Improving Word Alignment Quality Using Linguistic Knowledge

Bettina Schrader

Cognitive Science Doctorate Programme: Rules and Patterns
Institute for Cognitive science, University of Osnabrück
Kolpingstraße 7, D-49069 Osnabrück, Germany
bschrade@uos.de

Abstract

Word alignment of bilingual parallel corpora is usually generated using only statistical information. External linguistic information like e.g. a dictionary or linguistic structural annotation of the texts is used rarely, despite its usefulness. Additionally, it has to our knowledge never been examined systematically how linguistic information can be employed for word alignment improvement. In this paper, we present our experiments on finding out which linguistic information has which effect on word alignment quality, and we evaluate our experiments using precision and recall calculated for dictionaries that were generated after word alignment. The experiments show that information on e.g. lemmas and word category is useful to increase recall without lowering precision. Additionally, we discuss whether linguistic information can be used to compensate weak points of standard word alignment systems, and which features an ideal procedure should possess.

1. Introduction

Word alignment is an important technique in the exploitation of bilingual parallel corpora for lexicography, statistical machine translation, and cross-linguistic information retrieval (CLIR). It is used to automatically detect word pairs of translational equivalence, i.e. it computes which word in target language L2 is a translation of a word in source language L1.

Different word alignment techniques have been developed (cf. (Brown et al., 1990)), usually based on statistical information. Additionally, several researchers have experimented with combining linguistic and statistical information (Nießen and Ney, 2000). Still, the usefulness of linguistic information for word alignment has to our knowledge never been examined systematically.

The purpose of the experiments we are presenting here is to find out which linguistic information, whether on lemmas, word category or systactic constituency, can be used efficiently for word alignment. Additionally, we investigate which flaws standard alignment techniques have, and how they can be compensated. Experiments are evaluated using precision and recall calculated for 50-60 sample word pairs per corpus taken from automatically generated dictionaries after word alignment was done.

The paper is organized as follows: First, we give an overview on standard approaches to word alignment. Secondly, we introduce our the corpora and describe which linguistic tools were used for linguistic preprocessing. Then, we report on the experiments conducted and discuss their results.

2. Standard approaches to word alignment

Standard word alignment approaches like the ones by (Brown et al., 1990), (Brown et al., 1993), (Vogel et al., 1999), or (Hiemstra, 1996) make use of statistical models to derive word alignments.

(Brown et al., 1990) have been the first to publish a word alignment procedure. It consists of a cascade of five statistical translation models of increasing complexity. The first model of (Brown et al., 1990), IBM-1, treats every sentence as a *bag of words*, where the position of a word in a sentence does not have any influence on its translation probability. IBM-2 to IBM-5 refine this notion by introducing statistical weights such as *distortion* and *fertility* to account for word order phenomena and 1-to-many alignments.

The two competing standard alignment models, by (Vogel et al., 1999) and (Hiemstra, 1996) correspond most closely to the IBM-1 model: The HMM-model by (Vogel et al., 1999) treats a sentence mainly as a bag of words, but the probality of an alignment is influenced by the preceding alignment. (Hiemstra, 1996) uses a pure bag of words model. In contrast to (Brown et al., 1990) and (Vogel et al., 1999), he doesn't focus on the translation model, but instead uses word alignment as a means to generate a dictionary for CLIR.

All three approaches to word alignment do not use explicit linguistic knowledge, whether in form of a dictionary or in form of linguistic structural information, because these approaches are set up to be language independent, i.e. they are supposed to work equally well for each possible language pair. Researchers have, however, found it necessary to experiment on improving word alignment systems with linguistic knowledge: (Nießen and Ney, 2000) e.g. *manipulate* their parallel corpora: word order in one language e.g. is changed to resemble more closely word order in L2, in order to circumvent distortion problems caused by syntactic differences between L1 and L2.

3. Corpora

Three parallel German-English corpora were used for the experiments: debate protocols of the European Parliament (MLCC), a subset of the Linux manpages (MANPAGES), and a small corpus consisting of patent abstracts (PATENTE).

All corpora were tokenized, POS-tagged, and lemmatized using the tree-tagger by (Schmid, 1994). Two corpora were chunked using an extension of the tree-tagger (Schmid, unpublished) for the English, and the tool by (Kermes, 2003) for the German texts. All corpora were sen-

tence aligned using an aligner that was developed as part of the IMS corpus workbench and that combines various sentence alignment strategies (Evert, , p.c.).

