

LEARNING A TRANSLATION LEXICON FROM NON PARALLEL CORPORA

MA Project Report (CLMA – 2008)

Ankit K Srivastava

INDEX	PAGE
Section 1: Introduction	2
Section 2: Literature Survey	6
Section 3: Methodology & Implementation...	12
Section 4: Experiments	23
Section 5: Discussion	35
Section 6: Conclusion	37
Section 7: References	38

Abstract. This project evaluates the performance of syntactic context windows against positional context windows in extracting word translations from non-parallel English and German newswire corpora, supplied with a list of seed words i.e. known translations.

1 Introduction

A translation lexicon is a mapping between two disjoint sets of symbols. Given some corpus sample over each set of symbols, one might induce the mapping by performing statistical analyses on the corpora to find correlations between the symbols (Hwa et al., 2006). The sets stand for the source and target languages, while the symbols most often refer to words. Simplistically, a translation lexicon can be thought of as a dictionary which contains a word in the source language cross-listed with the corresponding word in the target language.

A translation lexicon is an important component of multilingual processing applications such as machine translation systems and multilingual information retrieval systems. Such components are also used in cross-lingual resource building. For example, (Yarowsky and Ngai, 2001) have shown that, for training a French part-of-speech tagger, one could “acquire” an automatically annotated French training corpus by projecting it from the output of an English part-of-speech tagger via lexical translations (i.e. using a English-French translation lexicon) between English and French.

Large collections of naturally-occurring text in machine readable form, better known as, corpora play an essential role in the process of constructing these word translations between languages. It was not long ago that researchers referred to the Brown Corpus (about the same size as a week of a newswire service or the complete works of William Shakespeare) as a “large” corpus. Today, one can easily surf the web and download millions of words in multiple languages in no

time at all (Armstrong et al., 1999). Such an extensive availability of huge amounts of text is beneficial to data-driven natural language processing applications including machine translation, parsing, and information retrieval. Researchers in large corpora are overcoming the so-called knowledge-acquisition bottleneck by processing these vast quantities of data and extracting useful information like word frequencies, word associations, typical predicate-argument relations, and lexical mappings (bilingual as well as multilingual).

Multilingual corpora can be classified into parallel and non-parallel. Statistical approaches to Machine Translation have achieved impressive performance by leveraging large amounts of parallel corpora. However, such data are available only for a few dozen language pairs in limited domains. On the other hand, comparable corpora (texts written in different languages that, roughly speaking, “talk about the same thing”) are by comparison plentiful for many languages. Most existing multilingual corpora are not parallel, but comparable. This fact is reflected in major evaluation conferences on cross language information retrieval, which only use comparable corpora for their multilingual tracks.

Compilation of translation lexicons is a crucial process for machine translation (MT) (Brown et al., 1990). The traditional method to automatically acquire a translation lexicon is through word alignment of large volumes of parallel corpora. The problem of extracting lexical mappings (word translations) from parallel corpora is well-studied; there exist numerous supervised and unsupervised word token alignment methods yielding highly accurate lexical mappings (Melamed, 2000; Och and Ney, 2003; Callison-Burch et al., 2004).

However, when the language samples are from non-parallel corpora, the problem is more challenging. Table 1.1 distinguishes the characteristics of parallel and non-parallel texts significant in lexicon extraction (Fung, 2000).

Table 1.1 Characteristics of parallel and non-parallel corpora

PARALLEL CORPORA	NON-PARALLEL CORPORA
Words have one sense per corpus	Words have multiple senses per corpus
Words have single translation per corpus	Words have multiple translations per corpus
No missing translations in the target document	Translations might not exist in the target document
Frequencies of bilingual word occurrences are comparable	Frequencies of occurrences not comparable
Positions of bilingual word occurrences are comparable	Positions of occurrence not comparable

Nonparallelness in corpora can be measured in terms of author, domain, topics, time period, and language. The quality of the mapping between word types in lexicon extraction depends directly on the degree of relatedness between corpora. The higher the degree of nonparallelness, the more challenging is the extraction of bilingual information. Parallel corpora represent one extreme of the spectrum wherein the only difference is along the dimension of language. At the other extreme, there are newspapers from different time periods, written by different authors, sometimes covering different domains. This paper studies extraction methodologies using this second type of nonparallel corpora. Parallel sentence pairs are no longer available.

Current methods exploit such non-parallel corpora in order to extract translations of new words (words with unknown translations) in the context of words with known translations. The theoretical framework behind the context approach is that words (from two different corpora) that are translations of each other should have similar co-occurrence patterns with other words in their respective corpora.

Word contexts are typically either syntactic or positional. This project explores and evaluates a method to extract a word-level translation lexicon from German and English non-parallel corpora by using dependency (syntactic) relations between unknown and known words. Starting with parsed English and German newswire text (phrase structure parse trees and dependency trees of Wall Street Journal '90-'92 and Deutsche Presse Agentur '95-'96), experiments are carried out to compare the quality of the lexicon acquired from syntactic contexts against that acquired from positional contexts.

Leveraging information learned from non parallel corpora has the potential to improve MT quality for the large number of languages for which we have sizeable monolingual corpora but no large parallel corpora.

2 Literature Survey

There have been several attempts to develop methods for extracting useful information like word translations directly from non-parallel corpora. Note, “unknown word” refers to any words whose translation is not known. Most of the proposed translation learning algorithms follow a 3-step process (Figure 2.1):

1. For each unknown word in the source and target languages, define the context in which that word occurs. The context will typically consist of a set of words.
2. Using an initial seed lexicon or a general bilingual dictionary, translate as many source context words into the target language.
3. Use a similarity metric to compute the translation of each unknown source word. It will be the target word with the most similar context.

Figure 2.1: The 3-step algorithm for lexicon extraction

Different approaches vary mainly in how they define the context of each word, how they acquire an initial lexicon, and how they compare the similarity of the contexts.

(Koehn and Knight, 2002) present work on the task of constructing a word-level translation lexicon purely from unrelated monolingual corpora. They combine several linguistic clues such as cognates, word frequency, similar context, and preservation of word similarity to find translations of words (only nouns) from English (Wall Street Journal) and German (German News Wire) monolingual corpora in a comparable domain. My approach uses the same corpora as described in their paper. They used the list of identically spelled words in the source and target languages as a seed lexicon. The paper concludes that for

efforts in building machine translation systems, some small parallel text should be available. From these, some high-quality lexical entries can be learned, but there will always be many words that are missing. These may be learned using the linguistic clues mentioned above. My methodology expands on the context clue described in this paper. While Koehn and Knight use a 4-word window (2 preceding and 2 succeeding) to define a context, I will be using syntactic (dependency) relations to define an unknown word context. Koehn and Knight's system will serve as the positional context window approach to be compared against.

