ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation
www.accurat-project.eu
Project no. 248347

Deliverable D2.1
Report on Application of Existing Alignment Methods to Comparable Corpora

June 30, 2010

Document Information

Deliverable number:	D2.1
Deliverable title:	Report on Application of Existing Alignment Methods to comparable Corpora
Due date of deliverable:	30/06/2010
Actual submission date of deliverable:	30/06/2010
Main Author(s):	David Guthrie, Ahmet Aker, Robert Gaizauskas, Evangelos Kanoulas, Monica Paramita, Mark Sanderson
Participants:	USFD, Tilde, RACAI
Internal reviewer:	Gregor Thurmair, Linguatec
Workpackage:	WP2
Workpackage title:	Multi-level Alignment Methods and Information Extraction from Comparable Corpora
Workpackage leader:	RACAI
Dissemination Level:	PU: Public
Version:	V0.2
Keywords:	Alignment, Machine Translation, Comparable corpora

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
0.1	28/04/2010	Draft	USFD	Draft	First internal draft
0.2	15/06/2010	Draft	USFD	Draft	For internal review
1.0	??/??/2010	V1.0	Linguatec	Comments and Corrections	Revised version

Executive Summary
This document reports on existing alignment strategies for translated texts. We investigate the applicability of strategies to corpora with different levels of comparability and establish criteria for their application.

Contents

List of Tables

List of Figures

Chapter 1

Review of Existing Alignment Methods

The term *alignment* is used in the context of machine translation to describe the pairing of text in one document with its translation in another. Alignment is commonly performed for texts that are translations of each other (called parallel texts), but it is also possible to produce a type of alignment between texts that are not parallel but may be comparable to each other (e.g. they are about the same topic or are in the same genre). Discovering ways to exploit comparable corpora to improve Machine Translation is of course the overall aim of the Accurat Project. In this chapter we begin (section 1.1) with a summary of work on alignment of parallel corpora at different levels of granularity. Much of this work is not directly applicable to comparable corpora, but it is of particular interest because of the approaches used. The following section of this report (section 1.2) reviews different approaches to alignment of comparable texts; we focus on approaches for the alignment of sentences, words, and phrases. The applicability of these methods to different levels of parallel texts is discussed in Chapter 2. Chapter 3 presents a case study of the application of four different alignment techniques to corpora of different levels of comparability.

1.1 Parallel Text Alignment

Texts which are translations of each other in distinct languages, call them L_1 and L_2 are often referred to as *parallel* and for these texts we use the term *alignment* to describe the process of identifying the correspondences between these texts. Research on alignment has generally focused on parallel texts which are direct and literal translations of each other, so that there exists a very strong correspondence between the texts and sentences in the L_1 text correspond to those roughly at the same position in the L_2 text. Parallel texts are commonly aligned at the level of sentences initially and once this mapping between sentences and their translations is established it is often used to induce a finer-grained mapping of the words or phrases that are translations of each other. In this section we give an overview of existing techniques for alignment of parallel texts first at the sentence level and then at the word and phrase levels.

1.1.1 Sentence Alignment

Sentence alignment of parallel texts has often been a prerequisite before these texts are used for statistical machine translation and consequently has received a great deal of research. The sentence alignment task for parallel texts in normally approached ignoring the possibility of crossing correspondences for parallel texts because the order of sentences rarely changes during direct translation. Most methods do, however, allow for one to many or many to one alignments as would be the case if a sentence was translated using two sentences. Approaches to sentence alignment have generally made use of the length of the sentences in the two texts, the distribution of words, or some combination of these factors.

Early work in this area [Gale and Church, 1991, 1993] demonstrated that sentences from parallel text can be aligned with high accuracy by matching sentence sequences that have similar length. The intuition is that sentences that are longer in one language should

correspond to sentences that are longer in another language. This intuition has been shown to be fairly reliable when languages are similar and translations are very literal. Gale and Church [1993] compute a maximum-likelihood estimation using dynamic programming and select the alignment from all possible alignments that has the highest probability. This method makes use of the difference between the length of sentences in characters and uses a reduced set of possible alignments: 1:0, 0:1, 1:1, 2:1, 1:2, 2:2. The basic method achieves a 4% error rate on a parallel corpus of United Bank of Switzerland economic reports in English, French, and German. This work additionally shows that if the entire corpus is not required it is possible to pick the top 80% most probably alignments and achieve an error rate of 0.7%. A similar approach was independently proposed in Brown et al. [1991] who formulated the problem using a hidden Markov model (HMM) and compared the length of sentences in words rather than in characters. Although these length based methods perform well they are not robust with respect to noisy parallel data and to achieve high accuracy they require that texts already be aligned (or anchored) with markers into relatively small units, say paragraphs, before the sentence alignment algorithm begins.

An influential method of alignment was introduced by Church [1993] called *char_align* that does not require the presence of initial markers in the parallel texts, but attempts to produce a character alignment of parallel texts rather that a sentence alignment. The method is based on the presence of *orthographic cognates*, words that have a similar spelling between the two languages because they have similar meaning; for example, the English word *quality* and the French word *qualité* share 6 of 7 letters in common. Alignment is performed by matching identical sequences of character 4-grams in the two texts. This relies on the languages having similar alphabets and writing styles, but Church [1993] suggest that even very different languages can share a large number of proper nouns, numbers, and punctuation. Matched character sequences are used as markers in the texts and are weighted by their frequency, so rare character grams that match are more indicative of a

true mapping. An extension of this method was proposed by Dagan et al. [1993] called *word_align* that uses the output of char_align to produce an alignment of words.

The presence of orthographic cognates across languages is also exploited in Melamed [1999], who uses them to produce a sentence alignment. However instead of looking for matching sequences of characters as in Church [1993], cognates are identified in Melamed [1999] by measuring the similarity of the spellings between pairs of words using the longest common subsequence ratio (LCSR).

$$LCSR(X,Y) = \frac{\text{length}[LCS(X,Y)]}{\max[\text{length}(X), \text{length}(Y)]} \tag{1.1}$$

where LCS is the longest common subsequence between two strings and characters in this subsequence need not be contiguous. Methods exist for the efficient computation of LCS between two strings in $O(n \log \log n)$ time [Bergroth et al., 2000]. Melamed [1999] allows for the use of a bilingual dictionary to identify corresponding words if one is available and notes that it would be possible to make use of phonetic cognates, based on the similarity of two words sounds for languages with different alphabets. Some of the corresponding word pairs are pruned based on their ambiguity and then they are used (along with sentence boundary information) to align the sentences. An iterative method of sentence alignment is then employed which greedily searches for chains of word correspondences. This method achieves 98% accuracy on English – French Hansard Data and is more robust than many length based approaches because it does not require data to be pre-aligned at the paragraph level.

Chen [1993] uses a lexical approach to sentence alignment that works by constructing a simple word to word translation model on the fly and then choosing the alignment that maximizes the likelihood of generating the corpus given the translation model. The translation model is a simple word based model that is bootstrapped from a small corpus of

100 sentences pairs that have been manually aligned and then the EM algorithm is used to incrementally reestimate parameters. Chen estimates the error rate is 0.4%. This method is more robust to large sections of text that do not have a translation than early methods based solely on sentence length.

Kay and Röscheisen [1993] employ a different kind of lexical approach to sentence alignment based on the similarity of word distributions in their respective texts. A partial alignment of words is used to induce sentence level alignments. The method works by finding pairs of sentences which contain many possible lexical correspondences. Initially, the first and last sentences of the parallel texts are paired and marked as anchors and then an iterative procedure is applied:

1. The Cartesian product of the list of sentences in one language with the list in the other language is computed to get all possible pairings. Sentences pairs are excluded if they cross anchors or their distances from anchors differ by a large amount.

2. The next step is to compute all pairs of words that co-occur in these potential alignments. For this task, the authors employ the use of the Dice coefficient:

$$Dice(i, j) = \frac{2 \cdot C(s_i, t_j)}{C(s_i) \cdot C(t_j)} \tag{1.2}$$

   where $C(s_i, t_i)$ is the number of co-occurrences of the ith word in the source language and the jth word in the target language and $C(s_j)$ is the total number of occurrences of the source word i and $C(t_j)$ is the total number of occurrences of word j in the target language. The words with the highest similarity scores are recorded after removing low frequency words, say below 3 occurrences, because they may be unreliable.

3. For every word pair from the last step, the pairs of sentences that contain these words

have an association score that is incremented. Sentences with an association above a minimum number of times are fixed as anchors and the algorithm repeats.

An automatic method of obtaining normalized forms of words is used, so that different morphological variants will match within the same language. It is important to note that like the length based methods described earlier, this approach does not make use of any extra lexical resources or even explicit cognate matching across languages. After 4 passes of this algorithm on a corpus of pairs of Scientific American articles in German and English, Kay and Röscheisen [1993] achieve a correct alignment that covers 96% of the sentences.

A similar method to that of Kay and Röscheisen [1993] is employed in Haruno and Yamazaki [1997], but using mutual information and t-score to judge similarity of words (rather than the Dice coefficient). Interestingly, Haruno and Yamazaki [1997] note that function words impede alignment for their task of English – Japanese alignment and so their method makes use only of content words. Additionally, this work shows that the use of an online dictionary to find matching word pairs improves alignment because many words do not occur with enough repetition to identify correspondences. This method is tested on a diverse collection of parallel English and Japanese texts and achieves an average precision and recall of 95%.

An open source lexicon-based toolkit for sentence alignment called *Champollion*[1] was introduced by Ma [2006]. This method uses a dictionary to identify translated words between two documents and then weights these lexical matches by assigning higher weights to less frequent words, much as Church [1993] did for character grams. The intuition here is that infrequent words (e.g. proper names) give much stronger evidence that two sentences align than frequent words (e.g. closed class words). Ma [2006] make use of a true bilingual lexicon in the toolkit, but there is nothing in the method that would prevent the use of an automatic lexicon built of cognates and similar spellings across languages. Champollion

[1] `http://champollion.sourceforge.net/`

achieves approximately 97% precision and recall even on relatively noisy Chinese – English parallel corpora.

A two pass alignment technique is introduced in Moore [2002] that makes use of both lexical and length based approaches for sentence alignment. First, Moore uses a version of the length based alignment of Brown et al. [1991] to search for the best alignment of sentences. This search is limited by only looking for alignments that are near the same position in both texts, but if the best alignment is near the edges of this range the search space is widened. Of these alignments the 1 to 1 alignments with the highest probability are used for training a word translation model based on IBM Model 1 [Brown et al., 1993]. A combination of the translation model with the length based model is used to perform the final sentence alignment. This method achieves recall and precision above 99% on parallel software manuals and additionally does not require the use of a parallel corpus, anchor points, cognates, or a bilingual lexicon and requires only modest computational time. Code for Moore's *Bilingual Sentence Aligner* is publicly available[2].

