Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm

Philipp Koehn and Kevin Knight
koehn@isi.edu, knight@isi.edu
Information Science Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292

Abstract

Selecting the right word translation among several options in the lexicon is a core problem for machine translation. We present a novel approach to this problem that can be trained using only unrelated monolingual corpora and a lexicon. By estimating word translation probabilities using the EM algorithm, we extend upon target language modeling. We construct a word translation model for 3830 German and 6147 English noun tokens, with very promising results.

1. Introduction

Selecting the right word translation among several options in the lexicon is a core problem for machine translation. The problem is related to word sense disambiguation, which tries to determine the correct sense for a word occurrence (e.g. *river bank* vs. *money bank*).

While the definition of word sense is a tricky issue, the picture is much clearer in translation. If we observe human translators, we can collect up the different ways in which a German word is usually translated into English. In some contexts, certain translations will be more appropriate than others. Determining the sense of a word, as opposed to its translation, is a more subjective enterprise — different experts tend to divide and sub-divide word senses differently. Of course, word sense disambiguation and word-level translation are related. If one cannot determine whether an instance of the word *bank* refers to a river or a financial institution, it is unlikely that one will be able to translate the word accurately into Japanese, and vice versa.

In some ways, word-level translation is easier than word-sense disambiguation. For example, WordNet [Miller et al., 1993] breaks the English word *interest* down into 5 senses. But 3 of these senses all translate to the German word *Interesse* [Resnik and Yarowsky, 1997], so to translate the word correctly in most cases, it may not be necessary to distinguish between these 3 senses. In other ways, word-level translation is harder. Human translators select word translations that accurately describe the source meaning, but they also want to generate fluent target language output. That means a certain word translation may be preferred if it fits in well with other word translations. Also, the target language may have finer sense distinctions than can be foreseen in the source language. For instance the English word *river* translates as *fleuve* in French when the river flows into the ocean, and otherwise as *rivière* [Ide and Véronis, 1998].

We propose a novel framework for selecting the right translation word in a given sentence context. Using two completely unrelated monolingual corpora and a lexicon, we construct a word translation model for 3830 German and 6147 English noun tokens, with very promising results.

Our method is completely unsupervised: it is not necessary that the two corpora can be aligned in any way. Such monolingual corpora are readily available for most languages, while parallel corpora rarely exist even for common language pairs. Also, no manual sense tagging or definition of senses are required. The corpora we used for the experiments in this paper are the Wall Street Journal (6,892,443 noun tokens) and German newswire (306,982 noun tokens). As lexicon we use the freely available online dictionary LEO[1].

For testing purposes we use parallel corpora (or bitexts), generated from the monthly bulletin of the European Central Bank (ECB[2]) and de-news[3], a daily German news service written by student volunteers. Note that we use the bitexts only for evaluation purposes; they are not required for the construction of the model.

2. Related Research

There has recently been increased interest in empirical word sense disambiguation methods. Most research is reported on supervised methods, which use sense-tagged corpora. A good quantitative comparison of various methods is given by Mooney [1996]. While good results can be achieved, acquiring sufficiently large labeled corpora is prohibitively expensive.

[1] http://www.leo.org/cgi-bin/dict-search
[2] http://www.ecb.int/
[3] http://www.isi.edu/ koehn/publications/de-news/

Impressive unsupervised-learning results rivaling supervised methods are also reported by Yarowsky [1995], who trains decision lists for binary sense disambiguation. His bootstrapping method is unsupervised except for the use of a seed definition, which can be obtained manually or from dictionary entries.

Yarowsky deals only with words with very distant senses such as *plant* (*living* vs. *factory*) or palm (*tree* vs. *hand*). It is not clear how well his method will work with words such as the German *Gebiet*, for which our lexicon lists as English translations the following words: *area, zone, district, realm, territory, field, region, department, clime*, and *tract*. Also, seeds for these fine distinctions cannot be easily obtained from dictionaries, and must be created manually.

Schütze [1998] also proposes an unsupervised method, which is in essence a clustering of different usages of a word. The obtained clusters relate to some degree with word senses. It is questionable, however, whether such a method could come up with the proper clusters for the French translations of the word *river*. Also, the mapping of clusters to certain translations requires manual input.

While both Yarowsky and Schütze minimize the amount of supervision, it is still tremendous in the face of thousands of ambiguous lexicon entries. Both report results only on very few examples (less than 20).