(Gaussier et al., 2004) present a geometric view on bilingual lexicon extraction from comparable corpora. The paper reinterprets the standard approaches (the 3-step process described above) and formulates three new ones (Extended Method, Multilingual PLSA, and Canonical Correlation Analysis) inspired by latent semantic analysis (LSA) developed within the information retrieval community to treat synonymous and polysemous terms. Clustering is used in multilingual probabilistic LSA so that translation pairs with synonymous words appear in the same cluster, while translation pairs with polysemous words appear in different clusters. Experiments were conducted on an English-French corpus. A bilingual dictionary was used in addition to linguistic preprocessing (tokenization, lemmatization, POS-tagging) on both the source and target language corpora. Infrequent words occurring less than 5 times were discarded when building the indexing terms and the dictionary entries. The context vectors were defined by considering terms occurring in the same context window of size 4 (i.e. a neighborhood of ± 2 words around the current word). The similarity metric used is Dice or Jaccard coefficients in addition to the cosine similarity. The paper mentions the widely-believed assumption if two words are mutual translations, then their more frequent collocates (taken here in a very broad sense)

is likely to be mutual translations as well. My approach to augmenting a translation lexicon is founded on this assumption.

(Diab and Finch, 2000) describe an alternate approach to creating an initial seed lexicon in contrast to that implemented by Koehn and Knight, 2002. It tackles the problem from scratch by searching for a translation mapping which optimally preserves the intralingual association measure between words. This is based on the assumption that pairs of words highly associated in one language should have translations that are highly associated in the other language. In other words, if two terms or words have close distributional profiles in one language, then their corresponding translations' (i.e. in the second language) distributional profiles should be close in a comparable corpus. It uses Spearman rank order correlation between the context vectors restricted to highly frequent tokens as an association measure. The statistics used in this approach is taken from within the respective source and target language corpora and not across them. A fixed sliding window of 2 tokens is used and the problem domain is compared to a substitution cipher. The approach is viewed as subjecting one of the corpora to a word-substitution cipher, and attempting to discover that cipher by using statistics of the distribution of texts in each corpus separately. No morphological or lexical analysis is applied to either corpus during this investigation. This approach depends neither on a bilingual dictionary nor on a seed lexicon. This paper is of importance for its description of the different types of corpora used in translation models; parallel, unrelated nonparallel, and comparable nonparallel corpora. It asserts the independence of corpora sizes of the comparable texts used to construct translation lexicon.

(Rapp, 1995) and (Rapp, 1999) are based on the assumption that there is a correlation between the patterns of word co-occurrences in corpora of different

languages. (Rapp, 1995) suggests that identification of word translations should be possible with non-parallel comparable as well as non-parallel unrelated corpora. (Rapp, 1999) asserts the fact that most statistical clues useful in the processing of parallel texts cannot be applied to non-parallel texts. (Rapp, 1995) proposes an approach very similar to the model presented in (Diab and Finch, 2000). Rapp builds his model based on the assumption that if two words strongly co-occur where strength is defined in terms of frequency – then their translations, in comparable and unrelated corpora, will also co-occur with a high frequency. He proposes a model for German-English non-parallel corpora (comprising both comparable and unrelated corpora) which assumes a fixed window size of 11 terms. He uses the city block metric to measure the distance between vectors, or entries in the contingency table. In (Rapp, 1999), the assumption remains the same as in the earlier work by the author, yet he varied the window size for the words to be $4n$ (i.e. count a context word separately for each the four (two left and two right) possible positions around an unknown word), and he introduced the usage of linguistic tools to the model such as lemmatization, morphological analysis, a bilingual lexicon and seed words. He also eliminated function words from his investigation.

(Fung and Yee, 1998) propose an approach based on the vector space model for translating new words in Chinese English comparable corpora. The motivation behind the work is to make use of the easier access to nonparallel resources and arrive at accurate translations for newly encountered words. The basic intuition of their work is that a content word is closely associated with words in its context. They form a vector for a word in terms of its context words, where the vector dimensions are defined by the frequency of occurrence of the context word with the content word in the same sentence, within a corpus. In the similarity measures described in the paper, the magnitude of the data items (term frequencies) is

contributing directly to the similarity measure. The frequencies are normalized using the commonly known IR method of Term Frequency (TF) and Inverse Document Frequency (IDF). Token frequencies do not contribute directly to the distance measure, rather their ranks with respect to one another, hence, the non-parametric measure of rank correlation. The approach that Fung and Yee propose seems to depend essentially on word pairs from a machine translation system, where these word pairs act as “bridges” between the terms, as well as seeds to bootstrap the word to word translation system. They claim that the association between words and seed words that occur in their context is preserved in comparable corpora. A distinguishing characteristic of this approach is that they pay special significance to content words.

(Peters and Picchi, 1997) propose a method for word-level translation for comparable corpora in Italian and English. It relies heavily on the availability of linguistic resources such as bilingual dictionaries and morphological analyzers. They report success for their approach, which is measured in a preliminary investigation for cross-language retrieval. The paradigm is slightly different since the model assumes interaction with a user to supply the seed words. It is considered a semi-automatic approach.

(Fung and Cheung, 2004) present a machine learning approach capable of extracting parallel sentences from non-parallel corpora, which the authors claim to be more disparate from the previously considered comparable corpora. They exploit bootstrapping on top of the IBM Model 4. They use EM for lexicon learning. While this paper mainly deals with extracting parallel sentences for a MT system from non-parallel corpora, it is noteworthy because they use lexical information like word overlap and cosine similarity to match documents before selecting them to extract sentences for MT. Although their approach to word

translation extraction is neither unique nor novel, it proves that the existing methods can be successfully deployed in a working MT system.