Only *secure* sentence pairs from all corpora were used in the experiments, and manipulated to include only the kind of linguistic information that was necessary. For one experiment e.g., tokens were included in the input only if they formed part of a nominal or prepositional chunk. Sentence pairs were considered secure if they occurred in a sequence of at least three 1:1-alignments. This condition was applied to ensure that the text used in the experiments was 100% correct.

For word alignment, we used the alignment tool by Hiemstra, 1996), as it automatically generates a bilingual dictionary in easy-to-read format.

3.1. MLCC

This parallel text is part of the corpus *Multilingual and Parallel Corpora for Cooperation* (MLCC) provided by ELRA[1] and consists of debate protocols of the European Parliament between 1992 and 1994. They were preprocessed and added to the IMS corpus workbench independently of our experiments.

After sentence alignment and restricting the data set to secure sentence pairs, it consists of 1,713,796 tokens in 78,130 sentence pairs. In the course of the experiments, the set of sentence pairs has been reduced further to a random sample of 2500 sentences due to software restrictions.

3.2. MANPAGES

The MANPAGES corpus consists of texts from the Linux online help for shell commands that are available in English and German. They have been reformatted removing all paragraphs except the sections NAME / NAME, BESCHREIBUNG / DESCRIPTION and ÜBERSICHT / ZUSAMMENFASSUNG / SYNOPSIS as only these sections consist of coherent text. After preprocessing and applying the restriction on secure alignments, the MANPAGES consist of 14,759 tokens in 860 sentence pairs.

3.3. PATENTE

The smallest corpus consists of patent abstracts in German and English that were provided by courtesy of the German Patent Office. After preprocessing and reduction to secure alignments, the corpus is made up of only 125 sentence pairs with 3,204 tokens. Although this size is much too small for a statistical alignment method, it is used for the experiments as the translations provided are very good and close to the original texts.

4. Experiments

We test in several experiments how information on word category, lemmas and syntactic costituency influences word alignment quality. Two experiments and the baseline are carried out on all three corpora, while the other experiments are done on only one or two of the corpora for reasons given in each experiment description.

[1] http://www.icp.inpg.fr/ELRA/index.html

4.1. Baseline

To be able to compare the experiment results to what a pure word alignment procedure is capable, a baseline has been created: all corpora have been word aligned using only the sentence aligned text, i.e. no linguistic information has been used.

4.2. Functional Class Words

First, we removed all words belonging to a functional class such as determiner or preposition from the texts. Words of the lexical classes nouns, adjectives, and verbs, remained in the corpus. POS-tags are used to distinguish between both groups of words.

The reason for removing function words is that they are uninteresting from a lexicographic point of view as they don't carry lexical meaning. Additionally, the number of function words per language is fixed, so that they are probably listed in any existing dictionary, and can be aligned easily using one.

4.3. Lemmas

In morphologically rich languages, words may only differ from each other due to their inflections, while their meaning stays the same. If such word forms are aligned, each of them will be treated as unique and will be aligned as such, i.e. two word forms of the same lemma in L1 can be set into translational equivalence with two tokens from L2 that may or may not share the same lemma. This happened e.g. in the baseline for German *Verhandlung/ Verhandlungen* (English: *negotiation/ negotiations*): With this

Verhandlung		Verhandlungen	
translation	probability	translation	probability
you	0.65	negotiations	0.98
followed	0.31	process	0.02
All	0.03		

Table 1: Baseline dictionary excerpt: MLCC corpus

consideration in mind, we should not align word forms but rather abstract away from inflections and use lemmas for alignment.

In morphologically poor languages, on the other hand, favouring lemmas does not influence word alignment as much. We therefore refrained from lemmatizing the English texts. We have, however, lemmatized the German texts and aligned it with the unlemmatized English texts. Additionally, function words have been removed.

4.4. Lexicon

We also tested whether alignment quality is improved if we add data from an English-German dictionary, in this case the (Langenscheidts Handwörterbuch, 1991). For each corpus, a vocabulary list was compiled containing all nouns that occurred both in the corpus and in the dictionary, and the list was appended to the corpus. This procedure was necessary as the word aligner did not support direct lexicon lookup during the alignment process.

