Knowledge Sources for Word-Level Translation Models

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

We present various methods to train word-level translation models for statistical machine translation systems that use widely different knowledge sources ranging from parallel corpora and a bilingual lexicon to only monolingual corpora in two languages. Some novel methods are presented and previously published methods are reviewed. Also, a common evaluation metric enables the first quantitative comparison of these approaches.

1 Introduction

We are currently experiencing a new wave of research in statistical machine translation, based on the influential work of the IBM Candide project (Brown et al., 1990). Statistical machine translation uses models for word-level translation and models for reordering of words, which may be based on syntactical structure (Yamada and Knight, 2001). These models are typically trained on parallel corpora. This paper will focus on training word-level translation models.

Figure 1 illustrates the issues of word-level translation models: For each word in one language (say, *interest*), many possible translations may exist. In turn, each of these may have several translations back, and so on. The task of the lexical component of statistical machine translation systems is not only to find the possible word-level translations, but also to estimate probabilities of how likely each translation occurs.

The purpose of this paper is twofold:

- First, we evaluate the premise of statistical machine translation by analyzing parallel corpora and hand-crafted translations.
- Second, we examine how we can augment or replace parallel corpora with other knowledge sources: bilingual lexicons and monolingual corpora.

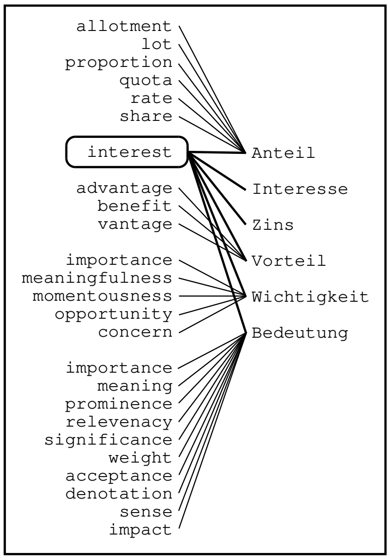

Figure 1: Multiple Translations per Word

In a large-scale experiment we assess the impact of a variety of resources, algorithms, and techniques on the task of deriving probabilistic dictionaries. These can be used easily in the context of both statistical and symbolic machine translation systems. We review and improve the state of the art with respect to algorithms that incorporate a probabilistic component into existing bilingual dictionaries and algorithms that derive such dictionaries from parallel and monolingual data. We assess on a German-English data set the performance of such algorithms and gauge the importance of each resource.

2 Relevance of Word-Level Translations

To get a better understanding of the problem, we performed a small-scale investigation of parallel text to measure the relevance of word-level translation. We examined two German-English aligned text sources, the monthly bulletin of the European Central Bank (ECB)[1] and the Minutes of the European Parliament (Europarl)[2]. A small excerpt is displayed in Figure 2.

Imagining the English text as a translation of the German, we examined what happened to each of the 443 German words during the assumed translation to English. We defined six categories for this purpose:

Translation in dictionary – The word is translated to a English word that could be found in an adequate bilingual dictionary.

Translation unusual – The translation word is translated to an English word, which is generally not a good translation, but valid in the given context. Example: *last session* (literal: *yesterdays session*)

POS changed in translation – The translation word is a literal translation, except that it is of a different part-of-speech. Example: *economic activity* (literal: *economy activity*)

Part of a phrase – The word is part of a phrase that as a whole is not translated literally. Examples: *are consistent* (literal: *stand in harmony*), *both ... and* (literal: *as well as*), *suggests* (literal: *gives indication to*)

Dropped for syntactic reasons – The original word is part of a syntactic construction that does not exist in the target language, for example some articles are dropped when translated from German to English.

Dropped otherwise – Words that are dropped, often for no clear reason at all. This may even slightly change the meaning of the sentence. Example: *in the euro [currency] area*

We found that only about 68% of the German words (also 68% of the German nouns) were translated to an English word that may be found in an adequate German-English bilingual

[1]`http://www.ecb.int/pub/period.htm`

[2]`http://www3.europarl.eu.int/`

German side of the parallel corpus

Die neuen Daten und Umfrageergebnisse, die seit Ende Juni 1999 vorliegen, stehen im Einklang mit den zuvor gehegten Erwartungen, daß die Wirtschaftstätigkeit im Euro-Währungsgebiet in der ersten Jahreshälfte 1999 zunächst nicht weiter zurückgegangen ist, sich dann stabilisiert hat und sich in der zweiten Jahreshälfte beschleunigen dürfte.
Die Wachstumsrate der Geldmengen- und Kreditaggregate bis einschließlich Juni 1999 unterstützt diese Beurteilung weitgehend, obwohl einige Aufwärtsrisiken für die künftige Preisstabilität nicht ausgeschlossen werden können.