(Otero and Campos, 2005) relies on the extraction of bilingual pairs of lexico-syntactic templates from parallel corpora. These templates are then used to extract word translations from non-parallel corpora. The authors seek to exploit the pros of both parallel (high accuracy) and non-parallel (high coverage) corpora. It is a 2-step strategy. First, a representative set of bilingual correspondences between unambiguous lexico-syntactic templates is extracted from small parallel texts. Next, these template pairs are used as local contexts to extract word translations from comparable, non-parallel corpora. A word w_1 in the source language can be the translation of a word w_2 in the target language if w_1 occurs in local contexts that are translations of the local contexts w_2 occurs in. While this method is the closest to the approach described in this paper in that it uses syntactic contexts, their approach is slightly different in the technique they use to define the lexico-syntactic templates and the fact the parallel corpora in small amounts must exist for the same language pairs for which comparable corpora is used.

3 Methodology and Implementation

According to Harris' hypothesis (Harris, 1985), counting co-occurrences within a window of size N is less precise than counting co-occurrences within local syntactic contexts. Syntactic contexts are less ambiguous and more sense-discriminative than contexts denoted as windows of size N . The main purpose of this project is to examine and evaluate the performance of syntactic context windows against positional context windows in lexicon extraction.

My approach to learning word translations from non-parallel corpora involves using syntactic dependencies to define the context of an unknown word. In addition to Harris' hypothesis stated above I consider the fact that dependency structures rely on words and their dependents as beneficial to translation of languages with different word orders and spontaneous words.

The algorithm is based on the observation that words which co-occur a large number of times in a particular domain in one language have their translations (in the same domain in another language) also co-occurring. This correlation is shown to hold true for certain non-parallel corpora (Koehn & Knight, 2002).

The system consists of the following modules (Figure 3.1):

1. Corpus Cleaning
2. Phrase Structure Parse Trees (PCFG)
3. Phrase Structure to Dependency Structure Conversion
4. Data Sets
5. Seed Lexicon
6. Context Vectors
7. Vector Similarity

8. Evaluation

Modules 1 through 4 comprise the pre-processing stage and remain constant throughout the experiments.

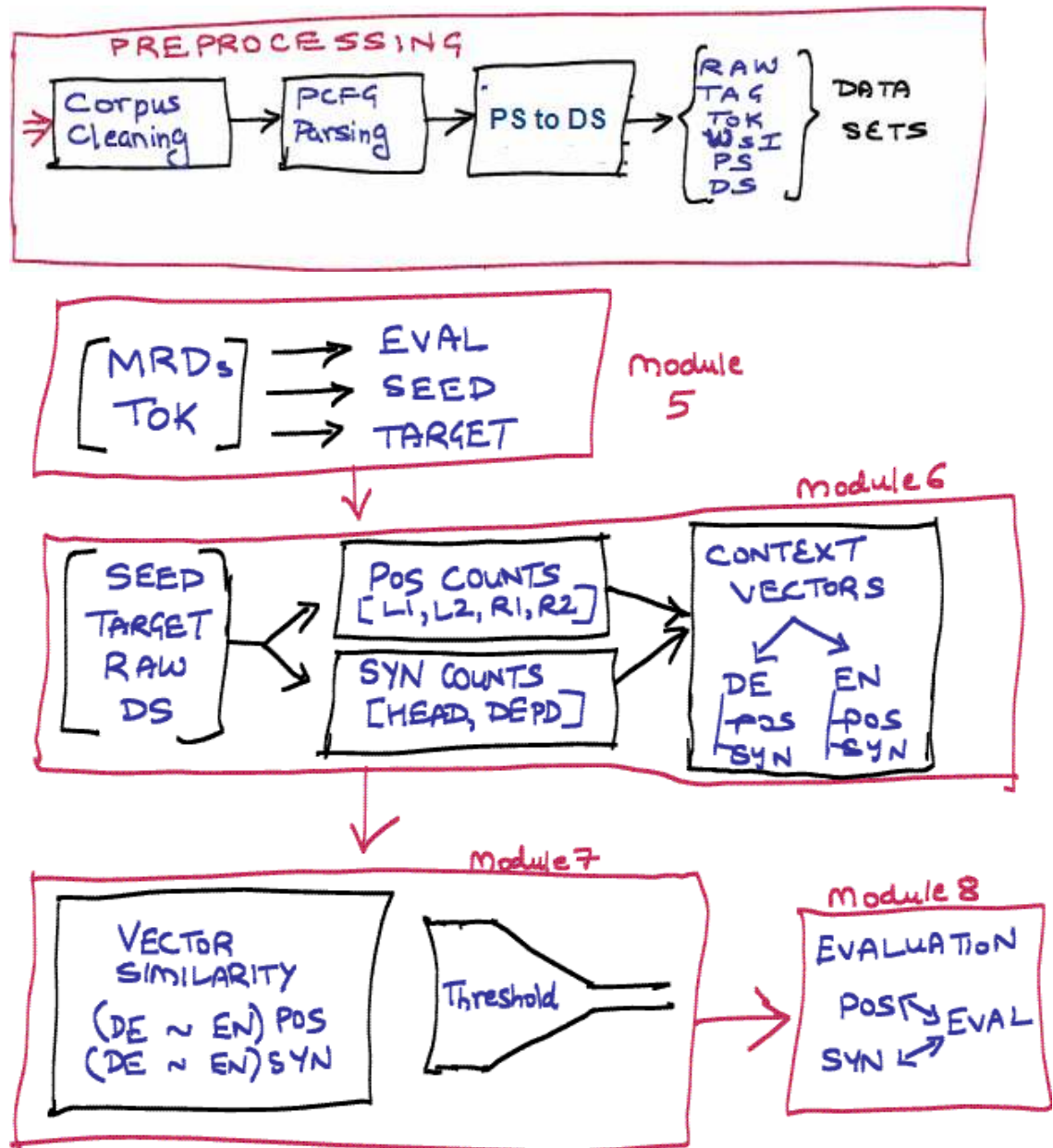


Figure 3.1: Modules in the Lexicon Extraction System

Module 1: Corpus Cleaning

The corpora used are 530 days of news stories from Deutsche Presse Agentur (DPA) on the German side and 446 days of news stories from Wall Street Journal (WSJ) on the English side. [Table 3.1]

Table 3.1: Unprocessed Corpora

	ENGLISH	GERMAN
Data Source	Wall Street Journal (WSJ)	Deutsche Presse Agentur (DPA)
Years Covered	1990,1991 and 1992	1995 and 1996
Size in # files	446 days of news text	530 days of news text
Corpus source	Linguistic Data Consortium (LDC93T3A)	Linguistic Data Consortium (LDC95T11)

The first step is to filter out the SGML tags and extract the actual text from the newswire corpora. Lines enclosed between <TEXT> and </TEXT>, <LP> and </LP> are considered useful data. Note at this stage, a single sentence may run over several lines of text. [Table 3.2]

Table 3.2: Corpora after extracting RAW text

	ENGLISH	GERMAN
Size in MegaBytes	209 MB of clean text	157 MB of clean text
Num. of lines	~4.0 million	~2.7 million
Num. of sentences	~1.5 million	~1.0 million

Next we preprocess the extracted raw corpus to get it ready for parsing. While the English parser has an inbuilt text Tokenizer, the German side does not.