This experiment was carried out on the two corpora PATENTE and MANPAGES, only. MLCC proved too big for the addition of vocabulary in initial tests.

4.5. Morphology

Correctly aligning German compounds with their English equivalents is a problem for word alignment as German compounds usually correspond to English multi word units, i.e. they do not stand in a 1:1 relationship. The German compound "Dämpfungsscheibenanordnung" e.g. corresponds to the three subsequent tokens "dampening disk assembly" in English.

Splitting the compound in its components would solve this problem, however: "Dämpfungsscheibenanordnung" consists of the three elements "Dämpfung", "scheibe", and "anordnung" that can easily be aligned with the three elements of the corresponding English expression in three 1:1 alignments[2]. Therefore, we decomposed all German complex nouns of the PATENTE corpus using the morphological tool DEKO (Schmid et al., 2001) and replaced them by their decomposed sequence of elements before aligning the corpus.

This experiment was carried out only on the smallest corpus, the PATENTE corpus, as compound decomposition is a very time-consuming task.

4.6. Chunks

In our final experiment, we tested whether shallow syntactic information is useful for word alignment, too. For this reason, the corpora MLCC and MANPAGES were chunked, and all tokens that did not belong to a nominal or prepositional chunk were deleted.

The MANPAGES corpus has proven too sloppily translated to allow for successfull chunking, so that we have not run this experiment on this corpus.

5. Evaluation

For the evaluation, we constructed tokenlists and compared them to the dictionaries generated during word alignment. Precision and recall were chosen as evaluation measures, and we examined only the translation direction German → English.

For each corpus, we compiled a tokenlist containing the 50-60 most frequent nouns of the corpus[3]. and translated them manually. This sample size is small enough to allow for manually examining the data, and sufficiently big to allow an analysis of the experiment results. We restricted the tokenlists to nouns, because new words are often created as such. We defined precision and recall such that:

$$\text{precision} = \frac{\text{\# correct translations}}{\text{\# suggested translations}}$$

and

$$\text{recall} = \frac{\text{\# correct translations}}{\text{\# manually assigned translation}}$$

The number of translations is given by the number of words of the English translation. In the case of a multi word unit like “child process”, each element is counted as correct translation candidate, i.e. “child process” counts with two correct translations.

Translation candidates of the dictionaries were ignored if their translational probability was below 10%.

Precision (%)	MLCC	MANPAGES	PATENTE
Baseline	59	64	35
Function words	54	58	43
Lemmatization	50	46	46
Lexicon	–	47	53
Morph. Decomposition	–	–	37
Chunks	55	–	42

Table 2: Precision values for all experiments

As can be seen in the tables, the precision of the dictionaries created during the experiments is lower than the value of the baseline. The only exception is the results of the PATENTE corpus, where all experiment precisions are higher than in the baseline.

Recall, on the other hand, is higher in all experiments on all corpora and increases up to 98%.

recall (%)	MLCC	MANPAGES	PATENTE
Baseline	90	84	67
Function words	95	84	91
Lemmatization	95	87	88
Lexicon	–	90	89
Morphology	–	–	76
Chunks	98	–	71

Table 3: Recall values for all experiments

To find out why precision values for the experiments are lower than the precision of the baseline, the dictionaries were more closely examined: We found out that the number of translation candidates per token is higher in the experiment dictionaries than in the baseline. Additionally, the baseline dictionary has a lower coverage than the other experiment dictionaries.

Precision as calculated here obviously does not describe dictionary quality completely enough: For once, it punishes alternatives - the more translation candidates are given per token, the lower precision will be. Secondly, precision is higher if a word is missing from the dictionary then if it is listed with at least one wrong suggestion (See example in table 4), i.e. differences in coverage are not taken into account.

Headword: Ergebnis (result)				
Experiment	word	probability	word	probability
Baseline	no suggestions			
Function Words				
Lemmatization	results	0.97	portable	0.03
Lexicon	result	1.00		

Table 4: Dictionary excerpts: Manpages corpus

If we take the problems with calculating precision into account, we assume that linguistic processing does not influence precision negatively despite evaluation numbers.

[2] Linking elements have been omittd for this example.