1.1.2 Word and Phrase Alignment

Word alignment in statistical translation refers to the mapping between the source words and target words in a set of parallel sentences. Alignments at the word level are typically much more complicated than sentence level alignments and can include re-orderings, mappings of one-to-many words, omissions, and insertions. Most work on alignment stems from the source-channel approach to statistical machine translation [Brown et al., 1993] where for any pair of parallel sentences we search for the alignment that maximizes the probability of the source language sentence given the target language sentence.

$$\hat{a}_1^J = \underset{a_1^J}{\operatorname{argmax}} P(s_1^J, a_1^J | t_1^I) \tag{1.3}$$

[2] `http://research.microsoft.com/en-us/people/bobmoore/`

where a_1^J is an alignment describing a mapping from source language string s_1^J to target language string t_1^I. Five models for the computation of this probability were proposed by Brown et al. [1993] and have greatly influenced research in this area. The models increase in complexity from a simple 1 to 1 model that does not take into account the order of words – to models that allow words to be aligned with sequences of words. These models have also been extended in various ways with some improvement in alignment accuracy or speed; for instance, Zens et al. [2004] uses a symmetric model that is trained in both directions and introduces a smoothed lexicon that explicitly takes into account word base forms.

The IBM models proposed in Brown et al. [1993] have been implemented and made publicly available in a toolkit developed by Franz Josef Och called *Giza++*[3] [Och and Ney, 2000, 2003]. The toolkit implements the five IBM models as well as a hidden Markov model (HMM) and various refinements.

A robust multi-stage word alignment method is presented in Tufiş et al. [2010] which makes use of a rich variety of lexical and morphological features using part-of-speech processing, lemmatization, and chunking. This method uses two different alignment strategies and combines the results using an SVM classifier to filter out improbable alignment links. This first alignment method, called *YAWA*, is a multi-stage aligner that incrementally adds alignment links at every stage. This aligner uses word pairs log-likelihood scores from a sentence aligned parallel corpus to first align content words and then makes use of chunkers in both languages to align phrasal chunks (i.e. noun phrases, verb phrases, prepositional phrases, etc.). The second alignment approach, called *MEBA*, also generates links step by step, but possible links are computed using a linear combination of several features. Many of these features are novel to this work and make use of a wide variety of information including: locality of words, Giza++ alignments, part-of-speech affinity, string similar-

[3] `http://code.google.com/p/giza-pp/`

ity, and content word bigrams. The combination of the aligners using an SVM classifier to select probable links, achieves an error rate of 16% on an English – Romanian parallel corpus. A previous version of this method [Tufiş et al., 2005], with a higher error rate, competed in the 2005 ACL Workshop on"Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond" and finished first out of 37 competing systems. Similar discriminative feature based techniques have also demonstrated very low error rates for other language pairs [Lacoste-Julien et al., 2006; Moore, 2005; Moore et al., 2006].

Limitations of the word based models for machine translation has lead to the incorporation of phrased based models that derive multiple word to multiple word alignments [Och and Ney, 2004; Marcu and Wong, 2002; Koehn et al., 2003], so that whole sequences of words can be translated as a singe unit. A state-of-the-art phrase based machine translation system called *Moses*[4] [Koehn et al., 2007] has been developed as part of the EuroMatrix project and is open source and publicly available. The Moses system extracts phrasal alignments from corpora which have first been word aligned using the Giza++ toolkit.

These phrase based methods, much like the word based methods, still require parallel aligned sentences in order to identify alignments and so, for the most part, cannot be directly applied to non-parallel or comparable corpora (where we generally do not have a mapping between sentences). Extracting alignments from comparable corpora is the focus of the next section.

1.2 Comparable Text Alignment

For many language pairs and in specialized domains, it is difficult to obtain the volume of parallel texts necessary to train a machine translation system. This has lead to research into the use of more readily available corpora which are not parallel, but may be *compa-*

[4] `http://www.statmt.org/moses/`

rable to each other (i.e. because they are about the same topic) or they may be simply unrelated monolingual corpora. Alignment of non-parallel corpora requires that our definition of alignment is broad enough to include identifying any correspondences between the texts, be they sentences, phrases, or merely words, because it is likely that only a small percentage of the texts will correspond. Much of the research in this area is for the purpose of extracting bilingual dictionaries (either for use by human translators or for machine translation systems), but we can view these dictionaries as an alignment between the words or phrases in the non-parallel texts. This section presents a review of research on extracting these types of alignments from non-parallel and comparable corpora grouped by their focus on sentences, words, or phrases.

1.2.1 Comparable Sentence Alignment

Sentence alignment for comparable or non-parallel corpora can be a much different task than for parallel texts. For comparable corpora the task is to identify any parallel sentences these corpora may contain without knowledge about the position in the documents these alignments are likely to occur. This task is closely related to cross language information retrieval; for instance, we can view sentences in one language as queries and attempt to retrieve the sentences that are most similar in the other language. Although comparable corpora are not direct translations of each other, they may contain parallel sentences. For example, comparable corpora made up of collections of newswire may contain stories that are on the same topic and express the same idea in sentences across languages. Much of the research in this area breaks the problem up by first identifying documents that are likely to be comparable to each other in a large collection of text and then identifying any parallel sentences this subset may contain [Zhao and Vogel, 2002; Yang and Li, 2003; Barzilay and Elhadad, 2003; Utiyama and Isahara, 2003; Fung and Cheung, 2004; Munteanu et al., 2004; Munteanu and Marcu, 2005; Tillmann, 2009; Tillmann and Xu, 2009].

Munteanu et al. [2004] and Munteanu and Marcu [2005] introduce a method of identifying parallel sentences in comparable corpora to aid statistical machine translation. This work considers a scenario in which some parallel training data is available to train a machine translation system and attempts to improve this system using comparable corpora. The approach in this work is to extract additional parallel sentences from the comparable corpora and use these to augment the initial parallel training data. Munteanu and Marcu [2005] suggest that available comparable corpora may be much closer to the domain or topic of the text you are interested in translating than the parallel training corpus, so the additional extracted parallel sentences may make a large impact on translation quality. This work makes use of a translation dictionary which is learned automatically from parallel corpora along with the comparable corpora collection which is comprised of news feeds in English, Arabic, and Chinese. Their algorithm for identifying parallel sentences can be broken down into 3 main steps.

1. The first step is *article selection*, where candidate documents are selected that are likely to contain parallel sentences. For every document in one language in the comparable corpora, the top 5 translations of every word (using the dictionary) are used to create a query to search all documents in the other language. This search is limited to articles published within 5 days of the source text and only top 20 ranked articles are returned.

2. Next, candidate sentence pairs are selected. All possible sentence pairs are generated and then they are filtered using simple heuristics. For instance, if the ratio of their lengths is greater than two or if less than half of the words have a translation in the sentence using the dictionary.

3. Lastly, a maximum entropy classifier is used to determine if the sentence pairs are parallel or non-parallel. The classifier uses a collection of "general features" (e.g.

sentence lengths, length difference, length ratio, the percentage of words which have translations) and also "alignment features" (these are computed using a symmetric modified version of IBM Model 1 Brown et al. [1993] to create features like the number of words that have connections and the longest contiguous connected span). The classifier is trained by creating collections of positive and negative examples created from a parallel corpus.

Munteanu and Marcu [2005] report that this method of extracting parallel sentences has a precision around 97% and a recall of 40% on their test collection. The authors additionally show an improvement to translation quality when adding the extracted parallel sentences to the initial parallel corpora used to train their machine translation system.

A similar method of identifying parallel sentences from a collection of comparable corpora is presented in Fung and Cheung [2004]. The method differs in that its focus is on more disparate or weakly comparable corpora by exploiting bootstrapping using extracted parallel sentences to extend their lexicon and then re-searching for sentence pairs. Much like Munteanu and Marcu [2005], this work makes use of a bilingual dictionary, but it employes the use of the cosine similarity measure to identify likely comparable texts and likely sentence pairs. The method then uses IBM model 4 to learn word translations and add new translations to the dictionary and recompute sentence pairs. The final step in this iterative method is to use the extracted parallel sentences to search again for documents with sentences similar to each of these sentences and then repeat the procedure. Convergence occurs when the number of sentence pairs and the size of the lexicon remain constant. Human evaluation of the sentence pairs is used to determine whether the pairs are parallel and they report that their approach gives a 50% improvement in precision over using a non-bootstrapping approach.

An additional contribution of Fung and Cheung [2004] is a proposal for a way to quantify the "parallelness" of bilingual corpora. This measure is computed using the matched

bilingual sentences pairs that their method produces as the sum of the mutual information of the set of word pairs that appear in the corpora, C.

$$S = \sum_{(w_e, w_f) \in C} \frac{f(w_e, w_f)}{f(w_f) f(w_e)} \tag{1.4}$$

where $f(w_f)$ and $f(w_e)$ are respectively the occurrence frequency of the word w_f in the source language and the frequency of the word w_e in the target language. The term $f(w_e, w_f)$ is the co-occurrence frequency of the pair of words in all matched sentence pairs. They show that using this formula "more parallel" corpora score higher.

Another variation of these approaches for under resourced languages or domains is to search for similar documents and sentences using features that are less reliant on the initial translation dictionary. Do et al. [2009] investigate aligning parallel sentences in French and Vietnamese comparable texts with possibly a very small or even non existent translation dictionary. As in the approaches described above the first step is to identify similar documents. This is achieved in Do et al. [2009] by searching for documents using: publication date, document length, special words (numbers, punctuation, and similarly written named entities) and the *Champollion* [Ma, 2006] sentence aligner to identify the proportion of sentences in the documents which align. The sentence aligner and special words (or lexicon) are then used to further filter the sentence pairs. They produce a corpus of 50 thousand parallel sentences and use this to train the Moses [Koehn et al., 2007] machine translation system. Impressively, the authors report BLEU scores higher than using Google Translate in both French to Vietnamese and also Vietnamese to French translation.

Abdul-Rauf and Schwenk [2009a] and Abdul-Rauf and Schwenk [2009b] also extract parallel sentences from comparable corpora, but make use of a full French to English machine translation system. The Moses machine translation system in trained on parallel

corpora and a language model is built using the English Gigaword corpus of newswire Graff [2003]. Comparable documents are first translated using the model and then each of the sentences in this data are used to search for similar sentences using an information retrieval toolkit. The retrieved similar sentences are then judged to be parallel or non-parallel based on word error rate (Edit Distance), and translation error rate. This work also introduces an interesting method of sentence tail removal, to better account for the situation where two sentences are matched as parallel, but one has some extra information added to the end of the sentence. Their method removes the extra information, so that phrases will align more accurately. Using this method an 11 million words of aligned parallel sentences are extracted from comparable corpora and added to the training for the translation system. Results show an improvement in translation quality of 2.5 BLEU points when adding the additional sentences.