Thus the German corpus is first tokenized (locate sentence boundaries, etc.) before invoking the Stanford parser. Our German text Tokenizer has been adapted from a script provided by Josh Schroeder for the 2nd workshop on SMT (ACL 2007).

Module 2: Phrase Structure Parse Trees

Constituency parsing is carried out using the Stanford PCFG parser (which is available for both English and German languages at <http://nlp.stanford.edu/software/lex-parser.shtml>). The English parser was trained using the Wall Street Journal while the German Parser was trained using the Negra Corpus.

Next we split the English data file and the tokenized German data into small files (ranging from 5 to 10 days of news stories each) and invoke the Stanford parser. After several experiments on time and space efficiency, an upper bound was placed on the length of the sentences parsed. English sentences longer than 50 words and German sentences longer than 30 words were ignored. A summary of the data size is given in [Table 3.3].

Table 3.3: Data Retention after Phrase Structure Parsing

	ENGLISH	GERMAN
# total (IN)	~1.5M sentences	~1.0M sentences
# unparsed	~40K sentences	~229K sentences
# parsed (OUT)	1,522,327 sentences	808,156 sentences

Running the Stanford parser took about 2 months. The next step was to clean the parsed output, discard error messages, etc. The format of the file was one parsed sentence per line.

Module 3: Phrase Structure to Dependency Structure Conversion

Next we use a head percolation table (Collins, 1997) to convert phrase structure parses (PS) to dependency structure parses (DS).

In case of English, the head table was used as provided in the script by Prof. Fei Xia (heuristics were designed on the Penn Treebank (WSJ)). In case of German, a new head table (derived from Stanford Dependency Parser code) was written to reflect the German syntax and labels used. Certain number of sentences failed to be converted into dependency syntax on account of incorrect format, etc. The results are in [Table 3.4].

Table 3.4: Data Retention after Dependency Parsing

	ENGLISH	GERMAN
# total (IN)	1,522,327 sentences	808,156 sentences
# unparsed	329 sentences	10 sentences
# parsed (OUT)	1,521,998 sentences	808,146 sentences

The process was run on parallel processes by splitting data into small files.

Module 4: Data Sets

Now that preprocessing and parsing is done, we compile the different data sets to be used in following modules and experiments. Table 3.5 gives an idea about the actual data size used in our system.

Table 3.5: Statistics in the processed corpora

	ENGLISH	GERMAN
# TYPES	276,402 words	388,291 words
# TOKENS	36,251,168 words	14,311,788 words
# SENTENCES	1,521,998 sentences	808,146 sentences

Executing the process at this step ensures that all data sets span the same sentences and words (thus ignoring the sentences that were dropped in the parsing modules).

- a. **Raw** Text (one sentence per line)
- b. Part-of-Speech **Tagged** Text (one sentence per line)
- c. Word **Frequency** Lists (alphabetical order and frequency order)
- d. **Word-Sentence Index** (the sentence number and position where each word occurs in corpus).
- e. **Phrase** Structure Trees (constituency)
- f. **Dependency** Trees (head-dependent relations only)

Module 5: Seed Lexicon

A seed lexicon is essential before one can start making the context vectors. The seed lexicon comprises of known words, i.e. words for which we know the translations. It makes up the dimensions of the context vectors on both the English and German side.

Module 5 is the process of selecting seed words (i.e. the vector dimensions) and target words (i.e. words for which we will make the vectors). Accordingly we will also compile an evaluation corpus to judge the system performance, (a subset of those target words for which we have the reference translations). There are **two ways to select seed words**: [a] extract identically spelled words (assume they are mutual translations) occurring across corpora (this mainly covers named entities), [b] use translations from a dictionary. Additionally, there are other factors like the length of the words in the seed lexicon, corpus frequency of the words, and the size of the seed lexicon itself. These parameters are discussed in detail in the Experiments section.

Module 6: Context Vectors

In Module 6, we make context vectors for selected target words in terms of frequency of pre-determined seed words. We run two simultaneous simulations – [a] the seed words occur within a context window size 4 of a target word (positional) and [b] the seed words have a grammatical relation (head / dependent) with the target word (syntactic).

Thus in this project, two systems are compared namely the positional context method (Koehn & Knight, 2002) and the syntactic context method (head-dependent relations). The first step is to count word-context (target-seed) occurrences in both the English and German corpora using Raw & Dependency data sets. We are collecting two types of context counts – positional and syntactic – for each word in the corpus. Positional context consists of 4 separate types namely L1, L2, R1, and R2. Syntactic context consists of 2 separate types namely HEAD and DEPENDent. The following example (Figure 3.2) illustrates how the different types are computed.

“He **tried** disguises endlessly, like an actor working out the shadings of a character,” says Mr. Rice.
WORD: tried
POSITIONAL: “ **_L2** He **_L1** disguises **_R1** endlessly **_R2**
SYNTACTIC: He **_DEPD** disguises **_DEPD** endlessly **_DEPD** like **_DEPD**
 says **_HEAD**

Figure 3.2: Context Vector Example

The second step is to select the target words, i.e. words for which we do not know the translations and for which we would compute context vectors from the counts collected in the above step. The qualifying criteria for target words is no punctuation, no numbers, no special characters, not in seed lexicon, have frequency > 2. It was observed that filtering out words which occur only once or twice in the corpus reduces the number of words by half. Table 3.6 lists the actual number of target words.

Table 3.6: Words before and after implementing the filtering criteria

	ENGLISH	GERMAN
# words before filtering (input)	276,402	388,291
# words after filtering (target)	74,434	106,366

Now that we have the seed words (known), target words (unknown), and the raw counts of occurrence of target words in context of seed words, the third step is to fill in the context vectors. Just to give an idea of how many raw counts we have for our set of target words in context of our seed words, see Table 3.7.