The idea of using a second language monolingual corpus for word sense disambiguation is exploited by Dagan and Itai [1994]. They use a target language model to find the correct word-level translation. We expand on this notion and achieve better results, as reported below.

Research in statistical machine translation [Brown et al., 1993] demonstrates that word-level translation models can be learned from large parallel corpora. While there is hope that such corpora are becoming increasingly available, there may never be enough data for each language pair and domain.

Finally, current commercial MT systems seem to rely on always choosing the best word translation, supported by a lexicon of frequent compounds (such as *interest rate*). While this is useful for some instances of the word-level translation problem, it also creates a huge knowledge acquisition bottleneck.

3. Translation Probabilities

We describe an approach that uses a monolingual corpus in the target language to estimate word translation probabilities. These take the form $p_w(f|e)$, the overall probability that the English word e will be translated as f, regardless of context[4]. Brown et al. [1991] show how to estimate $p_w(f|e)$ parameters from a bilingual corpus. Since the translation probabilities cannot be observed directly in non-parallel corpora, one simple idea is to use the frequencies of the translation words in the target language itself.

For instance, if we look at the English noun *question*, our dictionary lists three possible German translations: *Frage*, *Zweifel*, and *Anfrage*. We can obtain the following counts in our German news wire corpus.

count	translation	sense
241	Frage	*query*
47	Zweifel	*doubt*
44	Anfrage	*request*

So we can estimate that $p_w(Frage|question) = 241/332 = .725$, and so forth. This method often allows us to estimate reasonable translation probabilities. Armed with these translation probabilities, we can decide to always pick the most likely translation word, regardless of context. In testing this approach on the nouns in the evaluation bitexts, we achieve 68.5% word translation accuracy for the ECB and 74.4% for the de-news test set.

However, this simple method frequently fails badly, as for the English noun *interest*, for which we obtain the following counts:

count	translation	sense
187	Anteil	*share, stake*
151	Interesse	*curiosity*
113	Zins	*money paid for money*
66	Bedeutung	*importance*
60	Teilnahme	*participation*
30	Vorteil	*advantage*

Actually, the most common translation *Interesse* ranks only second, behind the very rare translation *Anteil*. This happens because the German word *Anteil* is also the translation of the frequent English words *share, quota, lot, rate, proportion* etc. Most of the occurrences of *Anteil* in the German corpus do not in fact relate to *interest*.

In order to get better translation probabilities estimates, we have to take into account which occurrences of the word translation actually relate to the source word in consideration, and not others. We will address this issue in Section 5.

4. Modeling Context

The simple method described in Section 3 makes no use of context, as it always selects the same translation for a word. One way of deciding among several word translation options is to use a language model of the target language. For example, the machine translation system Gazelle [Knight et al., 1995] uses a word bigram language model to choose among sentence translations. The idea is that the translation of one word will affect the translation of another.

To illustrate this method, consider translating the German compound *Unschuldsvermutung* into English. The ambiguity of *Vermutung* and of the syntactic form of compounds in English yields four different translations. We counted their frequencies in the

[4] We follow here the usual notation of translating a foreign word f to an English word e.

World Wide Web using the search engine Altavista (`http://www.altavista.com`).

count	translation
1	innocence assumption
165	assumption of innocence
24	innocence presumption
6669	presumption of innocence

Clearly, this suggests that the most idiomatic translation is *presumption of innocence*. Also note that the distinction *assumption* vs. *presumption* would not likely be made by a manual German sense tagger for *Vermutung*.

This approach is along the lines of the work by Dagan and Itai [1994], who also use a target language model to disambiguate word translations. They propose the use of syntactic relationships such as subject-verb, verb-object, adjective-noun to disambiguate word translations.

We focus in our experiments on nouns to simplify our experimental setup. This method can be easily extended to include word forms, but for now we strip the corpus of these. Then we collect counts of adjacent words in our reduced English corpus. These counts allow us to estimate language model probabilities $p_{LM}(e_2|e_1)$ that a certain noun e_2 follows a previously observed noun e_1. With the resulting language model we can compute probabilities for sequences of candidate word translations. This is done by

$$p_{LM}(e_1, ..., e_n) = p_{LM}(e_1)p_{LM}(e_2|e_1)...p_{LM}(e_n|e_{n-1})$$

Thus, we can pick the word translations that occur in higher scoring candidate sequences (or sentences) than others. For this, we add up all the scores of all sequences that contain the word translation, compare this sum against the sums for the competing translations, and pick the highest.

