al-Onaizan, Salim Roukos
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

The IBM Arabic-English Word Alignment Corpus

Abraham Ittycheriah and Yaser Al-Onaizan and Salim Roukos
IBM T.J. Watson Research Center
1101 Kitchawan Road
Yorktown Heights, NY 10598
{abei,onaizan,roukos}@us.ibm.com

Abstract

This report documents our work on producing a hand word aligned corpus for Arabic-English. We describe the corpus, the guidelines given to the annotators, and a measurement of the intra- and inter-annotator agreement. This corpus has been used as training material for both word alignment algorithms and machine translation algorithms.

1 Introduction

Word alignment has become an essential part of many translation systems. Early methods in statistical machine translation used unsupervised methods to obtain word alignments by assuming a hidden link between words in a sentence pair (Brown et al., 1993). This assumption together with an EM (expectation-maximization) algorithm formed an elegant solution to automatically obtain word alignments. Usually in the natural language processing field, supervised algorithms are investigated first and unsupervised approaches augment and enhance the performance of such algorithms. Word alignment algorithms have now arrived at the same conclusion (Ittycheriah and Roukos, 2005).

In this report, we first cover the pre-annotation issues such as tokenization, number classing for each of the languages. A brief overview of the annotation tool is then presented followed by a description of the annotation guidelines; we also report on the training corpus used for alignment as well as a test corpus for evaluating alignment algorithms. We also report a measurement of the amount of labor for word alignment, and the annotation agreement for the task. Finally, the output format of the annotation is presented.

2 Language Issues

Some minimal processing in each language is required before word-alignment can be performed. We keep this to a minimum, but we note that the tokenization and other processes used here are done by machine and as such there are unfortunately few instances of errors.

2.1 English

White-space tokenization and punctuation separation are performed on each string. Dashes are separated except in words that begin with "Al" which is a common name prefix in Arabic and in the cases of well known english phrases such as "so-called" or "state-of-the-art". Other examples like "Israeli-Palestinian" may be expressed in Arabic as three separate words in which case it would have been nice to separate these words but is not done in this version of the corpus. Also note that the tokenization is a probabilistic algorithm and the output is sometimes inconsistent.

Alpha-numerics are classed into a token prefixed with $NUM_ in English and all files except Ummah have the content in parentheses.

2.2 Arabic

The arabic text is tokenized and normalized. The tokenization is done via deterministic rules, where only punctuations are white-space separated. In the normalization step, the following mappings are performed:

- Map Arabic punctuations and digits into their English equivalents.
- Map Alef with Hamza above or below to a bare Alef.
- Map Alef-maqsura to Yaa.
- Remove Arabic diacritics and Kashida (tatweel).

Alpha-numerics are classed into a token prefixed with $num_ and all files except Ummah have the content in parentheses.

3 Annotation Tool

A web-based annotation tool was used and is shown in Figure 1. The tool displays the sentences in columns. A word is selected by clicking on the word and ranges can be selected by holding down the shift key. Clicking on a word in the second column draws a line to connect the word(s) completing a single link. Pressing <Alt> while clicking removes an alignment.

3.1 Status bits

Initially, the following status bits were defined, but in consideration of the time spent marking the status of the links we used only 'g', and 'x' as status indications. Once an alignment is made using the tool, the status automatically changes to 'g' from 'x'. For completeness, listed below are the definition of each status bit.

- good - g - Most words are expected to be marked as good. Once a link is made by the annotator, the status is automatically updated to indicate 'g'.
- fair - f - This status is used to indicate a 'loose' translation, where in the opinion of the annotator a better word could have been used.
- error - e - Errors are a more severe version of 'loose' translation and this word choice does not carry the meaning of the source sentence.
- extra - x - This word is not translated in the other sequence. This status is only applied when linking to either SourceNull or TargetNull.
- spontaneous - s - This word is to be considered spontaneous in this sequence. This status is usually on linking to SourceNull or TargetNull but could be applied to the internal words of a phrase.
- unsure - u - The word is unknown to the annotator.

The plan at the time of the annotation was to iterate and put the status indications as necessary, although currently there are no plans to complete the status indicators.

4 Annotation Guidelines

The following guidelines were used in the word alignment task.

4.1 Determiners

The definite article in English should be aligned with the same word the head noun is aligned to if the Arabic word is also definite (i.e., starts with Al#). An example is shown in Figure 2 for source word (3) which connects to the target words (5) and (9).

If the English head noun is definite but the Arabic head is indefinite, then either 'the' is connected to the Arabic head or not connected at all. An example where it is connected is the target word (13) which is connected to the arabic word without explicit evidence of the link.