1.2.2 Comparable Word Alignment

Extracting bilingual dictionaries or lexicons from non-parallel corpora can be viewed as a type of word alignment for comparable texts where we would not expect to align all words in a text because many of the words are from non-parallel pieces of text that have no translations in the corpus. Much of the research on this topic makes use of the idea that lexical terms in their respective languages have similar co-occurrence patterns and these patterns can be used to find words that are likely to be translations of each other [Fung, 1995; Rapp, 1995; Fung and McKeown, 1997; Fung and Yee, 1998; Rapp, 1999; Diab and Finch, 2000; Fung, 2000; Chiao and Zweigenbaum, 2002].

Rapp [1999], for example, produce a large lexicon of German to English word translations from comparable corpora using only an initial seed dictionary. The method first creates co-occurrence vectors for every word in the target language (in this case English) with all other words. Words are considered to co-occur if they occur within a 3 word

window of each other and the position of this occurrence is recorded. So, for an english vocabulary of size $|V|$, for every word a vector containing 6 vectors of size V are stored which hold words which appeared in position -3,-2,-1,1,2 and 3 (e.g. the vector of position -2 means all words that appeared two words to the left of the target word) and these 6 vectors are concatenated to produce one long vector of size $6 \times |V|$. Instead of storing raw co-occurrence counts, the log-likelihood ratio [Dunning, 1993] of every co-occurrence is stored to give an indication of how much more likely this pair is to occur than by chance. A similar co-occurrence matrix is constructed for all words in the source language, but only for those words which have translations using the initial seed dictionary and the ordering of the counts in the vectors is set to match the target language vectors. Rapp [1999] makes use of only the first translation in the dictionary for every term, but a similar method by Fung and Yee [1998] makes use of all possible translations. The final step is to take each word in the source text and compare its vector with all vectors in the target language to determine which word is most similar (in this case city block distance gave the best results). A precision of 72% is reported for 100 test words by comparing the automatically constructed translations to a reference dictionary. Variations to this approach have used modifiers and predicates rather than co-occurrences [Yu and Tsujii, 2009] and Gaussier et al. [2004] used geometric projection techniques to attempt to improve the alignment.

A very different method is presented in Diab and Finch [2000], where again vectors of co-occurrences are constructed for both languages (they measure co-occurrences for every word with only the 150 most frequent tokens and a window size of 2), but this method is particularly notable because it does not make use of any parallel texts or a dictionary, only the 150 most frequent words. Instead of mapping the vectors of words using dictionary entries, as in Rapp [1999], a distance matrix is created for the words in each language using

the Spearman Rank Correlation formula[5]:

$$\mathcal{S}(\vec{v}_1, \vec{v}_2) = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{4n(n^2-1)} \tag{1.5}$$

where $(x_i - y_i)$ is the distance between the i^{th} item's rank in vector x and that item's rank in vector y and n is the length of the vectors. The words are then aligned across languages using the gradient decent algorithm and looking for a mapping that makes the distance matrices most similar. Interestingly, the authors test this algorithm by taking two comparable corpora in the same language and pretending that one of the documents is written in a different language; the test is then to see if words map to the identical words across corpora simply based on their distributional characteristics. An accuracy of greater than 92% is reported for the 2000 most frequent words from the test corpora.

A evaluation of various approaches is performed in Koehn and Knight [2002] for a German to English translation system. This work creates an initial seed lexicon of 1000 words by choosing all words that appear identically in both languages. They evaluate this dictionary and determine it is 88% accurate. This lexicon is used with the algorithms of Rapp [1999] and Diab and Finch [2000] on a collection of comparable corpora, but using greedy matching to align words rather then by using dynamic programming to choose an optimal alignment. In addition to these methods, two additional methods are used to align words. The first uses word frequency to align words and the other chooses aligned words based on the longest common subsequence measure. These methods are used to align nouns only and are evaluated against a bilingual dictionary. The results show that the similarity measure of Rapp [1999] performs best, but that an improvement can be achieved by combining all of these methods. The combined system achieves an accuracy

[5]The formula we present here is not the one that appears in Diab and Finch [2000], but is instead the corrected version that appeared in Koehn and Knight [2002]

of 39% on 1000 lexical entries from a test corpus.

A similar combination of approaches is performed in Haghighi et al. [2008], where a co-occurrence similarity measure [Rapp, 1999] is combined with a string similarity measure. Canonical correlation analysis and the Viterbi EM algorithm are used to identify the most probable alignment (rather than the greedy matching approach used in Koehn and Knight [2002]). The authors report an accuracy of 61.7% for a German / English dictionary in a task similar to Koehn and Knight [2002].

1.2.3 Comparable Phrase Alignment

Weakly comparable corpora may contain very few or no parallel sentences, but still may contain alignments that are longer than a single word. This section reviews research on the extraction of multiword sequences or phrase alignments from comparable corpora. Some of the research in this area has no doubt been influenced by the current move towards phrase based statistical machine translation, where instead of providing the translation system with word translations we would like to provide the system with phrase translations, possibly learned using comparable corpora. In theory, phrase based methods are particularly appealing because they are not limited to the alignment multi-word sub-sentential phrases, but can choose to align the longest sequences of words possible, from single words to entire sentences.

Munteanu and Marcu [2002] present an approach which makes use of suffix trees to perform phrasel alignment in comparable corpora. The suffix tree is a data structure for encoding a string, A, by storing all possible suffixes so that it is possible to efficiently determine if a string, B, is a substring of A and where it occurs. In this research, the authors separately encode the source and target texts as suffix trees of words (rather than characters) that do not cross sentence boundaries. A small bilingual dictionary (6,900 words) automatically derived from a parallel corpus is then used to attempt to match the

longest possible substrings between the two trees. In this approach, phrases can only align across languages if all words in the phrases have some translation to each other in the dictionary. This method is run on a 1.3 million word English – French comparable corpus and produces almost 40 thousand parallel sequences of length 3 o 7 words. 95 out of 100 randomly selected alignments are judged to be correct. An extension of this method is presented to learn translations for words that do not contain mappings in the dictionary. The idea behind this extension is to identify matching aligned sequences that occur before and after words and then assume that the words must be translations as long as the left and right matching contexts are of a certain length (e.g. at least 3 words). They report a precision of 30% for alignments identified for unknown French words.

A related method of phrasal alignment that is more robust to longer sequences of unknown words is presented in Sharoff et al. [2006]. This work introduces a tool to identify possible phrase translation equivalents of words using comparable corpora for presentation to human translators. The method takes a dictionary based approach but looks for translations of words that are related because they occur in similar contexts. The method begins with a phrase to translate and expands each word by looking for all words similar to it in the same language using co-occurrence statistics [Rapp, 1999]. This expanded set of words is then translated using a bilingual dictionary and then all terms are further expanded in the target language using co-occurrence statistics. All possible combinations of the multi word terms are then computed and ranked by frequency of occurrence in the target language after some basic grammar guidelines are applied to prune this set (e.g. phrases don't end in a preposition). The reported precision of this tool is low but human evaluation suggests that it has high recall and is useful to human translators who can browse the top results retuned by the system.

A signal processing inspired approach is presented in Munteanu and Marcu [2006] to extract parallel sub-sentential fragments from comparable corpora. This approach requires

a dictionary which they produce from parallel texts in two different ways. The first dictionary is produced using Giza++ to generate an alignment for use as a lexicon and the second dictionary is produced identically except that it is refined using log-likelihood ratio scores to only pick words that co-occur with high probability. They show that when this pruned dictionary is used with their phrase alignment method, it gives better results than using the full Giza++ derived dictionary. Their method starts by searching for similar documents using the dictionary to translate all words and then using these words as a query to an information retrieval toolkit. The top 20 documents are returned and from these, all sentences pairs are created and those with few translations of each other (again using the dictionary) are discarded. This initial process is very similar to the procedure used in Munteanu and Marcu [2005]. The next step is to take each sentence pair and greedily align all words that appear in the dictionary based on their log-likelihood ratio. Words which do not have alignments in the dictionary are aligned with the word that gives the largest log-likelihood ratio. These alignments along with their probabilities can be viewed as a signal and Munteanu and Marcu [2006] next apply a smoothing filter to this signal which sets the probabilities for each alignment based on the average of several values surrounding it. Sequences of 3 or more words with positive log-likelihood scores are then taken to be "phrases" and aligned. The smoothing filter allows for some phrases to be aligned even when a word in the phrase is unknown. They test this method by training a baseline English – Romanian statistical machine translation system and then adding phrase alignments extracted from non-parallel news sources as extra training data. They report an improvement in BLEU scores over the baseline when adding the aligned phrases and furthermore note that this method performs better than the method that extracts parallel sentences [Munteanu and Marcu, 2005] when the comparable corpora are unlikely to contain parallel sentences..

Kumano et al. [2007] take a different approach to phrasal alignments by attempting

to identify the most probable alignments without using a dictionary. The authors extend the work of Marcu and Wong [2002] who use a phrase-based joint probability model to identify phrase alignments between parallel text pairs. Kumano et al. [2007] extend this idea by allowing the model to handle monolingual phrases that do not have a translation in their corresponding document. Reliability of alignments are determined using the log-likelihood ratio and only the alignments with a positive correlation are considered. The authors extracted phrases from 2,000 pairs of Japanese-English news articles and report *0.8* precision of their alignment method. This method relies on the comparable documents being paired initially to their counterparts and the results reflect the fact that all of the documents were only 5 to 8 sentences long. The authors report that using longer documents will decrease the performance of the method because the expansion of the window for phrase correspondences will lead to less reliable co-occurrences.

