Learning a Translation Lexicon from Monolingual Corpora

Philipp Koehn and **Kevin Knight**

Information Sciences Institute, University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292
koehn@isi.edu, knight@isi.edu

Abstract

This paper presents work on the task of constructing a word-level translation lexicon purely from unrelated monolingual corpora. We combine various clues such as cognates, similar context, preservation of word similarity, and word frequency. Experimental results for the construction of a German-English noun lexicon are reported. Noun translation accuracy of 39% scored against a parallel test corpus could be achieved.

1 Introduction

Recently, there has been a surge in research in machine translation that is based on empirical methods. The seminal work by Brown et al. [1990] at IBM on the Candide system laid the foundation for much of the current work in **Statistical Machine Translation** (SMT). Some of this work has been re-implemented and is freely available for research purposes [Al-Onaizan et al., 1999].

Roughly speaking, SMT divides the task of translation into two steps: a word-level translation model and a model for word reordering during the translation process.

The statistical models are trained on parallel corpora: large amounts of text in one language along with their translation in another. Various parallel texts have recently become available, mostly from government sources such as parliament proceedings (the Canadian Hansard, the minutes of the European parliament[1]) or law texts (from Hong Kong).

Still, for most language pairs, parallel texts are hard to come by. This is clearly the case for low-density languages such as Tamil, Swahili, or Tetun. Furthermore, texts derived from parliament speeches may not be appropriate for a particular targeted domain. Specific parallel texts can be constructed by hand for the purpose of training an SMT system, but this is a very costly endeavor.

On the other hand, the digital revolution and the wide-spread use of the World Wide Web have proliferated vast amounts of monolingual corpora. Publishing text in one language is a much more natural human activity than producing parallel texts. To illustrate this point: The world wide web alone contains currently over two billion pages, a number that is still growing exponentially. According to Google,[2] the word directory occurs 61 million times, empathy 383,000 times, and reflex 787,000 times. In the Hansard, each of these words occurs only once.

The objective of this research to build a translation lexicon solely from monolingual corpora. Specifically, we want to automatically generate a one-to-one mapping of German and English nouns. We are testing our mappings against a bilingual lexicon of 9,206 German and 10,645 English nouns.

The two monolingual corpora should be in a fairly comparable domain. For our experiments we use the 1990-1992 Wall Street Journal corpus

[1] Available for download at http://www.isi.edu/~koehn/publications/europarl/

[2] http://www.google.com/

on the English side and the 1995-1996 German news wire (DPA) corpus on the German side. Both corpora are news sources in the general sense. However, they span different time periods and have a different orientation: the World Street Journal covers mostly business news, the German news wire mostly German politics.

For experiments on training probabilistic translation lexicons from parallel corpora and similar tasks on the same test corpus, refer to our earlier work [Koehn and Knight, 2000, 2001].

2 Clues

This section will describe clues that enable us to find translations of words of the two monolingual corpora. We will examine each clue separately. The following clues are considered:

- **Identical words** – Two languages contain a certain number of identical words, such as computer or email.
- **Similar Spelling** – Some words may have very similarly written translations due to common language roots (e.g. Freund and friend) or adopted words (e.g. Webseite and website).
- **Context** – Words that occur in a certain context window in one language have translations that are likely to occur in a similar context window in the other language (e.g. Wirtschaft co-occurs frequently with Wachstum, as economy does with growth).
- **Similarity** – Words that are used similarly in one language should have translations that are also similar (e.g. Wednesday is similar to Thursday as Mittwoch is similar to Donnerstag).
- **Frequency** – For comparable corpora, frequent words in one corpus should have translations that are frequent in the other corpus (e.g. for news corpora, government is more frequent than flower, as its translation Regierung is more frequent than Blume.

We will now look in detail how these clues may contribute to building a German-English translation lexicon.

2.1 Identical words

Due to cultural exchange, a large number of words that originate in one language are adopted by others. Recently, this phenomenon can be seen with words such as Internet, or Aids.

These terms may be adopted verbatim, or changed by well-established rules. For instance, immigration (German and English) has the Portuguese translation immigração, as many words ending in -tion have translations with the same spelling except for the ending changed to -ção.

We examined the German words in our lexicon and tried to find English words that have the exact same spelling. Surprisingly, we could count a total of 976 such words. When checking them against a benchmark lexicon, we found these mappings to be 88% correct.