Table 3.7: Number of relevant raw counts for each type

	ENGLISH	GERMAN
# target_seed in Positional L1	859,197	479,206
# target_seed in Positional L2	1,410,389	684,971
# target_seed in Positional R1	921,081	551,337
# target_seed in Positional R2	1,361,135	656,769
# target_seed in Syntactic HEAD	1,268,863	637,278
# target_seed in Syntactic DEPD	1,368,253	733,111

We have 74,434 English (target) vectors and 106,366 German (target) vectors. Each vector has x dimensions which refers to the number of seed words. The value for each dimension is computed as follows:

$$P(\text{target} \mid \text{seed}) = \mathbf{abs} [\mathbf{log} \{ \text{freq}(\text{target} \ \& \ \text{seed}) / \text{freq}(\text{seed}) \}]$$

Module 7: Vector Similarity

In order to ensure we can compare German vectors with English vectors and compute their similarity, each dimension (seed) words is assigned a unique integer for reference. For example, the German word “erforschung” is assigned an ID 4032, the same as that of its English translation, “investigation.” Note however, the seed lexicon contains two translations for the German word “erforschung” and two translations for the English word “investigation.” Hence we also have ID 5075 assigned to “erforschung” and “exploration,” and an ID 898 assigned to “untersuchung” and “investigation.” This takes care of seed words having multiple senses or translations which is to be expected in non-parallel corpora.

Also, note we are considering each of the 6 context types (L1,L2,R1,R2,HEAD,DEPD) separately. Thus in reality, a syntactic context vector consists of $2n$ (n seeds * 2) dimensions, while a positional context vector consists of $4n$ (n seeds * 4) dimensions. To help reduce such a large overhead, we work with sparse vectors, i.e. if a particular dimension has a zero value, we do not store it.

There are several similarity metrics used in this domain. We use three similarity metrics – [a] **cosine similarity**; the closer the value is to 1 i.e. $\cos(0)$ the similar it is (Figure 3.4), [b] **city block metric** (Rapp 1999, Koehn & Knight, 2002); add the absolute differences of all components such that smaller the value, similar it is (Figure 3.3), and [c] **match metric**; a naïve method which simply counts the number of non-zero dimensions common between two vectors.

4 Experiments

In this section we describe the data used and the different parameters which were tuned to evaluate our system performance. As mentioned above, there are two main systems to be compared – positional context vectors (Koehn & Knight, 2002) and syntactic context vectors (contribution of this project).

DATA SAMPLES We will first display data samples used as input and output in the preprocessing modules (Modules 1 through 4).

1A Unprocessed Data – English

```
<DOC>
<DOCNO>
WSJ900402-0195
</DOCNO>
<DOCID>
900402-0195.
</DOCID>
<HL>
  Who's News:
  Timken Co.
</HL>
<DATE>
04/02/90
</DATE>
<SO>
WALL STREET JOURNAL (J), NO PAGE CITATION
</SO>
<CO>
  WNEWS TKR
</CO>
<LP>
  TIMKEN Co. (Canton, Ohio) - Larry R. Brown, managing
partner of the law firm Day, Ketterer, Raley, Wright & Rybold
of Canton, Ohio, was named vice president and general
counsel, a new post at this specialty steels and bearings
company.
</LP>
<TEXT>
</TEXT>
</DOC>
```

1B Unprocessed Data – German

```
<DOC>
<DOCID> dpger950109.0002 </DOCID>
<STORYID cat=s pri=r sel=eudpa> x0031 </STORYID>
<FORMAT> &D3; &D1; </FORMAT>
<HEADER> Ergebnisse der 24. 01-09 0180 </HEADER>
<PREAMBLE>
Ergebnisse der 24. Segelflug-Weltmeisterschaften &QC;
</PREAMBLE>
<TEXT>
<p>
    Beim Vormittagsfixing in London kostete die Feinunze Gold
    372,45
    (375,90) Dollar. Der Kilobarrren wurde in Frankfurt mit 18 710
    (18
    760) DM fixiert. Dpa ak gr
</p>
</TEXT>
<TRAILER>
AP-NY-01-09-95 0434EST &QL;
</TRAILER>
</DOC>
```

2A Raw unparsed text – English

TIMKEN Co. (Canton, Ohio) - Larry R. Brown, managing partner of the law firm Day, Ketterer, Raley, Wright & Rybold of Canton, Ohio, was named vice president and general counsel, a new post at this specialty steels and bearings company.

2B Raw unparsed tokenized text – German

Beim Vormittagsfixing in London kostete die Feinunze Gold 372,45
(375,90) Dollar . Der Kilobarrren wurde in Frankfurt mit 18 710
(18 760) DM fixiert. Dpa ak gr

3A Parsed (Phrase Structure) – English

Parsing [sent. 1 len. 50]: TIMKEN Co. -LRB- Canton , Ohio -RRB- -
 - Larry R. Brown , managing partner of the law firm Day ,
 Ketterer , Raley , Wright & Rybold of Canton , Ohio , was named
 vice president and general counsel , a new post at this specialty
 steels and bearings company .

(ROOT (NP (NP (NNP TIMKEN) (NNP Co.)) (PRN (-LRB- -LRB-) (NP (NNP
 Canton)) (, ,) (NP (NNP Ohio)) (-RRB- -RRB-)) (: --) (NP (NP (NNP
 Larry) (NNP R.) (NNP Brown)) (, ,) (PP (VBG managing) (NP (NP (NN
 partner)) (PP (IN of) (NP (DT the) (NN law) (NN firm)))) (NP-TMP
 (NNP Day)))) (, ,) (SBAR (S (NP (NP (NNP Ketterer) (, ,) (NNP
 Raley) (, ,) (NNP Wright) (CC &) (NNP Rybold)) (PP (IN of) (NP
 (NNP Canton) (, ,) (NNP Ohio) (, ,)))) (VP (VBD was) (VP (VBN
 named) (NP (NP (NN vice) (NN president)) (CC and) (NP (NP (JJ
 general) (NN counsel)) (, ,) (NP (DT a) (JJ new) (NN post)))) (PP
 (IN at) (NP (DT this) (NN specialty) (NNS steels) (CC and) (NNS
 bearings) (NN company))))))))) (. .))

3B Parsed (Phrase Structure) – German

Parsing [sent. 1 len. 14]: Beim Vormittagsfixing in London
 kostete die Feinunze Gold 372,45 (375,90) Dollar .