[3] 50 tokens each were chosen for PATENTE and MANPAGES; the tokenlist for the corpus MLCC contains 60 items as it is bigger than the other two corpora.

The analysis of the experiments shows as well, however, that some word alignment problems remain: Using linguistic information by means of text manipulation always means restricting oneself to one kind of knowledge, as a statistical model like that by (Hiemstra, 1996) allows for only one level of linguistic description – chunks e.g. can be used iff sentences are made up only of chunk material. Hence there is always some loss of information. Additionally, information on sentence-internal structure, like e.g. chunk boundaries, cannot be preserved and used as alignment clues: If we restrict the input to words occurring in noun or prepositional chunks and mark chunk boundaries, the alignment tool treats chunk boundaries in the same way as words.

Finally, a simple bag of words model is not able to align single words with multi word units correctly, as is necessary in the case of German compounds and their corresponding English multi word units. Even a morphological decomposition of compounds does not help much, as is seen in the experiments. The reason is that we cannot expect that the equivalent of a compound is a complex expression in itself - the German compound "Abstandselement" e.g. is equivalent to simplex "spacer". Additionally, even if the translation of a compound is morphologically complex, it need not be compositional as well: German "Schutzelement" is translated by "shield cushion" - where there is no correspondence between German "Element" and English "cushion" ("cushion" translated to German means "Kissen", "pillow").[4]

6. Conclusion

In this paper, we systematically investigated which linguistic information can be used for improving word alignment quality: Lexical information and information on lemmas, word category, morphology and syntactic constituency were used to manipulate three parallel corpora before aligning them. Afterwards results were evaluated calculating precision and recall for the dictionaries generated during word alignment, and the dictionaries were examined in more detail. Experiment results show that linguistic information is useful in increasing recall. Precision as calculated here is not sufficient to determine the influence of linguistic information on word alignment in terms of correctness of the established translation correspondences. We have reason to assume, however, that precision was not decreased during the experiments.

However, using linguistic information for sophisticated text manipulation does not compensating flaws of a standard word alignment approach: Using it means loss of information elsewhere, and sentence-internal structure cannot be used as alignment clues.

A word alignment system should be able to parse linguistically annotated text, so that one level of linguistic description, e.g. lemma information, can be used to align while preserving all other information, e.g. on word forms. Additionally, it should be able to parse and preserve sentence-internal structure, e.g. chunks: if two chunks c1 and c2 are equivalent toeach other, then the words in c1 and c2 will be equivalent to each other as well. Concerning multi word units, it should be possible to align across levels, so that a word in L1 (e.g. a German compound noun) is aligned with its corresponding chunk in L2 (an English multi word expression).

[4] All examples are taken from the PATENTE corpus.

7. Acknowledgements

We thank the German Patent Office for supplying the texts for the PATENTE corpus, and Langenscheidt for the permit to use the electronic version of the Langenscheidts Handwörterbuch, 1991).

8. References

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, 1990. A statistical approach to machine translation. *Computational Lingusitics*, 16:79–85.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, 1993. The mathematics of machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Evert, Stefan. personal communication.

Hiemstra, D., 1996. *Using statistical Methods to create a bilingual Dictionary*. Master's thesis, Universiteit Twente.

Kermes, Hannah, 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 9, number 3.

Langenscheidts Handwörterbuch, 1991. Langenscheidts Handwörterbuch Deutsch / Englisch, Englisch / Deutsch. 3. Auflage.

Manning, Christopher D. and Hinrich Schütze, 1999. *Foundations of statistical natural language processing*. Cambridge, Massachusetts, London: MIT Press.

Nießen, Sonja and Hermann Ney, 2000. Improving SMT quality with morpho-syntactic analysis. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbruecken, Germany.

Schmid, Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*. Manchester, England.

Schmid, Helmut, unpublished. The IMS Chunker. Unpublished manuscript.

Schmid, Tanja, Anke Lüdeling, Bettina Säuberlich, Ulrich Heid, and Bernd Möbius, 2001. DeKo: Ein System zur Analyse komplexer Wörter. In *GLDV - Jahrestagung 2001*.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann, 1999. HMM-based word alignment in statistical translation. In *Proceedings of the International Conference on Computational LInguistics*. Copenhagen, Denmark.