The correctness of word mappings acquired in this fashion depends highly on word length. This is illustrated in Table 1: While identical 3-letter words are only translations of each other 60% of the time, this is true for 98% of 10-letter words. Clearly, for shorter words, the accidental existence of an identically spelled word in the other language word is much higher. This includes words such as fee, ton, art, and tag.

Length	Number of words		Accuracy
	correct	wrong	
3	33	22	60%
4	127	48	69%
5	129	22	85%
6	162	13	93%
7	131	4	97%
8	86	4	96%
9	80	4	95%
10	57	1	98%
11+	50	3	94%

Table 1: Testing the assumption that identically spelled words are in fact translations of each other: The accuracy of this assumption depends highly on the length of the words (see Section 2.1)

Knowing this allows us to restrict the word length to be able to increase the accuracy of the collected word pairs. For instance, by relying only on words at least of length 6, we could collect 622 word pairs with 96% accuracy. In our experiments, however, we included all the words pairs.

As already mentioned, there are some well-established transformation rules for the adoption of words from a foreign language. For German to English, this includes replacing the letters k and z by c and changing the ending -tät by -ty. Both these rules can be observed in the word pair Elektrizität and electricity.

By using these two rules, we can gather 363 additional word pairs of which 330, or 91%, are in fact translations of each other. The combined total of 1339 (976+363) word pairs are separated and form the seed for some of the following steps.

2.2 Similar Spelling

When words are adopted into another language, their spelling might change slightly in a manner that can not be simply generalized in a rule. Observe, for instance website and Webseite. This is even more the case for words that can be traced back to common language roots, such as friend and Freund, or president and Präsident.

Still, these words – often called **cognates** – maintain a very similar spelling. This can be defined as differing in very few letters. This measurement can be formalized as the number of letters common in sequence between the two words, divided by the length of the longer word.

The example word pair friend and freund shares 5 letters (fr-e-nd), and both words have length 6, hence there spelling similarity is 5/6, or 0.83. This measurement is called **longest common subsequence ratio** [Melamed, 1995]. In related work, string edit distance (or, Levenshtein distance) has been used [Mann and Yarowski, 2001].

With this computational means at hand, we can now measure the spelling similarity between every German and English word, and sort possible word pairs accordingly. By going through this list starting at the top we can collect new word pairs. We do this is in a greedy fashion – once a word is assigned to a word pair, we do not look for another match. Table 2 gives the top 24 generated word pairs by this algorithm.

German	English	Score	
Organisation	organization	0.92	correct
Präsident	president	0.90	correct
Industrie	industries	0.90	correct
Parlament	parliament	0.90	correct
Interesse	interests	0.89	correct
Institut	institute	0.89	correct
Satellit	satellite	0.89	correct
Dividende	dividend	0.89	correct
Maschine	machine	0.88	correct
Magazin	magazine	0.88	correct
Februar	february	0.88	correct
Programm	program	0.88	correct
Gremium	premium	0.86	wrong
Branche	branch	0.86	wrong
Volumen	volume	0.86	correct
Januar	january	0.86	correct
Warnung	warning	0.86	correct
Partie	parties	0.86	correct
Debatte	debate	0.86	correct
Experte	expert	0.86	correct
Investition	investigation	0.85	wrong
Mutter	matter	0.83	wrong
Bruder	border	0.83	wrong
Nummer	number	0.83	correct

Table 2: First 24 word pairs collected by finding words with most similar spelling in a greedy fashion.

The applied measurement of spelling similarity does not take into account that certain letter changes (such as z to s, or dropping of the final e) are less harmful than others. Tiedemann [1999] explores the automatic construction of a string similarity measure that learns which letter changes occur more likely between cognates of two languages. This measure is trained, however, on parallel sentence-aligned text, which is not available here.

Obviously, the vast majority of word pairs can not be collected this way, since their spelling shows no resemblance at all. For instance, Spiegel and mirror share only one vowel, which is rather accidental.

2.3 Similar Context

If our monolingual corpora are comparable, we can assume a word that occurs in a certain context should have a translation that occurs in a similar context.

Context, as we understand it here, is defined by the frequencies of context words in surrounding positions. This local context has to be translated into the other language, and we can search the word with the most similar context.

This idea has already been investigated in earlier work. Rapp [1995, 1999] proposes to collect counts over words occurring in a four word window around the target word. For each occurrence of a target word, counts are collected over how often certain context words occur in the two positions directly ahead of the target word and the two following positions. The counts are collected separately for each position and then entered into in a context vector with an dimension for each context word in each position. Finally, the raw counts are normalized, so that for each of the four word positions the vector values add up to one. Vector comparison is done by adding all absolute differences of all components.