(ROOT (NUR (S (PP (APPRART Beim) (NN Vormittagsfixing) (PP (APPR
 in) (NE London))) (VVFIN kostete) (NP (ART die) (ADJA Feinunze)
 (NN Gold)) (NP (NM (CARD 372,45) (CARD () (CARD 375,90) (CARD)))
 (NN Dollar))) (\$. .)))

Parsing [sent. 2 len. 94]: Der Kilobarrren wurde in Frankfurt mit
 18 710 (18 760) DM fixiert. Dpa ak gr Melbourne (dpa) - 1.
 Wettfahrt : 1. Jose van der Ploeg (Spanien) , 2. Paul McKenzie
 (Australien) , 3. Xavier Rohart (Frankreich) , 4. Hank
 Lammens (Kanada) , 5. Emanuele Vaccari (Italien) , 6. Fredrik
 L & D4 ; & D4 ; f (Schweden) dpa ll gr Frankfurt am Main
 (dpa) - Der Deutsche Aktienmarkt hat am Montag uneinheitlich
 tendiert .

Sentence too long (or zero words).

SENTENCE_SKIPPED_OR_UNPARSABLE

4A Parsed (Dependency Structure) – English

sent: TIMKEN Co. -LRB- Canton , Ohio -RRB- -- Larry R. Brown ,
 managing partner of the law firm Day , Ketterer , Raley , Wright
 & Rybold of Canton , Ohio , was named vice president and general
 counsel , a new post at this specialty steels and bearings
 company .

tags: NNP NNP -LRB- NNP , NNP -RRB- : NNP NNP NNP , VBG NN IN DT
 NN NN NNP , NNP , NNP , NNP CC NNP IN NNP , NNP , VBD VBN NN NN
 CC JJ NN , DT JJ NN IN DT NN NNS CC NNS NN .

sent_leng=51 root_idx=11

1 2 # TIMKEN Co.
 2 11 # Co. Brown
 3 7 # -LRB- -RRB-
 4 7 # Canton -RRB-
 5 7 # , -RRB-
 6 7 # Ohio -RRB-
 7 11 # -RRB- Brown
 8 11 # -- Brown
 9 11 # Larry Brown
 10 11 # R. Brown
 11 11 # Brown Brown
 12 11 # , Brown
 13 11 # managing Brown
 14 19 # partner Day
 15 19 # of Day
 16 18 # the firm
 17 18 # law firm
 18 15 # firm of
 19 13 # Day managing
 20 11 # , Brown
 21 27 # Ketterer Rybold
 22 27 # , Rybold
 23 27 # Raley Rybold
 24 27 # , Rybold
 25 27 # Wright Rybold
 26 27 # & Rybold
 27 34 # Rybold named
 28 27 # of Rybold
 29 31 # Canton Ohio
 30 31 # , Ohio
 31 28 # Ohio of
 32 31 # , Ohio
 33 34 # was named
 34 11 # named Brown
 35 36 # vice president
 36 43 # president post
 37 43 # and post

38 39 # general counsel
 39 43 # counsel post
 40 43 # , post
 41 43 # a post
 42 43 # new post
 43 34 # post named
 44 34 # at named
 45 50 # this company
 46 50 # specialty company
 47 50 # steels company
 48 50 # and company
 49 50 # bearings company
 50 44 # company at
 51 11 # . Brown

4B Parsed (Dependency Structure) – German

sent: Beim Vormittagsfixing in London kostete die Feinunze Gold
 377,90 (377,55) Dollar .

tags: APPRART NN APPR NE VVFIN ART ADJA NN CARD CARD CARD CARD NN
 \$.

Sent_leng=14 root_idx=13

1 13 # Beim Dollar
 2 1 # Vormittagsfixing Beim
 3 1 # in Beim
 4 3 # London in
 5 13 # kostete Dollar
 6 8 # die Gold
 7 8 # Feinunze Gold
 8 13 # Gold Dollar
 9 12 # 377,90)
 10 12 # ()
 11 12 # 377,55)
 12 13 # -RRB- Dollar
 13 13 # Dollar Dollar
 14 13 # . Dollar

5A Head Percolation Table to convert PS to DS – English

S right VP/SBAR/S
SINV right VP/SINV
SQ right VP/SQ
SBAR right S/SBAR/SINV
SBARQ right SQ/SBARQ
RRC right VP/RRC
NX right NX/NN/NNS/NNP/NNPS
NP right NP/NN/NNS/NNP/NNPS/NX/EX/CD/QP/JJ/JJR/JJS/PRP/DT/POS/FW
NAC right NAC/NN/NNS/NNP/NNPS/NX
WHNP right WDT/WP/NP/NN/NNS/NNP/NNPS/NX/WHNP
QP right CD/QP
ADJP right JJ/JJR/VBN/ADJP/*?*
WHADJP right JJ/WHADJP
VP right VP/VB/VBN/VBP/VBZ/VBG/VBD/NP-PRD/ADJP-PRD/PP-PRD/*?*
PP left IN/TO/VBG/PP
PP-PRD left IN/TO/VBG/PP
WHPP left IN/WHPP
ADVP right ADVP/RB/RBR/RBS/WRB
WHADVP right WRB/WHADVP
PRT right RP/PRT
INTJ left UH/INTJ
UCP right UCP
X right X
LST left LS