4.2 Particles

English particles that change the meaning of the verb associated with them should be aligned with the word that is aligned to the verb if they don't have an equivalent particle in Arabic (e.g., up in 'give up', off in 'take off'). As an example, in Figure 2, the target word 'on' is connected to its verb.

4.3 Spontaneous Words

Arabic words that are not translated into English should be aligned to a special 'Null' token.

4.4 Phrase Alignment v. Word-to-Word Alignment

If the annotator feels that a word-to-word alignment is not possible between two larger units, then a phrase-to-phrase alignment is acceptable.

4.5 Acronyms and Abbreviation

Acronyms and Abbreviation in one language should be aligned to the entire phrase in the other language that refers to the same entity.

4.6 Attached Arabic prefixes or suffixes

Arabic is un-segmented. Therefore some prefixes or suffixes might have their equivalents in English as separate words. In such cases the annotator is instructed to align the un-segmented Arabic word to the English equivalent of the main Arabic word as well as the equivalent of the prefix.

5 Corpus Selection

The files that are being released in Version 1.0 are detailed below in Table 1. The test corpus is the MT03 machine translation test released by NIST. The four references have been combined into a single file: reference 1, 'ahd' is sentence 0...662; reference 2, 'ahe' is sentence 663...1325, etc. In (Ittycheriah and Roukos, 2005), the test set used is the first 50 sentences of the 'ahd' reference.

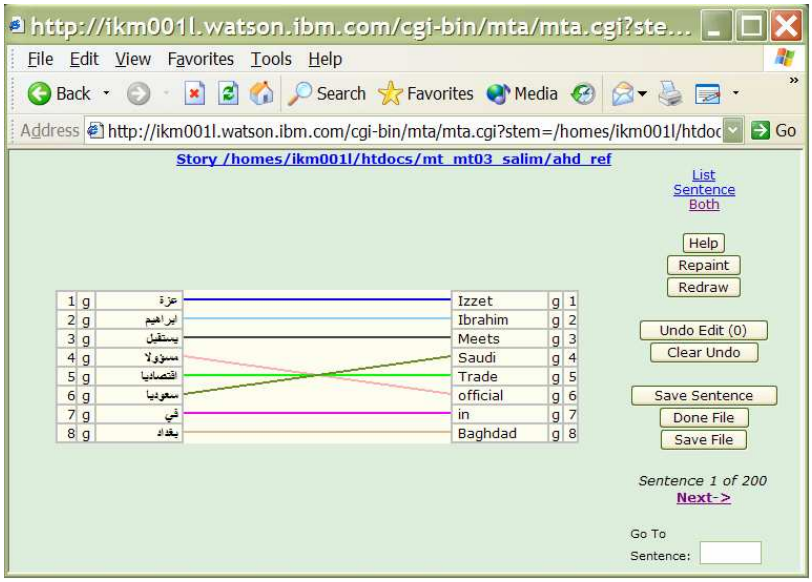

Figure 1: The word alignment tool.

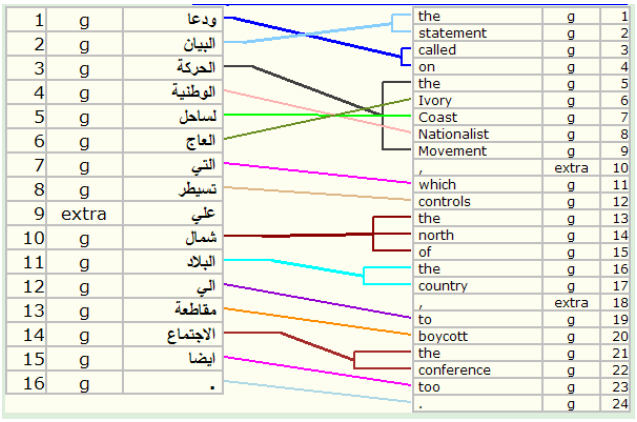

Figure 2: An example alignment.

6 Timing Studies

The annotator working a normal 8 hour day was able to annotate about 120-150 sentences per day. Very long or poor translations as judged by the annotator were marked and removed later from the corpus in order to ensure the rate of annotation was above 100 sentences per day.

7 Annotation Agreement

As reported in (Ittycheriah and Roukos, 2005), intra/inter-annotator agreement was measured on the test set in order to determine the feasibility of human annotation of word links. These are shown in Table 2. In the table, the column for 'Annotator 1 Correction' is the first annotator correcting his own word alignments after a span of a year. After two weeks, the annotator (Annotator 1') was given the same material with all the links removed and asked to realign and we see that there is more discrepancy in resulting alignments. The differences are largely on the head concept where determiners are attached and the alignment of spontaneous words. The performance with a second annotator is in the same range as the reannotation by a single annotator.