The approaches we have mentioned in this section have been principally concerned with finding phrasal alignments in comparable corpora for the purpose of improving the translation model in a statistical machine translation system. Snover et al. [2008] attempt to leverage the information available in comparable corpora to improve the language model in a noisy channel model of translation as well as the translation model. This work makes use of a slightly modified version of a probabilistic cross language information retrieval (CLIR) method introduced by Xu et al. [2001] to identify documents in the target language that are most likely to be similar (or comparable) to the source document that is to be translated. This CLIR method makes use of parallel sentence data to estimate the probabilities of word translations. The top returned documents are then used to adapt a general language model to give more weight to the words and phrases that occur in these documents. This is performed by building a bias language model by using only the returned documents and interpolating this model with a general language model. In experiments a weighting of 10% is used for the bias language model. I.e. setting λ to 0.1 in the following

formula.

$$P_{\text{Adapted}}(w_i|w_1...w_{i-1}) = (1-\lambda)P_{\text{General}}(w_i|w_1...w_{i-1}) + \lambda P_{\text{Bias}}(w_i|w_1...w_{i-1}) \quad (1.6)$$

Next the translation model is adapted by selecting phrases that occur in multiple documents selected by the CLIR and generating a new translation from every phrase in the source document to these phrases. All of the new translations are assigned a low uniform probability and then added to a general translation model to bias it. The evaluation of these methods shows only modest improvements in BLEU scores and translation error rates, but the authors note that there may be room for improvement by using variable weighting for the translation rules or possibly utilizing the probabilities from the CLIR system to weight the contribution of each document returned.

Morin and Daille [2010] proposes a method for alignment of phrases from comparable corpora that uses morphological information and syntactic structures to improve accuracy. This method is similar to the single word methods described in Section 1.2.2, but with some additions and modifications. First, term extraction is performed to identify candidate phrases in the source language using frequency of occurrence and part of speech information. Next, each phrase is expanded using morphological variants of words and recoding rules to produce variations of the phrase to translate. For instance one recoding rule is:

$$N_1 Adj \rightarrow N_1 PrepArt?N_2, \text{ where } N_2 \text{ is a nominalized neutral form of Adj} \quad (1.7)$$

This rule could be use for example to transform the phrase *index glycémique* to the phrase *index de la glycémie*. Each variation of the phrase is then translated using a dictionary; where all listed translations of every word are used to generate all possible com-

binations and reordering. These possible phrase translations are then filtered by their frequency in the target language. Their results show that the addition of the morphological and recoding rules significantly improves the recall and precision French – Japanese phrasal alignment.

Related work using comparable corpora has been performed using paraphrasing techniques to improve machine translation quality [Barzilay and McKeown, 2001; Kauchak and Barzilay, 2006; Callison-Burch et al., 2006; Nakov, 2008; Zhao et al., 2008; Marton et al., 2009]. For instance, Marton et al. [2009] focus on using paraphrasing to learning possible translations for out-of-vocabulary words and phrases in the source language and using these to augment existing phrase translation tables. Much like the other work in this section, we can view this as a type of phrasal alignment where by each phrase translation generated is an alignment. This work begins by taking each out-of-vocabulary phrase in the source language and storing all left and right contexts in which this phrase appears (usually one or two words on each side depending on the frequencies of the words). Next these contexts are used to gather all *paraphrase candidates*, words or phrases that appear between one of the left and right contexts in the training data. The distributional profile of each paraphrase candidate is then compared to the distributional profile of the original out-of-vocabulary phrase and the 20 paraphrases that are most similar are kept along with their similarity scores. The distributional profiles used for phrases are vectors containing the log-likelihood of co-occurrence counts with all words that occur within a 6 word window of that phrase, regardless of the position in the window (in contrast to Rapp [1999] where the relative position of words was also stored). The most similar paraphrases are then used to create phrase translation rules from the out-of-vocabulary word or phrase to their transitions. Evaluation is performed by constructing English to Chinese as well as Spanish to English baseline translation systems trained on parallel texts. The paraphrasing method is then used to augment these systems and results show a significant increase

in BLEU and TER over the baseline systems when the size of the initial parallel texts is relatively small (30 thousand words). Increasing the amount of parallel training data used for training reduces the effectiveness of the method, where parallel resources are scarce this method seems a viable way to improve translation quality.

Chapter 2

Discussion of Applicability

A diverse range of alignment techniques and methodologies are presented in the previous chapter. These techniques have been developed for use on various kinds of corpora, using a range of resources, and with a focus on achieving or improving a range of different tasks. The techniques we presented in Section 1.1, for example, have been developed for alignment of parallel corpora and generally cannot be used to produce alignments of comparable corpora, while those presented in Section 1.2 are appropriate for specific types of non-parallel texts. In this chapter we examine how these different alignment strategies can be applied, focusing specifically on their application to corpora of various levels of comparability. We also examine the resources required by theses techniques and other factors that may limit their usefulness or application to under resourced languages.

Techniques for sentence level alignment of parallel texts can be broadly grouped into length based and lexical based techniques. Lexical techniques are generally more tolerant of insertions and deletions of text or otherwise noisy parallel texts than length techniques, but both groups of techniques are, in most cases, not applicable to comparable corpora. Table 2.1 presents a summary of some different sentence level alignment strategies (with an example of work using this strategy). The table includes the resources required and

General Approach	Required Resources of Method	Applicability
Sentence Alignment Techniques		
Length based alignment [Gale and Church, 1993]	none	parallel text with markers
Lexical based alignment (translation model based) [Chen, 1993]	small sentence aligned parallel corpora	parallel text
Lexical based alignment (concurrence based) [Kay and Röscheisen, 1993]	none	parallel text
Cognate/lexical based alignment [Melamed, 1999]	LCS measure /cognate pairs	parallel text
Champollion [Ma, 2006]	bilingual lexicon, parallel corpora, or cognate pairs	parallel text
Lexical/length combination [Moore, 2002]	none	parallel text
Parallel sentence extraction [Munteanu et al., 2004; Fung and Cheung, 2004; Munteanu and Marcu, 2005]	bilingual lexicon or sentence aligned parallel corpora	strongly/weakly comparable texts (publication dates are used in Munteanu et al. [2004])
Parallel sentence extraction (special word based) [Do et al., 2009]	cognate information, publication date	strongly/weakly comparable texts
Parallel sentence extraction (query translation based) [Abdul-Rauf and Schwenk, 2009b]	full MT system trained on parallel corpus	strongly/weakly comparable text

Table 2.1: Applicability of Sentence Alignment Methods

also attempts to specify their applicability to different kinds of corpora. Notice that some lexical techniques make use of additional information such as word cognates, small sentence aligned corpora, or bilingual lexicons. These could be limiting factors for the applicability of techniques as these resources may not exist for a language or the languages may not contain cognates because they use very different writing systems.

Some techniques for sentence alignment that apply to comparable corpora are also

listed in Table 2.1. These techniques are all based on the premise that comparable corpora will contain some sentences that express exactly the same information across languages. The techniques make use of different types of information retrieval strategies to identify these sentences. These methods therefore apply to a specific type of corpora, namely those likely to contain parallel sentences. Collections of parallel corpora are clearly well suited, but it is difficult to assess in general if non-comparable, weakly comparable, or strongly comparable corpora would be appropriate for these methods. The important aspect is if these corpora are likely to contain parallel sentences and if so, for what proportion of the sentences in the corpora.

Word alignment techniques can be groped by those developed for alignment of words in sentence aligned parallel corpora and those developed for alignment of words in non-parallel or monolingual corpora. Table 2.2 gives an overview of some of these techniques and their applicability.

General Approach	**Required Resources of Method**	**Applicability**
Word Alignment Techniques		
IBM lexical based [Brown et al., 1993; Och and Ney, 2003]	none	parallel sentence aligned texts
Discriminative lexical/morphological model [Tufiş et al., 2010]	part-of-speech tagger, lemmatizer, chunker	parallel sentence aligned text
Lexical co-occurrence similarity [Fung, 1995; Rapp, 1999]	bilingual lexicon	monolingual/comparable texts
Lexical frequency rank similarity [Diab and Finch, 2000]	none	monolingual/comparable texts
Combination of lexical co-occurrence, frequency, and similarity [Koehn and Knight, 2002]	bilingual lexicon	monolingual/comparable texts

Table 2.2: Applicability of Word Alignment Methods

Techniques that operate on sentence aligned corpora are generally not applicable to comparable corpora. These techniques can be broadly groped into estimation techniques based on IBM translation models [Brown et al., 1993] and discriminative feature based techniques [Tufiş et al., 2005; Moore et al., 2006]. Feature based techniques can make use of additional language resources such as part-of-speech taggers, phrase chunkers, and bilingual lexicons, but there is nothing inherent in these methods that require their use.

Word alignment techniques for non-parallel corpora generally align words by making use of a bilingual dictionary to compute word co-occurrence information across languages [Fung, 1995; Rapp, 1995; Fung and McKeown, 1997; Rapp, 1999]; however, some techniques make use of other information such as modifiers and predicates [Yu and Tsujii, 2009] or word frequency information without the aid of a dictionary [Diab and Finch, 2000]. All of these techniques apply to monolingual corpora, weakly comparable, strongly comparable, and parallel corpora, although more similar corpora are likely to yield more high probability alignments. Although these techniques are more widely applicable than those requiring parallel sentences, they also produce a simpler type of alignment. The techniques which work over parallel sentences can produce 1 to 1 or 1 to many alignments, while techniques for non-parallel corpora generally produce only 1 to 1 alignments.

Table 2.3 gives a summary of some techniques for phrasal alignment. We use the term phrasal alignment to refer to techniques that can produce many word to many word alignments. Some phrasal alignment techniques expect the input corpora to be parallel and sentence aligned and even make use of these word alignment techniques as an initial step in their algorithms [Koehn et al., 2003; Och and Ney, 2004]. Their applicability is therefore very similar to the methods designed for word alignment of sentence aligned parallel texts.

Techniques for extraction of phrasal alignments from non-comparable corpora commonly make use of a bilingual dictionary, but often this dictionary is generated automatically from a small collection of sentence aligned parallel corpora using a word alignment

General Approach	Required Resources of Method	Applicability
Phrase Alignment Techniques		
Moses Phrase based [Koehn et al., 2007]	none	parallel sentence aligned texts
Bilingual Suffix Trees [Munteanu and Marcu, 2002]	bilingual lexicon or sentence aligned parallel corpora	comparable corpora
Dictionary and co-occurrence translation expansion [Sharoff et al., 2006]	bilingual lexicon	monolingual/comparable corpora
Signal processing based extraction of phrases [Munteanu and Marcu, 2006]	parallel sentence aligned corpora	monolingual/comparable corpora
Alignment using phrase based probability model [Kumano et al., 2007]	none	comparable corpora of short documents, 5 to 8 sentences
Syntax aided lexical alignment of terms [Morin and Daille, 2010]	bilingual lexicon, morphological and recoding rules	monolingual/comparable corpora
Paraphrase based alignment [Marton et al., 2009]	sentence aligned parallel corpora / phrases translations	monolingual/comparable corpora

Table 2.3: Applicability of Phrase Alignment Methods

method such as those implemented in Giza++ [Och and Ney, 2000, 2003]. Generating a dictionary this way has the advantage that it is also possible to make use of the word translation probabilities that these methods provide. The techniques presented in Kumano et al. [2007] do not require the use of a bilingual dictionary, but they do require that all documents are relatively short. These techniques are generally applicable to corpora with any level of comparability including monolingual corpora, but we would expect them to achieve better results on corpora that are likely to contain many phrases expressing the same information across languages

Chapter 3

Comparable Corpora Alignment Case Study

This chapter presents a case study of applying some of the techniques described in the last chapter to corpora for different levels of comparability. The techniques described in the previous chapters are typically applied to either parallel corpora or comparable corpora, but not both, and without evaluation based on the impact the level of comparability has on these results. In this chapter we explore these issues by implementing four general techniques for alignment and measuring their performance on corpora of different levels of comparability.