5B Head Percolation Table to convert PS to DS – German

AA right ADJD/ADJA
ADJX right ADJX/ADJA/ADJD
ADVX right ADVX
AP right ADJD/ADJA/CAP/AA/ADV
AVP right ADV/AVP/ADJD/PROAV/PP
C right KOUS
CAC right APPR/AVP
CAP right ADJD/ADJA/CAP/AA/ADV
CAVP right ADV/AVP/ADJD/PWAV/APPR/PTKVZ
CCP right AVP
CD right CD
CNP right NN/NE/MPN/NP/CNP/PN/CARD
CPP right APPR/PROAV/PP/APP
CS right S/CS
CVP right VP/CVP
CVZ right VZ
DM left PTKANT/ITJ/KON/FM
EN left NX
EN-ADD left NX
FKONJ left LK/VC/MF/NF/VF
FKOORD left LK/C/FKONJ/MF/VC
FX right FM/FX
MPN right NE/FM/CARD
MTA right ADJA/ADJD/NN
NM right CARD/NN
NN right NN
NP right NN/NE/MPN/NP/CNP/PN/CARD
NR right NR
NUR left S/CS/VP/CVP/NP/XY/CNP/AVP/CAVP
NX right NX/NE/NN/EN/EN-ADD/FX/ADJX/PIS/ADVX/CARD/TRUNC
P left SIMPX
PP left KOKOM/APPR/PROAV
PX left APPR/APPRART/PX
R left C/R/VC
S right VVFIN/VMFIN/VAFIN/VVIMP/VP/CVP/VAIMP/S/CS/NP/PRELS
SIMPX right LK/VC/SIMPX/C/FKoord/MF/
VC left VXINF
VF left NX/ADJX/PX/ADVX/EN/SIMPX
VP right VVINF/VVIZU/VVPP/VZ/VAINF/VMINF/VMPP/VAPP/PP
VXFIN left VXFIN/VVFIN
VXINF right VXINF/VVPP/VVINF
VZ right VVINF/VAINF/VMINF/VVFIN/VVIZU/PRTZU/APPR/PTKZU

PARAMETER TUNING We will now describe the different choices of the parameters experimented upon in the project.

This first parameter is the choice of seed words. As described in the previous section, we used two different sets of seed words – identically spelled words across English and German corpora, bootstrapping entries from a bilingual dictionary.

We first tried to use identically spelled words as the seed lexicon (Koehn & Knight 2002). However this proved dubious as evaluating this seed lexicon against a dictionary yielded an accuracy of 35% (894 out of 2,550 translations). Such a low starting point was unwanted. So we abandoned this idea initially and instead extracted one-word translations (i.e. both the English and German side consist of a single word) from 2 Machine Readable Dictionaries (MRDs) [see Table 4.1].

Dict 1: <http://odg.info>

Dict 2: <http://dict.tu-chemnitz.de/de-en/lists/index.html>

Table 4.1: Dictionary statistics

# total entries in DICT 1	140,966
# 1-WORD entries in DICT 1 & EN/DE corpora	43,152 (A)
# total entries in DICT 2	44,610
# 1-WORD entries in DICT 2 & EN/DE corpora	6,358 (B)
# 1-WORD entries from merged DICT 1,2 & EN/DE corpora	45,562 (A+B)

Each of these 45,562 entries was given a ranking score which was the average rank of frequency of the German word in German corpus and the English word in English corpus. (The most frequent word in a corpus is awarded a rank of 1 and so on).

$$\text{Rank (dictionary entry)} = (0.5 * \text{freq}_{\text{DE}}) + (0.5 * \text{freq}_{\text{EN}})$$

So our dictionary has 45,562 entries (ranked 2.5 through 26385) which we will split into seed set and evaluation set. The test set will help us evaluate the effectiveness of syntactic context methodology over positional context methodology in extracting a translation lexicon.

From the top one-third scored listing ($45,562 / 3 = 15,184$ entries), we randomly select 100 entries. *Note it is possible that a particular word may have multiple translations in the dictionary.* Hence, we cross reference the German words (from the 100 entries) with the entire list and record all the possible English translations. This expanded list will become our evaluation set. There are 364 entries in our **EVALUATION DATA**.

The seed lexicon is obtained by filtering out the evaluation data entries and extracting the top **6000** entries from the remaining ranked set. This is our **SEED LEXICON**. *Note there are 6000 unique entries, however many German and English words repeat across entries (multiple translations).* Thus actually there are 3,407 German and 3,684 English words in the seed lexicon.

A sample English Positional Vector for the word “abashed,” where each non-zero dimension is of the form `dim_name=value`, is as follows:

```
abashed : 3205_L1=6  1_L1=14  3_L1=14  43_L2=12  25_L2=12  13_L2=12
34_L2=12  3216_L2=7  1926_L2=7  134_L2=11  5810_R1=8  457_R1=8
3723_R1=8 2328_R1=8  2_R2=13 267_R2=13  78_R2=10
```

A sample German Syntactic Vector for the word “abbröckelte,” where each non-zero dimension is of the form dim_name=value, is as follows:

```
abbröckelte : 14_DEPD=10  16_DEPD=10  970_DEPD=10  4_DEPD=10
1694_DEPD=10  45_DEPD=10  6_DEPD=10  331_DEPD=6  747_DEPD=6
20_DEPD=10  670_DEPD=7  1222_DEPD=7  487_DEPD=7  227_DEPD=7
1882_DEPD=7  390_DEPD=7  611_DEPD=7  659_DEPD=7  1091_DEPD=7
1373_DEPD=7  19_DEPD=10  142_DEPD=8  4766_DEPD=8  71_DEPD=8
94_DEPD=8  203_DEPD=7
```

Although we have each of 106,366 German vectors to compare with each of 74,434 English vectors, only 100 German vectors (from the EVAL set in Module 5) are compared with the 74,434 English vectors. This is done to reduce the processing time and because we have the reference translations for only 100 German words (364 entries in the evaluation set).

We also implement a second seed set, a modified version of the earlier failed attempt to collect identically spelled words. Thus words which have the same spelling were sorted with respect to word length. We also collected words which could be considered translations through application of the following spelling transformation rules: (Koehn and Knight, 2002)

k and z in german => c in english

tät in german => ty in english

Number of common words of length 4 = 1259
Number of common words of length 5 = 1690
Number of common words of length 6 = 1676
Number of common words of length 7 = 1354
Number of common words of length 8 = 892
Number of common words of length 9 = 575
Number of common words of length 10 = 304
Number of common words of length 11 = 147
Number of common words of length 12 = 70
Number of common words of length 13 = 36
Number of common words of length 14 = 19
Number of common words of length 15 = 8
Number of common words of length 16 = 3
Number of common words of length 17 = 1
Number of common words of length 18 = 3
Number of common words of length 19 = 2
Number of words through spelling transformation = 562

We thus filtered out the smaller length words because longer length words have a higher probability of being accurate translations. We selected all words of length 8 and above thus obtaining 2,376 seed words in our seed set 2.

As a preliminary evaluation, we obtained the 364 mappings from the evaluation set, and compared the corresponding positional vector and syntactic vector similarity scores for each.