The advantage of such a model, in addition to being very simple, is that it can be applied to all the nouns we find in a text. Syntactic models, as used by Dagan and Itai [1994], are more restrictive. Still, nothing in the framework that we will describe in the following section prevents us from using their model.

When we apply language probabilities to our evaluation bitexts, we improve on the ECB corpus to 70.6% and on the de-news corpus to 76.6%.

5. Estimation from Unrelated Corpora

We now combine the notion of translation probabilities with the use of context. First, we generate an English noun bigram language model for our English corpus (the target language). Then we use the expectation maximization (EM) algorithm [Dempster et al., 1977] to estimate word-level translation probabilities.

Note that this approach is very similar to research in statistical machine translation [Brown et al., 1993]. There, sentence pairs are given and the word translation model is to be learned without knowing the word alignments. Here, the source sentence is given and the word translation model is to be learned without knowing the target sentence. This is feasible, because we use a lexicon to restrict the space of possible target sentences.

Outline — Consider the following sentence (translation: *Hans visits the bank counter at the end of the day*), annotated with the English noun translations. The correct translations are in bold type.

Hans besucht den *Bank*	*Schalter*	*am* *Ende*	*des* *Tages*
bench	**counter**	*bottom*	**day**
bank	*switch*	*finish*	
		end	
		ending	
		expiration	
		tail	

To compute probabilities for each candidate English noun sequence e_s, we first use Bayes rule:

$$p(e_s|f_s) = p(f_s)^{-1}p(e_s)p(f_s|e_s)$$

So, instead of using direct translation probabilities from German to English, we use a English language model $p_{LM}(e_s)$ and a translation model from English to German $p(f_s|e_s)$. The factor $p(f_s)$ can be discarded for the purpose of comparing different English noun sequences, since it is equal for all possibilities.

We now compute the remaining probabilities $p(e_s)p(f_s|e_s)$ using the language model p_{LM} and word translation probabilities p_w:

$$\begin{aligned} p(e_s)p(f_s|e_s) &= p_{LM}(e_1, ..., e_n)p_s(f_1, ..., f_n|e_1, .., e_n) \\ &\approx p_{LM}(e_1)p_{LM}(e_2|e_1)...p_{LM}(e_n|e_{n-1}) \cdot \\ &\quad p_w(f_1|e_1)...p_w(f_n|e_n) \end{aligned}$$

Estimation of Translation Probabilities — The translation probabilities p_w are initially set to an uniform distribution. The correct translation will have a higher probability, if it contains more frequent bigrams (*bank counter* vs. *bench switch*).

These noun sequence (or sentence) probabilities are normalized and then used to update the word translation probabilities. Intuitively, after finding the most probable translations of the sentence, we can collect counts for the word translations it contains. Since the English language model provides context information for the disambiguation of the German words, we hope to count only the appropriate occurrences.

In Figure 1 we give a more formal description of our use of the EM algorithm. Given a language model $p_{LM}(e)$ we wish to estimate the translation probabilities $p_w(f|e)$ that best explain the German corpus as a translation from English. The translation probabilities converge after 10 to 20 iterations of the EM algorithm.

This naive algorithm requires integration over c^n possible sentence translations (where c is the average number of translations for any given word). This is too much computation in practice. However, the forward-backward algorithm, which we have implemented, can

```
train language model for English p_LM
initialize word translation probabilities p_w uniformly
iterate
  set score(f|e) to 0 for all dictionary entries (f,e)
  for all German sentences f_s = (f_1, ..., f_n)
    for all possible English sentence translations e_s
      compute sentence transl. probability p_s(e_s|f_s)
        by p_w(f_1|e_1)p_w(f_2|e_2)...p_w(f_n|e_n)·
        ·p_LM(e_1)p_LM(e_2|e_1)...p_LM(e_n|e_{n-1})
    endfor
    normalize p_s(e_s|f_s) so their sum is 1
    for all sentence translations e_s
      for all words e_w in e_s
        add p_s(e_s|f_s) to score(f_w|e_w)
      endfor
    endfor
  endfor
  for all translation pairs (f_w, e_w)
    set p_w(f_w|e_w) to normalized score(f_w|e_w)
  endfor
enditerate
```

Figure 1: The EM Algorithm

accomplish the same in c^2n steps through the use of dynamic programming [Baum, 1972].