Fung and Yee [1998] propose a similar approach: They count how often another word occurs in the same sentence as the target word. The counts are then normalized by a using the tf/idf method which is often used in information retrieval [Jones, 1979].

The need for translating the context poses a chicken-and-egg problem: If we already have a translation lexicon we can translate the context vectors. But we can only construct a translation lexicon with this approach if we are already able to translate the context vectors.

Theoretically, it is possible to use these methods to build a translation lexicon from scratch [Rapp, 1995]. The number of possible mappings has complexity $O(n!)$, and the computing cost of each mapping has quadratic complexity $O(n^2)$. For a large number of words n – at least more than 10,000, maybe more than 100,000 – the combined complexity becomes prohibitively expensive.

Because of this, both Rapp and Fung focus on expanding an existing large lexicon to add a few novel terms.

Clearly, a seed lexicon to bootstrap these methods is needed. Fortunately, we have outlined in Section 2.1 how such a seed lexicon can be obtained: by finding words spelled identically in both languages.

We can then construct context vectors that contain information about how a new unmapped word co-occurs with the seed words. This vector can be translated into the other language, since we already know the translations of the seed words.

Finally, we can look for the best matching context vector in the target language, and decide upon the corresponding word to construct a word mapping.

Again, as in Section 2.2, we have to compute all possible word – or context vector – matches. We collect then the best word matches in a greedy fashion. Table 3 displays the top 15 generated word pairs by this algorithm. The context vectors are constructed in the way proposed by Rapp [1999], with the difference that we collect counts over a four *noun* window, not a four word window, by dropping all intermediate words.

German	English	Score	
Jahr	mr	5.03024	wrong
Regierung	government	5.54937	correct
Prozent	percent	5.57756	correct
Angabe	us	5.73654	wrong
Mittwoch	company	5.83199	wrong
Donnerstag	time	5.90623	wrong
Präsident	president	5.93884	correct
Dienstag	year	5.94611	wrong
Staat	state	5.96725	correct
Zeit	people	6.05552	wrong
Freitag	officials	6.11668	wrong
Montag	week	6.13604	wrong
Krieg	war	6.13604	correct
Woche	yesterday	6.15378	wrong
Krankheit	disease	6.20817	correct
Kirche	church	6.21477	correct
Unternehmen	companies	6.22896	correct
Ende	money	6.28154	wrong
Streik	strike	6.28690	correct
Energie	energy	6.29883	correct
Öl	oil	6.30794	correct
Markt	market	6.31116	correct
Wirtschaft	economy	6.34883	correct
Sonntag	group	6.34917	wrong

Table 3: First 24 word pairs collected by finding words with most similar context vectors in a greedy fashion.

2.4 Preserving Word Similarity

Intuitively it is obvious that pairs of words that are similar in one language should have translations that are similar in the other language. For instance, Wednesday is similar to Thursday as Mittwoch is similar to Donnerstag. Or: dog is similar to cat in English, as Hund is similar to Katze in German.

The challenge is now to come up with a quantifiable measurement of word similarity. One strategy is to define two words as similar if they occur in a similar context. Clearly, this is the case for Wednesday and Thursday, as well as for dog and cat.

Exactly this similarity measurement is used in the work by Diab and Finch [2000]. Their approach to constructing and comparing context vectors differs significantly from methods discussed in the previous section.

For each word in the lexicon, the context vector consists of co-occurrence counts in respect to 150 so-called **peripheral tokens**, basically the most frequent words. These counts are collected for each position in a 4-word window around the word in focus. This results in a 600-dimensional vector.

Instead of comparing these co-occurrence counts directly, the **Spearman rank order correlation** is applied: For each position the tokens are compared in frequency and the frequency count is replaced by the frequency rank – the most frequent token count is replaced by 1, the least frequent by n = 150. The similarity of two context vectors $a = (a_i)$ and $b = (b_i)$ is then defined by:[3]

$R(a, b) = 1 - \frac{6 \sum (a_i - b_i)^2}{4n(n^2-1)}$

The result of all this is a matrix with similarity scores between all German words, and second one with similarity scores between all English words. Such matrices could also be constructed using the definitions of context we reviewed in the previous section. The important point here is that we have generated a similarity matrix, which we will use now to find new translation word pairs.