3.1 Corpora used in the experiments

In our experiments we used the Romanian and English parallel corpora obtained from the JRC-Acquis multilingual parallel corpus[1] and the ACL 2005 collection (training and testing data)[2]. Both corpora are parallel i.e. Romanian sentences are aligned with their

[1] http://langtech.jrc.it/JRC-Acquis.html
[2] http://www.cse.unt.edu/~rada/wpt05/

Table 3.1: Number of sentences in both JRC-Acquis and ACL 2005 corpora.

	Number of sentences before pre-processing	Number of sentences after pre-processing
JRC-Acquis	417,901	286,172
ACL 2005	40,195	35,197

English translations. In addition, the ACL 2005 data is manually word aligned. The total number of sentences in both corpora is shown in Table 3.1. JRC-Acquis corpora contains 417 thousand and the ACL 2005 collection 40 thousand sentences. We pre-processed both corpora to render them to the input format the experimental systems require. In our experiments we use some existing tools such as GIZA++ and Moses. These tools require that the sentences have a limit of 40 words. Thus we removed from both corpora all sentence pairs where at least one of the sentences had more than 40 words. The last column of Table 3.1 shows the remaining sentences in both corpora after this process. In the JRC-Acquis corpora there are 286 thousand and in the ACL 2005 collection 35 thousand remaining sentences. In our experiments we used the sentences from both corpora leading to the total number of 321 thousand sentences.

Giza++ takes as input two text files, one for the source and the other one for the target language. The source text file contains all the sentences from the source language corpora and the target text file all the sentences from the target language corpora. Giza++ assumes that those text files are parallel.

3.1.1 Parallel corpora

When the text files are input to Giza++ it assumes that sentences in those text files are already aligned with each other. This means that a sentence in position X in the source language (English) file has a translation sentence at the same position in the target language (Romanian) file. If a sentence in source language has more than one translation

in the target language then we repeat that source language sentence in the source file so many times it has translations on the target language and vice versa. For instance, if a sentence in the source language has two translation sentences in the target language then we include the source sentence twice to the source file.

We transfer the corpora to this input format. Finally, Giza++ assumes that the sentences in the text files are tokenized. We use the OpenNLP tools[3] to perform the tokenization step.

3.1.2 Strongly comparable corpora

We use the parallel corpora described in Section 3.1.1 to derive a strongly comparable corpora. In our strategy for deriving the strongly comparable corpora we follow the assumption that strongly comparable documents contain the same or similar textual content but at different positions. We keep the source language (English) file as it is and modify only the target language (Romanian) file. In our corpora the maximum alignment between the sentences is two. Thus we swap every sentence in the Romanian text file by another one which is three lines further than the original one. For instance, we swap the first sentence with the third one, the second one with the fourth, the third one with the fifth, etc. To be precise we use the following formula to swap the sentences:

$$newSentecePosition = (positionOfCurrentSentence + 3)\%numberOfSentencesInTheFile \quad (3.1)$$

3.1.3 Weakly comparable corpora

Similar to strongly comparable corpora we derive our weakly comparable corpora. Here we assume that weakly comparable documents don't contain the same or similar textual

[3] `http://opennlp.sourceforge.net/`

content but might have small units which overlap with each other. This can happen if sentences from different documents are aligned with each other. Again we keep the source language (English) file as it is and modify only the target language (Romanian) file. We swap every sentence in the Romanian text file by a sentence that comes from another document. In our corpora the maximum size of a document (measured in number of sentences it contains) is 6022 sentences. Thus to ensure that we don't take sentences from the same document we swap every sentence in the file with another one where we have 6022 gaps between them:

$$newSenteceePosition = (positionOfCurrentSentence + 6022)\%numberOfSentencesInTheFile \quad (3.2)$$

3.2 Experiment 1: Word alignment with Giza++

As described in Section 1.1.2, Giza++ is a statistical word alignment tool for aligning words between two different text files, one written in the source language and the other one in the target language. We applied Giza++[4] on our parallel but also on our strongly and weakly comparable corpora to see how well it performs when there is no correct alignments between the input files. We evaluate the results for each setting (parallel, strongly and weakly comparable corpora) using the manually performed word alignments from the ACL 2005 training and testing data. In this ACL 2005 data there are in total 402,186 word alignment pairs. Some of these pairs are repeated several times. To see how many unique pairs we have we counted each pair only once and recorded a unique number of 72,308 word pairs (see 3.1). In the evaluation we take an aligned word pair from the ACL 2005 data, check if it occurs in the word alignment list produced by Giza++ and count the number of matches. The results are shown in Figure 3.2 to 3.3.

[4]We used all default settings for the parameters Giza++ uses.

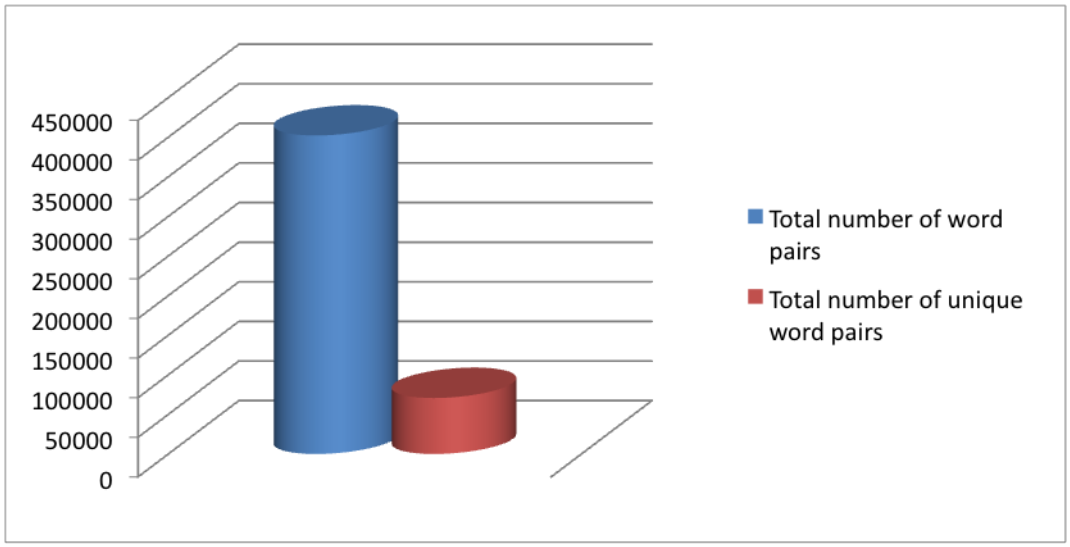

Figure 3.1: Word alignment test data statistics.

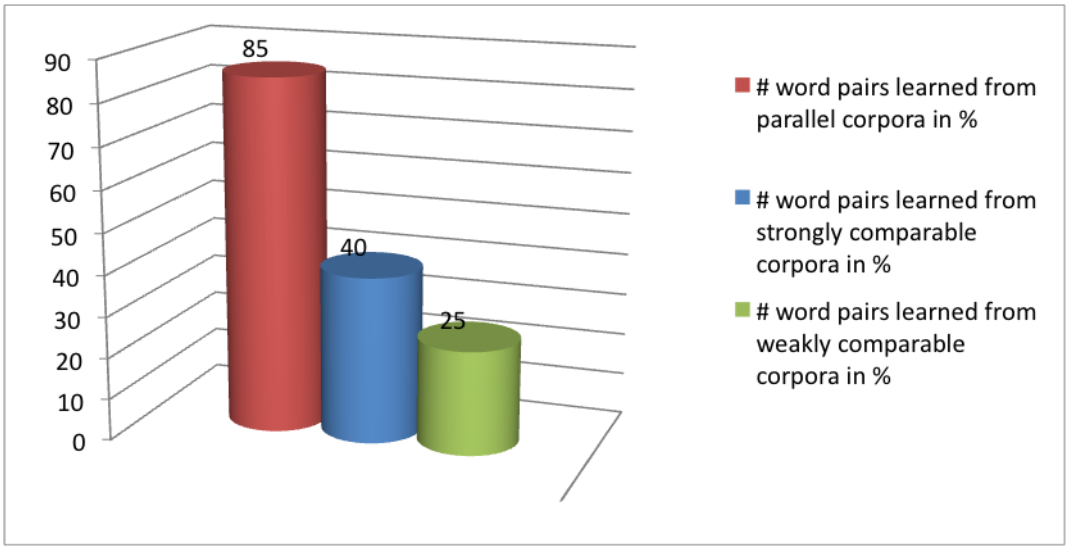

Figure 3.2: Results for alignment of all word tokens in test data of different levels of comparability.

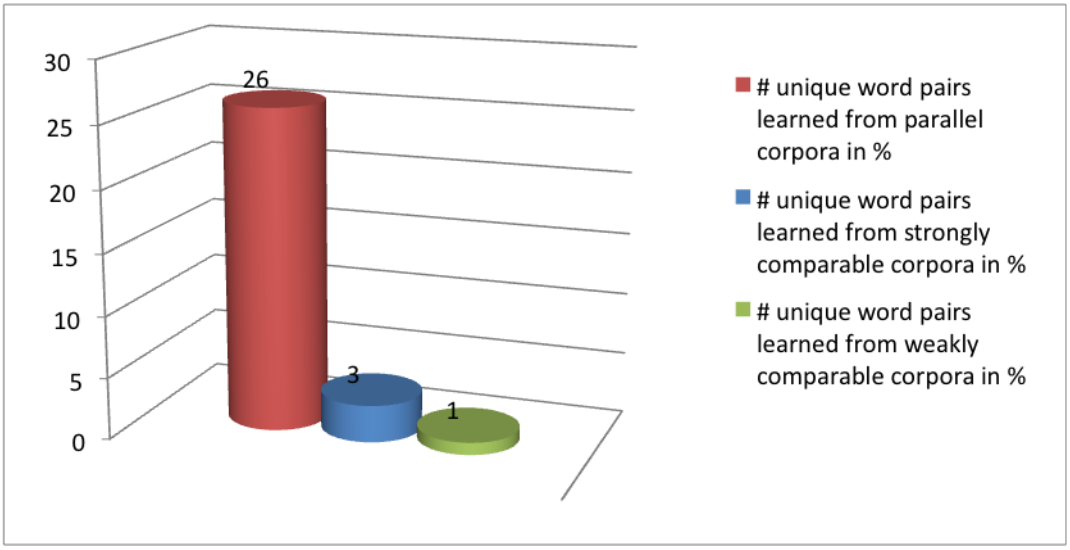

Figure 3.3: Results for alignment of *unique* word types in test data of different levels of comparability.