English side of the parallel corpus

The data and surveys which have become available since end-June 1999 are consistent with earlier expectations, according to which economic activity in the euro area first ceased to decline and then stabilized in the first part of 1999 and should accelerate in the second part of the year.
The evolution of monetary and credit aggregates up to June 1999 broadly supports this assessment, although some upward risks to future price stability cannot be ruled out.

More literal translation of German side of the parallel corpus

The new data and survey results that have been available since the end of June 1999 are consistent with the previously held expectations that the activity of the economy in the Euro currency area initially did not further decline in the first half of 1999, then stabilized itself and should accelerate in the second half of the year.
The growth rate of money size and credit aggregate until June 1999 inclusively supports this assessment widely, although some upward risks for the future price stability can not be excluded.

Commercial MT Translation (Systran)

The new data and survey data, which are present since at the end of of June 1999, are in conformity with expectations preserved before that the economic activity in the Euro-currency area in the first yearly half 1999 did not continue to decrease first, then stabilized and in the second yearly half accelerate itself might.
The growth rate of the money supply and credit units to including June 1999 supports this evaluation to a large extent, although some upward risks for the future price stability cannot be excluded.

Figure 2: Part of the ECB corpus

	Official Translation						Literal Translation					
	ECB German		Europarl German		Europarl Port.		ECB German		Europarl German		Europarl Port.	
	all	n.	all	n.	all	n.	all	n.	all	n.	all	n.
Translation in dictionary	70%	65%	66%	71%	61%	77%	91%	94%	90%	95%	93%	98%
Translation unusual	4%	13%	1%	2%	1%	2%	1%	2%	0%	0%	0%	0%
Part of idiomatic phrase	10%	2%	20%	16%	20%	11%	3%	2%	6%	3%	2%	2%
Dropped for syn. reasons	7%	0%	5%	0%	8%	0%	3%	0%	3%	0%	5%	0%
Dropped otherwise	3%	5%	5%	8%	9%	7%	1%	2%	0%	0%	2%	0%
POS changed	6%	15%	3%	3%	2%	1%	1%	0%	1%	2%	0%	0%

Figure 3: Breakdown of what happens to words during translation – both for official translations found in a parallel corpus and for a more literal translations which is the target of machine translation systems. Analysis for all words and nouns (n.) on three corpora: German-English European Central Bank bulletin, German-English and Portuguese-English European Parliament Minutes.

dictionary. A detailed break-down is provided in Figure 3.

There are many reasons for the low number of literal translations. It is quite frequent that a word gets translated in a way that is only justifiable in this particular context. Some words are simply dropped or change their part-of speech, e.g., a noun may be turned into an adjective. Sometimes this seems arbitrary, but in many cases it seems motivated by a more fluent text in the target language.

A serious problem for machine translation systems that rely on word-level translation models are phrases that cannot be translated literally. This ranges from idiomatic expressions such as *den Löffel abgeben* (literal: *to give up the spoon*, meaning: *to die*) to constructions whose translations are understandable, but just do not sound right, e.g. *im Einklang stehen* (literal: *to stand in harmony*, meaning: *to be consistent*).

Another issue are words that are dropped, changed, or added due to their idiosyncratic syntactic nature in a particular language. Hopefully, a more syntactic approach to translation will be able to deal with this.

Of course, there are many ways to correctly translate a text. In the second stage of our investigation, we emulate the behavior that can be expected from an MT system: a more literal translation. We tried to translate as many words as possible with translations that may be found in a bilingual lexicon.

We can achieve a much higher percentage of literal word translations this way, as detailed in Figure 3. About 90% of all words and about 95% of the nouns can be translated using terms from a dictionary. The lower number for all words is mostly due to syntactic reasons, e.g. determiners that are used in German, but not in English. For open class words such as nouns, the biggest remaining problem are phrases that cannot be translated literally.

We carried out the same analysis on Portuguese-English data with similar results.

The high accuracy of word-by-word translation suggests that we will be able to address the core of the machine translation problem with an approach that basically does word-level translation, occasional dropping and inserting, and reordering of words. This is what current statistical machine translation projects are shooting for.

3 Resources

A large number of resources have recently become available for machine translation research, both corpora and tools. We chose German-English translation for the experiments in this paper. The following details the resources used.

The LEO bilingual German-English dictionary[3] is an ongoing volunteer effort. While it is still not finished, it already provides an outstanding resource with over 230,000 entries. Bilingual dictionaries may not be easily available for other language pairs, especially for low-density languages. Also, these dictionaries are general-purpose and may not be suited for specialized domains such as medicine, law, or finance.