Again, as in the previous Section 2.3, we assume that we will already have a seed lexicon. For a new word we can look up its similarity scores to the seed words, thus creating a **similarity vector**. Such a vector can be translated into the other language – recall that dimensions of the vector are the similarity scores to seed words, for which we already have translations. The translated vector can be compared to other vectors in the second language.

As before, we search greedily for the best matching similarity vectors and add the corresponding words to the lexicon.

2.5 Word Frequency

Finally, another simple clue is the observation that in comparable corpora, the same concepts should be used with similar frequencies. Even if the most frequent word in the German corpus is not necessarily the translation of the most frequent English word, it should also be very frequent.

Table 4 illustrates the situation with our corpora. It contains the top 10 German and English words, together with the frequency ranks of their best translations. For both languages, 4 of the 10 words have translations that also rank in the top 10.

Clearly, simply aligning the nth frequent German word with the nth frequent English word is not a viable strategy. In our case, this is additionally hampered by the different orientation of the news sources. The frequent financial terms in the English WSJ corpus (stock, bank, sales, etc.) are rather rare in the German corpus.

For most words, especially for more comparable corpora, there is a considerable correlation between the frequency of a word and its translation. Our frequency measurement is defined as ratio of the word frequencies, normalized by the corpus sizes.

3 Experiments

This section provides more detail on the experiments we have carried out to test the methods just outlined.

[3] In the given formula we fixed two mistakes of the original presentation [Diab and Finch, 2000]: The square of the differences is used, and the denominator contains the additional factor 4, since essentially 4 150-word vectors are compared.

Rank	German	English	Rank
1	Jahr	year	3
2	Land	country	112
3	Regierung	government	18
4	Prozent	percent	1
5	Präsident	president	8
6	Staat	state	24
7	Million	million	22
8	Angabe	statement	335
9	Mittwoch	wednesday	298
10	USA	us	5
4	Prozent	percent	1
308	Herr	mr	2
1	Jahr	year	3
72	Unternehmen	company	4
10	USA	us	5
58	Markt	market	6
150	Aktie	stock	7
5	Präsident	president	8
52	Bank	bank	9
119	Umsatz	sales	10

Table 4: The frequency ranks of the most frequent German and English words and their translations.

3.1 Evaluation measurements

We are trying to build a one-to-one German-English translation lexicon for the use in a machine translation system.

To evaluate this performance we use two different measurements: Firstly, we record how many correct word-pairs we have constructed. This is done by checking the generated word-pairs against an existing bilingual lexicon.[4] In essence, we try to recreate this lexicon, which contains 9,206 distinct German and 10,645 distinct English nouns and 19,782 lexicon entries.

For a machine translation system, it is often more important to get more frequently used words right than obscure ones. Thus, our second evaluation measurement tests the word translations proposed by the acquired lexicon against the actual word-level translations in a 5,000 sentence aligned parallel corpus.[5]

The starting point to extending the lexicon is the seed lexicon of identically spelled words, as described in Section 2.1. It consists of 1339 entries, of which are (88.9%) correct according to the existing bilingual lexicon. Due to computational constraints,[6] we focus on the additional mapping of only 1,000 German and English words.

These 1,000 words are chosen from the 1,000 most frequent lexicon entries in the dictionary, without duplications of words. This frequency is defined by the sum of two word frequencies of the words in the entry, as found in the monolingual corpora. We did not collect statistics of the actual use of lexical entries in, say, a parallel corpus.

In a different experimental set-up we also simply tried to match the 1,000 most frequent German words with the 1,000 most frequent English words. The results do not differ significantly.

3.2 Greedy extension

Each of the four clues described in the Sections 2.2 to 2.5 provide a matching score between a German and an English word. The likelihood of these two words being actual translations of each other should correlate to these scores.

There are many ways one could search for the best set of lexicon entries based on these scores. We could perform an exhaustive search: construct all possible mappings and find the highest combined score of all entries. Since there are $O(n!)$ possible mappings, a brute force approach to this is practically impossible.

We therefore employed a greedy search: First we search for the highest score for any word pair. We add this word pair to the lexicon, and drop word pairs that include either the German and English word from further search. Again, we search for the highest score and add the corresponding word pair, drop these words from further search, and so on. This is done iteratively, until all words are used up.

Tables 2 and 3 illustrate this process for the spelling and context similarity clues, when applied separately.