From the figures we can see that when parallel corpora is used for obtaining word alignments Giza++ covers 85% of the duplicated manual aligned word pairs and 26% of the unique ones. The number drops down when strongly and weakly comparable corpora are used. In case of strongly comparable corpora the coverage of duplicated word pairs goes down to 40% and 3% for the unique word pairs. For weakly comparable corpora the coverage numbers get even smaller. Only 25% for duplicated word pairs and 1% for unique ones. These numbers indicate that although the current version of Giza++ performs very well on parallel data it is less appropriate for comparable data, but still correctly aligns many common words.

3.3 Experiment 2: Phrase alignment with Moses

We described Moses in Section 1.1.2. It performs phrase alignment between two files written in source and target language. It makes use Giza++ word alignment outputs and performs on them statistical phrase alignments. Similar to our first experiment we run

Moses[5] on the parallel, strongly and weakly comparable corpora. We evaluate the results for each setting (parallel, strongly and weakly comparable corpora) based on the phrase alignment we obtain from the ACL 2005 training and testing data.

We generate phrases for each sentence pair in the human word aligned ACL 2005 data to use for out test data in this experiment. To generate these phrases we process each sentence pair and look for the longest sequence of words aligned across sentences, but not necessarily in the same order. The stipulations are that phrases are made up of consecutive words and that all words in both phrases must align to words in the other phrase (not necessarily unique words). This allows for phrases to be of different lengths and contain different word orderings and eliminates phrases that cannot contain gaps, untranslated words, or any words that translate to a word that is not a contiguous part of the paired phrase. We always search for the longest possible phrases in each sentence. By following this we obtain 87456 phrase pairs in total made up of 71957 unique phrases (see 3.4). Similar to the experiment with Giza++ in the evaluation we report the number of phrases (produced from the ACL data) included in the phrase table created using Moses. The results are shown in Figure 3.5 to 3.6.

Figures 3.5 to 3.6 show that parallel data again leads to the highest coverage. From Figure 3.5 we can see that 18% of the phrases from the ACL 2005 data are found in the phrase table generated by Moses using the parallel corpora. In case of the unique phrase pairs (see Figure 3.6) 9% of the ACL 2005 phrases are found in the Moses trained phrase table. The results for the comparable corpora (both strongly and weakly) have only 2% or less coverage indicating that Moses is likely not appropriate for use directly on comparable data.

[5]We run Moses using the script file distributed with the Moses installation package.

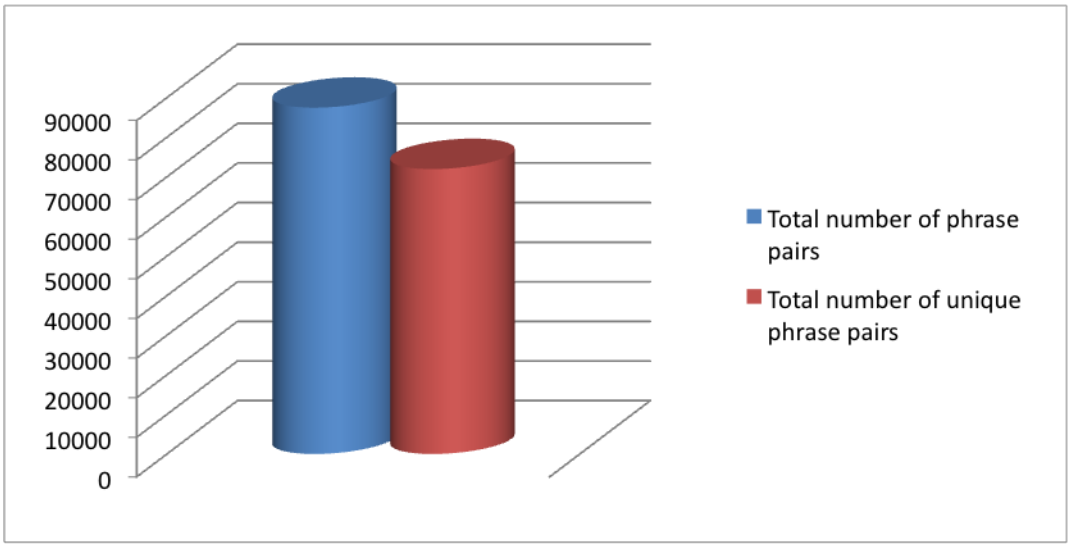

Figure 3.4: Phrase alignment test data statistics.

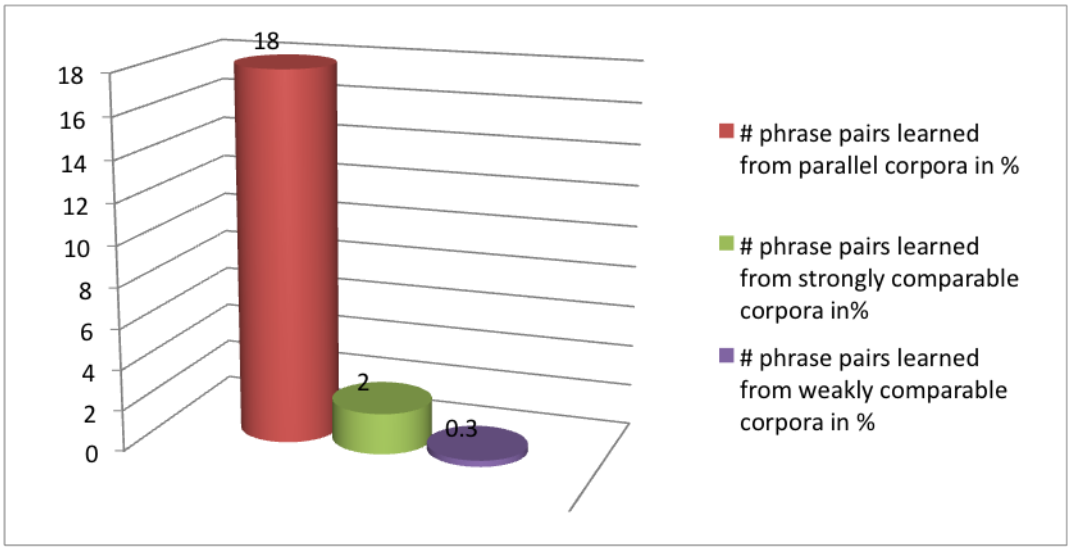

Figure 3.5: Results for alignment of *all* phrases in test data of different levels of comparability.

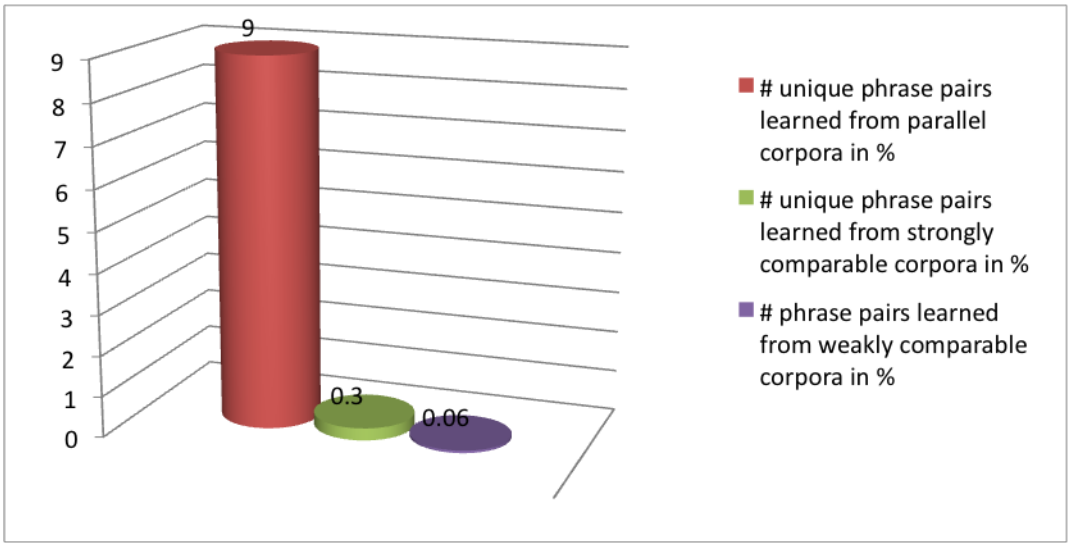

Figure 3.6: Results for alignment of *unique* phrases in test data of different levels of comparability.

3.4 Experiment 3: Cognate Based Alignment

In this experiment we test several methods for alignment of words based on identification of word cognates across languages (see Section 1.1.1). Our procedure works by searching for words that are spelled similarly across languages. The most similar target word (according to the metrics we tested) for each source word is taken to be its alignment. We then test our procedure using the same experimental setup as in the previous two sections, namely by testing what percentage of the gold standard human word aligned ACL 2005 data are correctly identified using this method. These methods do not make use of any alignment information therefore impose very few restrictions of the data used, for instance, the data can simply be unrelated monolingual corpora in both languages.

We experiment using 4 different methods of identifying cognate pairs: *longest common subsequence*, *longest common string*, *edit distance*, and *exact match*. The *longest common subsequence* (LCS) measure is described in Section 1.1.1, it measures the longest common non-consecutive sequence of characters between two strings. For instance, the words "*dol-*

lars" and "*dolari*" share a sequence of 5 non-consecutive characters in the same ordering. We implement the LCS measure using a dynamic programming approach Cormen et al. [2001], so that its computation is efficient and can be applied to a large number of possible word pairs. The LCS method is used to identify the most likely alignment for every source word based by choosing the word in target word that has the longest (normalized) subsequence in common. We normalize by the length of the longest word (see LCSR in Section 1.1.1). If more than one word gives the same score then ties are broken by choosing the word which occurs more frequently in the target data.