Parallel corpora are slowly becoming more available. Currently, they tend to derive from

[3] `http://dict.leo.org/`

government sources, such as parliament proceedings or laws, which may not be suitable for other domains. For our experiments we decided not to use the Europarl and ECB corpora mentioned in the previous Section, but the more general DE-NEWS corpus. It contains transcript and translation of German news reports from 1996-2000. The size of the corpus is 50,000 sentences pairs (one million words per language), which could be considered medium-sized – in comparison, the Canadian Hansard has about two million sentence pairs with forty million words per language. We sentence-aligned the corpus by an implementation of an algorithm proposed by Gale and Church (1993), with manual post-editing.

Fortunately, the evolution of the World Wide Web and widespread use of digital publishing has created a wealth of monolingual corpora. We use the Wall Street Journal (WSJ) and German news wire (DPA) corpora, which are both available through the Linguistic Data Consortium (LDC)[4]. Both consist of over one million sentences and about twenty million words each.

Also, recent research in natural language processing has equipped us with many useful tools. For instance, the performance of part-of-speech taggers is currently considered on-par with humans. For our experiments, we also used Morphy as morphological analyzer and part of speech tagger for German, and Eric Brill's part of speech tagger for English (Brill, 1994).

In summary, we used the following resources:

- Morphy, a German POS tagger and morphological analyzer[5]
- Eric Brill's English POS tagger[6]
- the DE-NEWS German-English corpus[7]
- the LEO German-English dictionary
- the Wall Street Journal corpus (English)
- the DPA news wire corpus (German)

Other knowledge sources that may be useful are natural language parsers or ontologies such as WordNet (Miller et al., 1993).

4 Experimental Setup

For this paper we want to investigate the role of different knowledge sources for the training of word-level translation models. We evaluate the methods by the accuracy of the suggested word-level translations with respect to a reference parallel corpus.

When translating the limited number of closed class words such as articles, syntactical issues and language ideosyncracies play a big role. The emphasis of the lexical component of a machine translation system is to perform well for the much larger number of open class words – nouns, verbs, adjectives and adverbs. In this work, we decided to focus on nouns.

We examined the behavior of 9,206 German and 10,645 English distinct nouns. Some of these have unique, while most have multiple translations. The lexicon consists of 19,782 word pairs. So, on average, there are two entries per word.

5,000 sentence pairs of the DE-NEWS corpus are used as evaluation set. We identified word translations using a bilingual lexicon. The task for the following methods is to find the same English translation for a German word, given the German, but not the English part of the evaluation corpus.

Given nouns in the German sentence, the lexicon constrains the possible matching English nouns sufficiently in nearly all cases. It is very rare that two or more nouns in the German sentence may map to the same word in the English sentence.

If there is no English translation for a German word within the lexicon, we do not place it in the evaluation set. While this excludes a considerable portion of the German words, we do not view this as a weakness of the evaluation metric. As we pointed out in Section 2, we are looking for a more literal translation as the goal of a machine translation system than the parallel corpus provides.

Of course, a method that does not use the lexicon may find good translations outside the dictionary and may get punished by this metric. For our data, however, this did not constitute a significant problem.

Another issue is that in some cases two translations might be perfectly fitting. A method that picks the translation that is not in the evaluation set is unfairly discounted. But since this is the same for all methods, it should have no effect on the comparison of the methods. Still,

[4] http://www.ldc.upenn.edu/

[5] http://www-psycho.uni-paderborn.de/lezius/

[6] http://www.cs.jhu.edu/ brill/

[7] http://www.isi.edu/ koehn/publications/de-news/

it does mean that 100% accuracy according to the metric may not be achievable.

We focused the following investigation on nouns. We can identify the nouns in the corpus using the part-of-speech taggers. These tools allow us to reduce the found surface forms to word stems. In addition, the German Morphy also allows us to split up compounds such as *Bundesverteidigungshaushalt* (*federal defense budget*).

5 Using Parallel Corpus and Lexicon

5.1 Background

The best results can be achieved provided both a parallel corpus and a bilingual dictionary. As with our evaluation corpus, we can use the bilingual lexicon to extract word-level noun translation pairs from the parallel corpus.

Having word-level translations in context, we can use supervised word sense disambiguation methods, which have been extensively studied. For instance, Mooney (1996) provides a good comparison these methods. See also the overview by Ide and Véronis (1998).

5.2 Experiment

In the method used here, we extracted the following context features for each noun occurrence:

- Up to three words of local context around the target word, using part-of-speech tags as back-off.
- Any open-class word (noun, verb, adjective, adverb) in the same sentence
- Any open-class word in the same document

These features allow us to train a decision list as described by Yarowsky (1994).

5.3 Results

The resulting classifier finds the correct word-level translation in our test data with 89.5% accuracy. The baseline method for this data, which is to choose always the most frequent word-level translation, as found in the training data, however, performs only slightly worse (88.9%).