[4] extracted from LEO, `http://dict.leo.org/`

[5] extracted from the German radio news corpus de-news, `http://www.mathematik.uni-ulm.de/de-news/`

[6] For matching 1,000 words, the described algorithms run up to 3 days. Since the complexity of these algorithms is $O(n^2)$ in regard to the number of words, a full run on 10,000 would take almost a year. Of course, this may be alleviated by more efficient implementation and parallelization.

3.3 Results

The results are summarized in Table 5. Recall that for each word that we are trying to map to the other language, a thousand possible target words exist, but only one is correct. The baseline for this task, choosing words at random, results on average in only 1 correct mapping in the entire lexicon. A perfect lexicon, of course, contains 1000 correct entries.

The starting point for the corpus score is the 15.8% that are already achieved with the seed lexicon from Section 2.1. In an experiment where we identified the best lexical entries using a very large parallel corpus, we could achieve 89% accuracy on this test corpus.

Clues	Entries	Corpus
Identical Words (1339 Seed)	-	15.8%
Spelling	140	25.4%
Context	107	31.9%
Preserving Similarity	2	15.8%
Frequency	2	17.0%
Spelling+Context	185	38.6%
Spelling+Frequency	151	27.4%
Spelling+Context+Similarity	186	39.0%
All clues	186	39.0%

Table 5: Overview of results. We evaluate how many correct lexicon entries where added (Entries), and how well the resulting translation lexicon performs compared to the actual word-level translations in a parallel corpus (Corpus). For all experiments the starting point was the seed lexicon of 1339 identical spelled words described in Section 2.1. which achieve 15.8% Corpus score.

Taken alone, both the context and spelling clues learn over a hundred lexicon entries correctly. The similarity and frequency clues, however, seem to be too imprecise to pinpoint the search to the correct translations.

A closer look of the spelling and context scores reveals that while the spelling clue allows to learn more correct lexicon entries (140 opposed to 107), the context clue does better with the more frequently used lexicon entries, as found in the test corpus (accuracy of 31.9% opposed to 25.4%).

3.4 Combining Clues

Combining different clues is quite simple: We can simply add up the matching scores. The scores can be weighted. Initially we simply weighted all clues equally. We then changed the weights to see, if we can obtain better results. We found that there is generally a broad range of weights that result in similar performance.

When using the spelling clue in combination with others, we found it useful to define a cutoff. If two words agree in 30% of their letters this is generally as bad as if they do not agree in any – the agreements are purely coincidental. Therefore we counted all spelling scores below 0.3 as 0.3.

Combining the context and the spelling clues yields a significantly better result than using each clue by itself. A total of 185 correct lexical entries are learned with a corpus score of 38.6%.

Adding in the other scores, however, does not seem to be beneficial: only adding the frequency clue to the spelling clue provides some improvement. In all other cases, these scores are not helpful.

Besides this linear combination of scores from the different clues, more sophisticated methods may be possible [Koehn, 2002].

4 Conclusions

We have attempted to learn a one-to-one translation lexicon purely from unrelated monolingual corpora. Using identically spelled words proved to be a good starting point. Beyond this, we examined four different clues. Two of them, matching similar spelled words and words with the same context, helped us to learn a significant number of additional correct lexical entries.

Our experiments have been restricted to nouns. Verbs, adjectives, adverbs and other part of speech may be tackled in a similar way. They might also provide useful context information that is beneficial to building a noun lexicon.

These methods may be also useful given a different starting point: 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 described methods.

References

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. (1999). Statistical machine translation. Technical report, John Hopkins University Summer Workshop `http://www.clsp.jhu.edu/ws99/projects/mt/`.

Brown, P., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Rossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85.

Diab, M. and Finch, S. (2000). A statistical word-level translation model for comparable corpora. In *Proceedings of the Conference on Content-based multimedia information access (RIAO)*.

Fung, P. and Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of ACL 36*, pages 414–420.

Jones, K. S. (1979). Experiments in relevance weighting of search terms. In *Information Processing and Management*, pages 133–144.

Koehn, P. (2002). Combining multiclass maximum entropy classifiers with neural network voting. In *Proceedings of PorTAL*.

Koehn, P. and Knight, K. (2000). Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proceedings of AAAI*.

Koehn, P. and Knight, K. (2001). Knowledge sources for word-level translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Mann, G. S. and Yarowski, D. (2001). Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL*, pages 151–158.

Melamed, D. (1995). Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Third Workshop on Very Large Corpora*.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of ACL 33*, pages 320–322.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL 37*, pages 519–526.

Tiedemann, J. (1999). Automatic construction of weighted string similarity measures. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.