The *longest common substring* (LCST) measure is computed similar to the LCS measure, but measures the longest common *consecutive* string of characters that two strings have in common. This measure can be thought of as finding a word which contains the longest n-gram of characters in common with a given word. The formula for the LCST is score we use is a ratio as in the previous measure:

$$LCSTR(X,Y) = \frac{\text{length}[LCST(X,Y)]}{\max[\text{length}(X), \text{length}(Y)]} \tag{3.3}$$

The *edit distance* measure (also referred to as Levenshtein distance) computes the minimum number of operations necessary to transform one string into the other. The allowable operations are insertion, deletion, and substitution. We use a bottom up dynamic programming approach to compute the edit distance in an approach similar to that used for computation of LCS.

We also make use of exact string matching measure. This measure only considers words as cognates if they are written exactly the same in both languages. This occurs often for numbers, punctuation, and many proper nouns. This measure is included because it is extremely fast to compute and gives an indication of what percentage of alignment pairs are likely to be written exactly the same in both languages. These initial experiments

over human word aligned English/Romanian data, so we would expect exact cognate based alignment methods to perform better for these two languages than for languages with more different writing systems (for example English/Greek), but nonetheless most language pairs will likely share some punctuation, symbols, numbers or other strings in common.

In all measures we attempt to match the longest string possible first using the characters as written and then by stripping diacritics from all characters and attempting to match again. A word pairs score is the maximum of it score with and without diacritical information. For instance for the words "*coalition*" and "*coaliţie*" share an extra '*t*' in common if we strip diacritics.

The results for these experiments are given in Figure 3.7 and Figure 3.8. These figures show the results of using the different methods of identifying cognates. The results are applicable to corpora to all levels of comparability (including weakly and non-comparable) because these methods do not use any word alignment information. Therefore it is reasonable to compare these numbers with the Giza++ alignment figures from Section 3.2 on comparing weakly comparable corpora. We can see that simply taking as alignments all pairs that match exactly we correctly identify almost 18% of all alignments in the test data

3.5 Experiment 4: Co-occurrence Based Alignment

In this section we present experiments on the alignment of words in corpora using of word co-occurrence information. We use a method of creating a bilingual dictionary of alignments very similar to the methods used in Rapp [1995, 1999]; Koehn and Knight [2002] and described in Section 1.2.2. This method does not make use of any alignment information between the texts, such as which sentences or documents are parallel, and therefore applies to corpora of different levels of comparability. This method requires

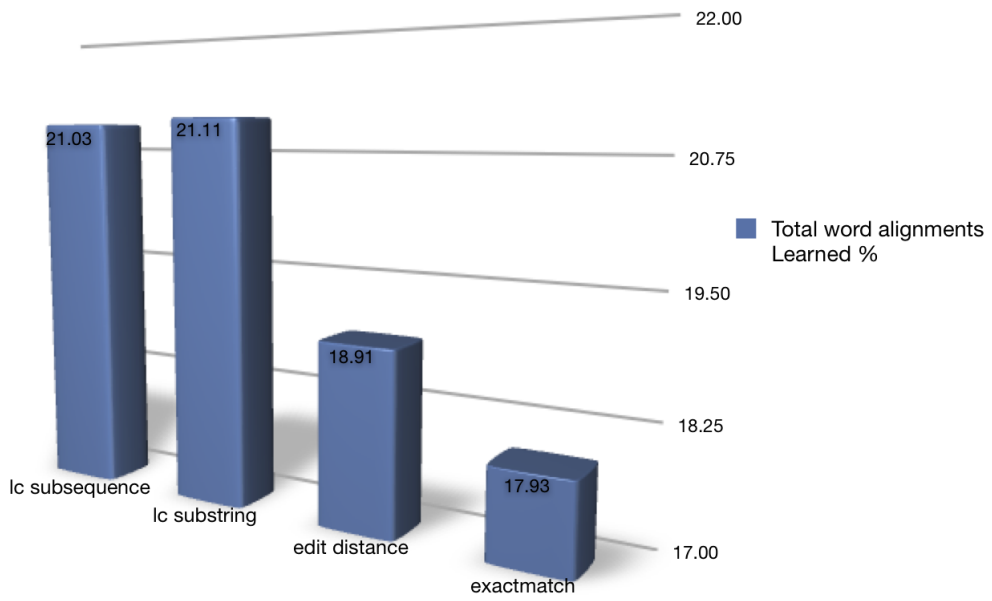

Figure 3.7: Results for alignment of *all* words in test data using different strategies for the alignment of cognates.

a small initial seed dictionary to use to translate some words form the source to the target language. We experiment with two automatic methods for producing this initial dictionary from the held-out JRC-Acquis training data. The first is to identify all exact match cognates that occur in the training, as described in the previous section, and use these word pairs as the dictionary. The second method we use is to run Giza++ on the sentenced aligned JRC-Acquis training data to generate possible translations of words and then produce a bilingual dictionary by choosing the most likely translation for every word. Both of these methods for generating dictionaries do not involve any of the data used for testing alignments and therefore to not compromise our goal that these experiments should be representative of results on corpora of all levels of comparability, including weakly and non-comparable corpora.

Our method proceeds as in Rapp [1999] by constructing word co-occurrences informa-

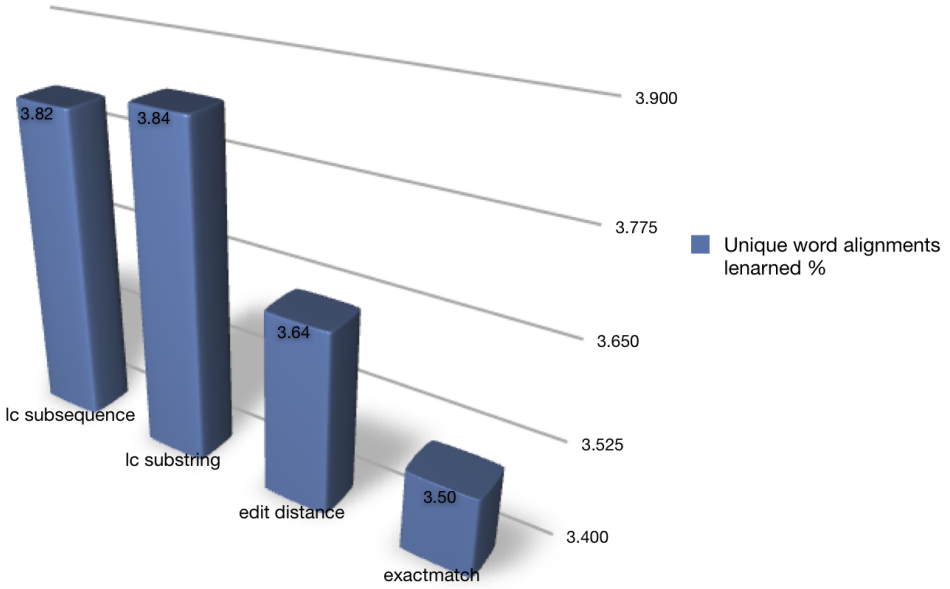

Figure 3.8: Results for alignment of *unique* words in test data using different strategies for the alignment of cognates.

tion independently for both the source text and the target text. For each word in the source texts we construct a vector of its co-occurrence near all words which have an entry in the bilingual dictionary. We keep track of the position this word occurred relative to the words in the dictionary as in Rapp [1999]; Koehn and Knight [2002] by storing a separate vector for words which appeared in 4 positions around the word, namely the positions -2, -1, 1, 2 relative to the source word. These 4 vectors are all of length equal to the number of words in the dictionary. The values in the array are taken to be the log-likelihood of these words occurring together in the corpus and then the vectors are each normalized so their elements sum to one. Vectors for each word of the target text are constructed in a similar way but with the values being log-likelihood with terms which are translations in the dictionary. For each word in the source text we search for its most likely alignment, by translating all words in its vector, ordering it to be the same as the target text vectors, and then computing similarity with each target language word vector. The word vector with the highest similarity score is taken to be its alignment. In these experiments we stem all Romanian[6] words in the dictionary and stem the Romanian words in the text only for mapping to the dictionary, but achieve greater dictionary coverage in this manner. In addition we remove all stop words and punctuation from the texts in both languages before we begin processing the text.

The results for these experiments are shown in Figure 3.9 and Figure 3.10. The percentages of test data alignments covered are promising and show a clear improvement over simple cognate based alignment and do not make use of any sentence alignment information or make any assumptions about the comparability level of the corpora. These methods can be applied to any corpora, even unrelated monolingual corpora.

[6]We make use of the Snowball Romanian Stemmer available at `http://snowball.tartarus.org`.

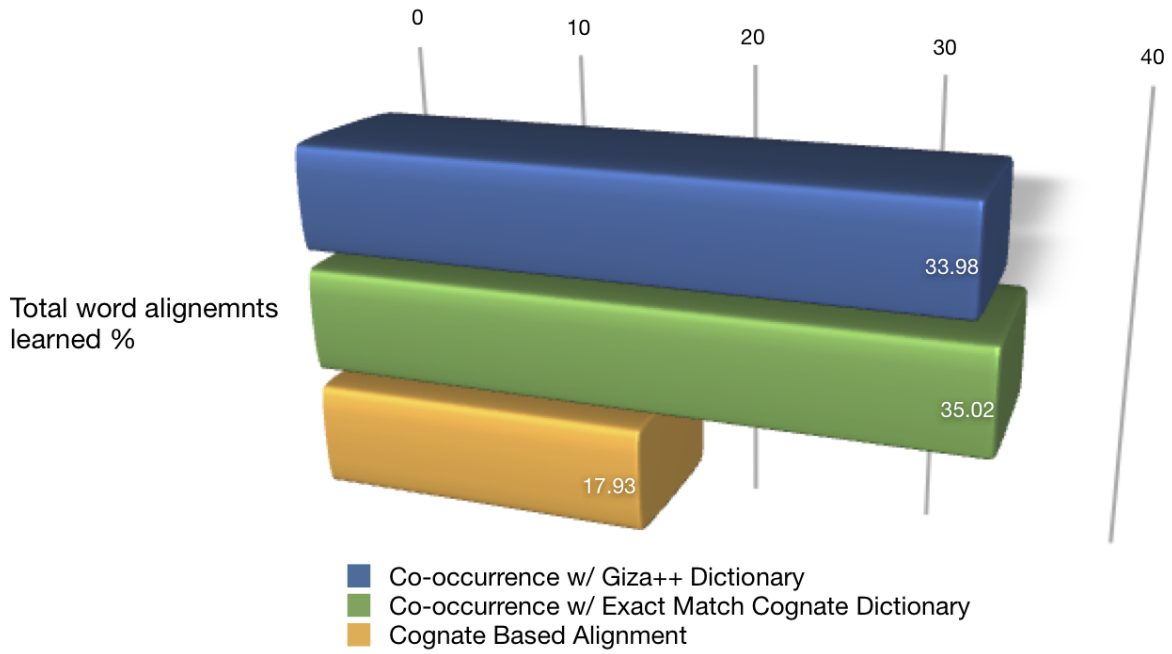

Figure 3.9: Results for alignment of *all* words in test data using word co-occurrence information.