Application — With both language model and translation probabilities in place, we can now find the best word translations for a given German sentence f_s by using the Bayes rule

$$argmax_{e_s} p_s(e_s|f_s) = argmax_{e_s} p_{LM}(e_s) p_w(f_s|e_s)$$

We combine the language model $p_{LM}(e_s)$ with the use of translation probabilities $p_w(f_s|e_s)$ to search for the best translation $p_s(e_s|f_s)$. Again, this is done in c^2n steps.

6. Results

We now evaluate the generated translation probabilities. First, we look at the translation table for *interest*, as generated by our algorithm:

prob.	translation	sense
33.2%	Interesse	*curiosity*
27.7%	Zins	*money paid for money*
19.8%	Anteil	*share, stake*
12.6%	Teilnahme	*participation*
6.0%	Bedeutung	*importance*
0.5%	Vorteil	*advantage*

These numbers are much closer to a realistic distribution, as the most frequent translations *Interesse* and *Zins* come out on top. The use of the language model clearly helped to discount most instances of *Anteil* that do not translate to *interest*. Our method generated respective translation tables for all 6147 English nouns.

Another way to test the quality of the generated word translation probabilities is to use them to translate German words in context and compare the results against other methods.

For this, we use the ECB and de-news bitexts. After sentence aligning them we can use our lexicon to identify how the nouns in the text were translated. We then measure how accurate the methods match these word-translation pairs. Since sometimes more than one translation of a word may be fully acceptable, we cannot expect 100% accuracy on this task, but it is still a very good metric of the relative performance.

We compare our method (EM) against just using the language model of Section 4 (LM) and just relying of the most frequent translation word in the raw count, as in Section 3 (MF). We also report the performance of a commercial system on this task. Note that there is a slight bias against the commercial system in this evaluation, since we only consider word-translation pairs that are in the dictionary used by our methods.

corpus	commercial	MF	LM	EM
ECB	77.9%	68.5%	70.6%	80.5%
de-news	73.3%	74.4%	76.6%	78.2%

On both texts, our method clearly comes out ahead. The de-news bitext contains 5610 noun word-translation pairs in 2713 sentences, the ECB contains 693 word-translation pairs in 155 sentences. The larger improvement of our EM method over the benchmarks may lie in the fact that it suffices to get a few frequent word translations right.

The results suggest that we can improve substantially upon pure target language modeling, as done by Dagan and Itai [1994]. Although we currently use a different target language model, adding word translation probabilities clearly benefits performance.

7. Discussion

We introduced a completely unsupervised method to estimate translation probabilities. The required monolingual corpora are readily available for most cases. Although a bilingual lexicon is still required and its quality impacts the performance, this should not be a problem for most language pairs in question.

The method works on a large scale: We were able to apply it to a much bigger number of ambiguous words than related research on word sense disambiguation. Our current experimental setup is restricted to nouns, but it will extend to verbs, adjectives, prepositions, etc.

We may improve performance with larger corpora, a larger context window, use of context in the source language, or better language modeling, for instance by exploiting syntactic relations between words. We plan to address these directions in future research.

References

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3:1–8.

Brown, P. F., Della-Pietra, S., Della-Pietra, V., and Mercer, R. (1991). Word-sense disambiguation using statistical methods. In *Proceedings of ACL 29*.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–313.

Dagan, I. and Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelyhood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

Ide, N. and Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

Knight, K., Chander, I., Haines, M., Hatzivassiloglou, V., Hovy, E., Iida, M., Luk, S. K., Whitney, R., and Yamada, K. (1995). Filling knowledge gaps in a broad-coverage MT system. In *International Joint Conference on Artificial Intelligence*.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1993). Introduction to WordNet: An online lexical database. Technical Report CSL 43, Cognitive Science Laboratory Princeton University.

Mooney, R. (1996). Comparative experiments on disambiguation word senses: An illustration of bias in machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Resnik, P. and Yarowsky, D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *ACL 35, SIGLEX Workshop at ANLP*.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Lingustics*, 24(1):97–123.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of ACL 33*, pages 189–196.