A sample of the evaluation data (i.e. the reference translations to be evaluated in our system) is as follows:

Table 4.2: Sample Evaluation Data

GERMAN	ENGLISH
blei	lead
forscher	researched, researcher, explorer, inquirer, researchers
nachrichten	newscast, news
gefecht	skirmish, battle, encounter
tragik	tragedy

Both city block and cosine similarity measures were used to compare the syntactic and positional context vector systems. The syntactic tally represents the number of times syntactic vectors predicted a better (closer) translation and the positional tally represents the same for positional vectors.

Table 4.3: System Evaluation

METRIC	SYNTACTIC TALLY	POSITIONAL TALLY
CITY BLOCK	301 / 364	63 / 364
COSINE SIMILAITY	216 / 364	148 / 364

5 Discussion

There was one major issue unresolved in this project. When similarity scores were computed by comparing each English vector to a German vector, none of the top 5 translations appear in the evaluating dictionary. We found that neither the positional nor the syntactic systems predicted the correct translation in the top 5. That is the words which were predicted by the system as translations were not actual translations. However, on comparing both systems against each other it was found that syntactic vectors did predict better translations than positional, even though it was not in the top 5. For example, a syntactic system predicted the correct translation at rank 23 while the positional system predicted it at rank 32.

A few error analyses were conducted. It was found that the words which occurred as top ranking translations were random. Some of them were high frequency words (thus implying their vectors would match with most word vectors on account of having such a large context that they match even if it is not a translation), but most were random unrelated words. Some of the potential translations don't have a large enough common context.

Experiments were also carried out to emulate the Koehn and Knight, 2002 paper. However the accuracy figures were much lower than reported. And still the top 5 scores were not achieved.

I have tried using both types of seed lexicon (identical spelling and dictionary) and the three types of similarity metrics (cosine, city block, matches). This led me to believe that maybe my evaluation corpus was wrong. Maybe the words I chose to evaluate were not fit for this corpus.

Note that the evaluation corpus contains 364 (100 unique entries) only. Each of the 100 German word vector is compared against 74,434 English word vectors. Major reason for this filtering is that it is computationally expensive (take months) to compare each German word with each English word in the corpus. A better evaluation set was also tried with no significant improvement.

6 Conclusion

German and English newswire corpora from different time periods were parsed and used to extract a translation lexicon by bootstrapping two different sets of seed words (identical spelling and high-frequency entries from a dictionary). Two competitive systems were developed for building and evaluating context vectors of words with unknown translations. Positional vector system involved weighting an unknown word with frequencies of known (seed) words in a surrounding window of size 4 (two words to the left and two words to the right). Syntactic vector system involved weighting an unknown word with frequencies of known (seed) words which were either its head or dependent in a sentence.

While ranking vector similarity scores resulted in better performance by the syntactic system, none of the systems were able to compute the correct translation as the most similar or even top 5 similar words.

Therefore, while the project helped prove that syntactic dependencies are a better predictor than surrounding words (positional), the data was inconclusive in predicting the correct translation accurately (i.e. none of the translations appeared in the top rankings of similar word vectors).

Future avenues of investigation include using more pre-processing (stemmers, morphological analyzers) to help make up for the data sparsity. The major issue with non parallel corpora is that there is no guarantee mutual translations will occur with sufficient frequency to make up large enough vectors. Nevertheless, subsequent error analyses and experiments may prove definite success.

7 Bibliography

S. Armstrong, K.Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D.Yarowsky (Editors). BOOK. Natural Language Processing Using Very Large Corpora, *Kluwer Academic Publishers, Netherlands, 1999*.

P. Brown, J. Cocke, S. Pietra, V. Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin. 1990. A Statistical Approach to Machine Translation. In *Computational Linguistics, Vol. 16*, pp. 79-85.

C. Callison-Burch, D. Talbot, and M. Osborne. 2004. Statistical Machine Translation with Word- and Sentence-Aligned Parallel Corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, July 2004*, pp. 175-182.

M. Collins. 1997. Three Generative Lexicalized Models for Statistical Parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97), Madrid, July 1997*, pp. 16-23.

M. Diab and S. Finch. 2000. A Statistical word-level Translation Model for Comparable Corpora. In *Proceedings of the Conference on Content-based multimedia information access (RIAO), 2000*.

P. Fung. 2000. A Statistical view on Bilingual Lexicon Extraction: From Parallel Corpora to non-Parallel Corpora. In J. Veronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*, pages 219-236. Kluwer Academic Publishers, 2000.

P. Fung and P. Cheung. 2004. Mining Very Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-04), Barcelona, July 2004*.

P. Fung and L. Yee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 36th Conference for the Association for Computational Linguistics, Montreal, August 1998*, pp. 414 – 420.

E. Gaussier, J. Renders, I. Matveeva, C. Goutte, and H. Dejean. 2004. A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 526-533.

Z. Harris. Distributional structure. BOOK. In J.J. Katz, editor, *The Philosophy of Linguistics*, pages 26-47. New York: Oxford University Press, 1985.

R. Hwa, C. Nichols, and K. Sima'an. 2006. Corpus Variations for Translation Lexicon Induction. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06), Cambridge, August 2006*.

P. Koehn and K. Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. In *Proceedings of the Workshop of the Association for Computational Linguistics Special Interest Group on the Lexicon (SIGLEX), Philadelphia, July 2002*, pp. 9-16.

I. D. Melamed. 2000. Models of Translational Equivalence among Words. In *Computational Linguistics, Vol. 26*, pp. 221-249.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics, Vol. 29*, pp.19-51.

P. Otero and J Campos. 2005. An Approach to Acquire Word Translations from Non-Parallel Texts. In *Proceedings of the 12th Portuguese Conference on Artificial Intelligence (EPIA-05), Portugal, December 2005*, pp.600-610.

C. Peters and E. Picchi. 1997. Using Linguistic Tools and Resources in Cross-Language Retrieval. In *Cross-Language Text and Speech Retrieval Papers from the 1997 AAI Spring Symposium, Technical Report SS-97-05, AAI Press*, pp. 179-188

C. Quirk, A. Menezes, and C. Cherry. 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pp. 271-279.

R. Rapp. 1995. Identifying Word Translations in Non-Parallel Texts. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics, Cambridge, June 1995*, pp. 320-322.

R. Rapp. 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 (ACL-99)*, pp. 519-526.

D. Yarowsky and G.Ngai. 2001. Inducing Multilingual POS Taggers and NP Brackets via Robust Projection across Aligned Corpora. In *Proceedings of the*

2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01), Pittsburgh, June 2001, pp. 200-207.