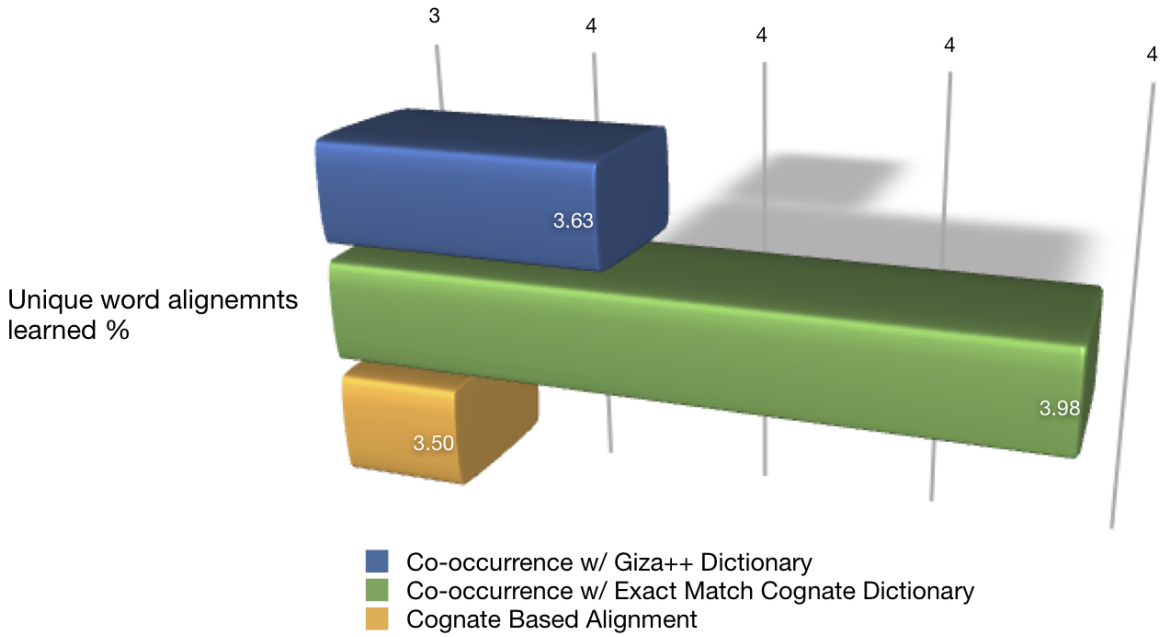

Figure 3.10: Results for alignment of *unique* words in test data using word co-occurrence information.

3.6 Conclusion

In this report we began with a review of existing alignments strategies designed for parallel corpora, comparable corpora, and non-comparable corpora. We focused particularly on the appropriateness of these techniques to corpora of different levels of comparability and established guidelines for their applicability and the resources required by the techniques. A case study was presented making use of four different alignment methods and applying them to corpora of different levels of comparability. We tested the accuracy of these methods for alignment of words or phrases by comparing the alignments produced to human word alignments. We showed that the most widely used existing alignment methods (Giza++ and Moses) are not well suited for use directly on strongly or weakly comparable texts, but for parallel corpora it is possible to 85% of the correct alignments using this method. Additionally, we showed that for weakly comparable corpora it is possible to correctly identify only around 35% of the alignments in text using word co-occurrence information. The results indicate that there is much room for improvement on alignment accuracy of strongly and weakly comparable texts and increasing this accuracy will major point of focus for this project.

Bibliography

Abdul-Rauf, S. and Schwenk, H. (2009a). Exploiting comparable corpora with TER and TERp. In *BUCC '09: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora*, pages 46–54, Morristown, NJ, USA. Association for Computational Linguistics.

Abdul-Rauf, S. and Schwenk, H. (2009b). On the use of comparable corpora to improve SMT performance. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23, Morristown, NJ, USA. Association for Computational Linguistics.

Barzilay, R. and Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 25–32.

Barzilay, R. and McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 50–57, Morristown, NJ, USA. Association for Computational Linguistics.

Bergroth, L., Hakonen, H., and Raita, T. (2000). A survey of longest common subsequence algorithms. In *String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium on*, pages 39 –48.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176, Morristown, NJ, USA. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24, Morristown, NJ, USA. Association for Computational Linguistics.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 9–16, Morristown, NJ, USA. Association for Computational Linguistics.

Chiao, Y.-C. and Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–5, Morristown, NJ, USA. Association for Computational Linguistics.

Church, K. W. (1993). Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Columbus, Ohio, USA. Association for Computational Linguistics.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. The MIT Press, 2nd revised edition edition.

Dagan, I., Church, K. W., and Gale, W. A. (1993). Robust bilingual word alignment for machine aided translation. In *In Proceedings of the Workshop on Very Large Corpora*, pages 1–8.

Diab, M. and Finch, S. (2000). A statistical word-level translation model for comparable corpora. In *In Proceedings of the Conference on Content-Based Multimedia Information Access*.

Do, T., Le, V., Bigi, B., Besacier, L., and Castelli, E. (2009). Mining a comparable text corpus for a Vietnamese-French statistical machine translation system. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 165–172. Association for Computational Linguistics.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 173–183.

Fung, P. (2000). A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. *Machine Translation and the Information Soup*, pages 1–17.

Fung, P. and Cheung, P. (2004). Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of EMNLP 2004*, Barcelona, Spain.

Fung, P. and McKeown, K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Fung, P. and Yee, L. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Annual Meeting-Association for Computational Linguistics*, volume 36, pages 414–420. Association for Computational Linguistics.

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 177–184, Morristown, NJ, USA. Association for Computational Linguistics.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102.

Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., and Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 526, Morristown, NJ, USA. Association for Computational Linguistics.

Graff, D. (2003). English Gigaword. Linguistic Data Consortium, catalog number LDC2003T05.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio. Association for Computational Linguistics.

Haruno, M. and Yamazaki, T. (1997). High-performance bilingual text alignment using statistical and dictionary information. *Nat. Lang. Eng.*, 3(1):1–14.

Kauchak, D. and Barzilay, R. (2006). Paraphrasing for automatic evaluation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 455–462, Morristown, NJ, USA. Association for Computational Linguistics.

Kay, M. and Röscheisen, M. (1993). Text-translation alignment. *Comput. Linguist.*, 19(1):121–142.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Morristown, NJ, USA. Association for Computational Linguistics.

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, pages 9–16, Morristown, NJ, USA. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Kumano, T., Tanaka, H., and Tokunaga, T. (2007). Extracting phrasal alignments from comparable corpora by using joint probability SMT model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 95–103.

Lacoste-Julien, S., Taskar, B., Klein, D., and Jordan, M. I. (2006). Word alignment via quadratic assignment. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 112–119, Morristown, NJ, USA. Association for Computational Linguistics.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *In Proceedings of LREC-2006*.

Marcu, D. and Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, volume 10, pages 133–139.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390. Association for Computational Linguistics.

Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. *Comput. Linguist.*, 25(1):107–130.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 135–144, London, UK. Springer-Verlag.

Moore, R. C. (2005). A discriminative framework for bilingual word alignment. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 81–88, Morristown, NJ, USA. Association for Computational Linguistics.

Moore, R. C., Yih, W.-t., and Bode, A. (2006). Improved discriminative bilingual word alignment. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 513–520, Morristown, NJ, USA. Association for Computational Linguistics.

Morin, E. and Daille, B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44(1–2):79–95.

Munteanu, D., Fraser, A., and Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2004)*.

Munteanu, D. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Munteanu, D. S. and Marcu, D. (2002). Processing comparable corpora with bilingual suffix trees. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 289–295, Morristown, NJ, USA. Association for Computational Linguistics.

Munteanu, D. S. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 81–88, Morristown, NJ, USA. Association for Computational Linguistics.

Nakov, P. (2008). Paraphrasing verbs for noun compound interpretation. In *Proc. of the Workshop on Multiword Expressions, LREC-2008*.

Och, F. J. and Ney, H. (2000). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics*, pages 1086–1090, Morristown, NJ, USA. Association for Computational Linguistics.

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

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30(4):417–449.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322. Association for Computational Linguistics.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Sharoff, S., Babych, B., and Hartley, A. (2006). Using comparable corpora to solve problems difficult for human translators. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 739–746, Morristown, NJ, USA. Association for Computational Linguistics.

Snover, M., Dorr, B., and Schwartz, R. (2008). Language and translation model adaptation using comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 857–866. Association for Computational Linguistics.

Tillmann, C. (2009). A beam-search extraction algorithm for comparable data. In *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 225–228, Morristown, NJ, USA. Association for Computational Linguistics.

Tillmann, C. and Xu, J.-m. (2009). A simple sentence-level extraction algorithm for comparable data. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 93–96, Morristown, NJ, USA. Association for Computational Linguistics.

Tufiş, D., Ion, R., Ceauşu, A., and Ştefănescu, D. (2005). Combined aligners. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 107–110, Ann Arbor, Michigan. Association for Computational Linguistics.

Tufiş, D., Ion, R., Ceauşu, A., and Ştefănescu, D. (2010). Reifying the alignments. In *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian.* Romanian Academy Publishing House, Bucharest.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning japanese-english news articles and sentences. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 72–79, Morristown, NJ, USA. Association for Computational Linguistics.

Xu, J., Weischedel, R., and Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 105–110, New York, NY, USA. ACM.

Yang, C. C. and Li, K. W. (2003). Automatic construction of english/chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54(8):730–742.

Yu, K. and Tsujii, J. (2009). Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 121–124, Morristown, NJ, USA. Association for Computational Linguistics.

Zens, R., Matusov, E., and Ney, H. (2004). Improved word alignment using a symmetric lexicon model. In *COLING '04: Proceedings of the 20th international conference on*

Computational Linguistics, page 36, Morristown, NJ, USA. Association for Computational Linguistics.

Zhao, B. and Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. In *IEEE International Conference on Data Mining (ICDM 2002)*.

Zhao, S., Niu, C., Zhou, M., Liu, T., and Li, S. (2008). Combining multiple resources to improve SMT-based paraphrasing model. In *Proceedings of ACL-08: HLT*, pages 1021–1029, Columbus, Ohio. Association for Computational Linguistics.