	Anno. 1 Correction	Anno. 1'	Anno. 2
Anno. 1	96.5	92.4	91.7
Anno. 1'	95.2	—	93.2

Table 2: F-measure for human performance on word alignment for Arabic-English.

Type	Filename	LDC Catalog #	# of sentences
Train	afa.align	LDC2003E05 LDC2003E09 LDC2004T17	4468
Train	annahar.align.fw	LDC2004E07	4320
Train	mar_ummah.align.fw	LDC2004T18	179
Train	ummah.align.fw	LDC2004T18	644
Train	treebank.align	LDC2005T02	4350
Total			13961
Test	mt03.ref		2652

Table 1: Corpus statistics.

8 Output Format

The following is the output format for the alignment tools as well as the hand-alignment annotation tool.

```
<bead beadid=num src_length=slen \\
   tgt_length=tlen doc=docid ...>
//SOURCE_LANGUAGE:  sw1 sw2 ... swL
//TARGET_LANGUAGE:  tw1 tw2 tw3 .. twM
<wordstat>
<src>
sw1 status
sw2 status
.
.
.
swL  status
</src>
<tgt>
tw1  status
tw2  status
.
.
.
twM  status
</tgt>
</wordstat>
<alignment>
n1,n2  m1,m7     prb  <comment>
n3     m2,m3,m4  prb  <comment>
n4     m5        prb  <comment>
n5,n6  m6        prb  <comment>
</alignment>
</bead>
...
```

Here the status stanzas indicate the word level status as described above ("g", "f", "e", "x", "s", "u") on each word. The status is to be interpreted at the word level, but in the case that two words are linked then their status will be tied and can be interpreted as the status of the link. In the case of a group, however, the usage of separate status on the source and target words allows the annotator to indicate the status of a word within a group but percludes reflecting that status on the group link. Status changes are reflected on both sides of a link only if the connection is without a phrase. Otherwise the change affects only the current word. The lack of link status for groups is a slight design flaw and it is not considered major but and may be corrected later.

The alignment stanzas indicate the alignment between the source and target words. There are three major columns and then a comment field which is not necessarily preserved or generated by the hand-alignments. The first two columns indicate position indices starting at 1 which indicates the first position of each sequence. The columns have comma separated entries to indicate a group of words. There are no ranges in these columns: each word position must be indicated. A NULL token is assumed to exist in both the source and target language and it is optionally marked by the human annotator. Leaving a word without an alignment but marked by a status of "s" (spontaneous) will indicate the alignment of this word to either the SOURCE_NULL or TARGET_NULL. Position indices '0' and '-1' are both aligned to SOURCE_NULL or TARGET_NULL depending on which column it occurs in.

The first alignment line indicates a group (n1,n2) on the source side being connected to a group on the target side (m1,m7). Groups can be constructed from arbitrary indices on either the source or target side but can not include SOURCE_NULL or TARGET_NULL. Although an alternative method for showing group connections is shown below, we prefer the connections shown above so as to make explicit the notion of the source group.

```
<alignment>
n1    m1,m7     prb  <comment>
n2    m1,m7     prb  <comment>
n3    m2,m3,m4  prb  <comment>
n4    m5        prb  <comment>
n5,n6 m6        prb  <comment>
```

</alignment>

9 Conclusion

This data was created in the hope of advancing research in word-alignment and machine translation. Roughly one third of the data is selected from the Arabic Treebank Part I, which is available from the LDC. The LDC provides linguistic analysis of the Arabic side such as segmentation, part-of-speech tagging, and syntactic trees. English syntactic trees are also available from the LDC for portions of this data. We hope that this will advance research in projection of information, word order and syntactic structure differences between Arabic and English. Many errors in annotation might still be present in the data despite our best efforts to reduce these.

10 Acknowledgements

This work was partially supported by the Department of the Interior, National Business Center under contract No. NBCH2030001 and Defense Advanced Research Projects Agency under contract No. HR0011-06-2-0001. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the U.S. government and no official endorsement should be inferred. This paper owes much to the collaboration of the Statistical MT group at IBM and in particular to Mohammed Nasr who did most of the annotation described here. We also thank LDC (Linguistic Data Consortium) and NIST (National Institute of Standards and Technology) for releasing the source and translated material used in this corpus.

References

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

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 89–96, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.