Let us already point out at this point the significance of these results: Without any word sense disambiguation we could achieve almost 90% accuracy. None of the following experiments will reach this performance. So, at least in the framework of these experiments, the main problem is still finding the overall best translation for a word, not the advanced task of finding the right translation in a given context.

6 Using Parallel and Monolingual Corpora and Lexicon

6.1 Background

Yarowsky (1995) proposes a bootstrapping scheme that uses a initial decision list trained on supervised data as a starting point. By labeling new word occurrences in a monolingual corpus, he was able to collect more evidence that enable the construction of a superior decision list.

6.2 Experiment

We can easily apply this idea to our problem: We already trained a decision list for word-level translations. Using this decision list on a German monolingual corpus and additional clues such as "one sense per discourse", we can label more occurrences of the German words with English translations. This, in turn, provides a larger training set to retrain the decision list.

6.3 Results

Applying this idea to our data, however, was not successful. Nearly all ambiguous German words have a strong majority translation. After applying the decision list to monolingual data, an even larger portion of the occurrences is labeled with the majority translation. The algorithm quickly converges to a decision list that always predicts the majority case.

Apparently the initial decision lists are not good enough to find strong context features for the minority translations. This is underscored by the accuracy just above the baseline. The original training set with at most a few hundred occurrences for each German word does not seem to be big enough. It might be more successful, if a larger parallel corpus is available.

7 Using only Parallel Corpus

7.1 Background

The Candide machine translation approach (Brown et al., 1990) is based on the noisy channel model. It takes the view that the foreign language sentence is just distorted English that has been corrupted by a translation process. It can be decoded by using the Bayes rule:

$$argmax_e p(e|g) = argmax_e p(g|e)p(e)$$

So, instead of directly collecting statistics on how likely an English translation is given German input, we use statistics on how likely a German translation is given English input and an English language model $p(e)$. We will not go into the details of the model. The important point is that word-level translation probabilities $p(g|e)$ are an important element in the model.

These word-level translation probabilities are collected from a parallel corpus. During training, the most likely word alignments are determined. These word alignments are then the basis of the word-level probability estimates $p(g|e)$. We simply count the number of occurrences of the English word e, and how often it is aligned to the German word g.

Since the alignment process is not limited by a bilingual lexicon, it is hard to find good word alignments for rare words in the corpus.

For alternative methods for building translation models from parallel texts, see the work by Melamed (2000).

7.2 Results

After training a machine translation system using the freely available Giza toolkit[8] (Al-Onaizan et al., 1999) (Model 3), we can translate the German side of the evaluation set using a stack decoder (Germann et al., 2001). It turns out that 76.9% of the German nouns are translated correctly, opposed to 89.5% with the method in Section 5.

Figure 4 illustrates clearly where the lack of a bilingual lexicon hurts the most: When only a few training examples for a given word are contained in the training corpus. Only when at least about 50 occurrences can be found in the parallel corpus, the performance closes in with the method that uses a bilingual lexicon. But if fails occasionally even for the two words with at least 1024 instances that where both unambiguous in the dictionary (*Kanzler/chancellor* and *Jahr/year*).

For the method in the previous sections we eliminated all German noun occurrences when none of the word translations provided by the lexicon in the corresponding English sentence. As we have seen in Section 2, we can expect this to happen in thirty percent of the cases. However, the method in this section still tries

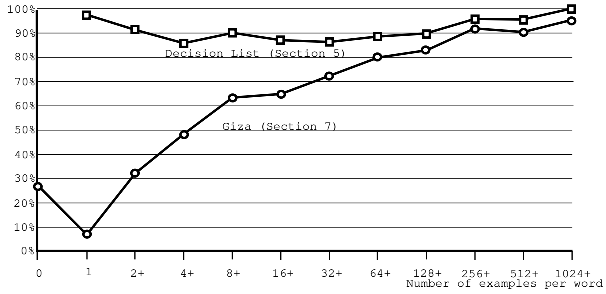

Figure 4: Accuracy for Decision List (Parallel Corpus, Lexicon) and Giza (only Parallel Corpus)

to find a word alignment in these cases. This results in noisy data.

Two minor points are worth mentioning: Given no evidence for a German word (0 training examples), the word is not translated but simply transfered verbatim to the English output. For these cases the performance of the system is a surprising 27%. This is due to words such as *Nation* or *Email* that are identical in German and English. Note that this is not generally the case: If we present all German words verbatim as English output, the score is just 11.9%.

Secondly, the machine translation system is able to find one perfectly good translation (*Arbeit/employment*) that was missing in the bilingual lexicon. This points out one weakness of an approach that relies too much on the lexicon.

8 Using Monolingual Corpora and Lexicon

8.1 Background

Parallel corpora are generally hard to come by. The corpora currently used in research are often parliament proceedings (the Canadian Hansard, the European Parliament), law texts (e.g. Chinese-English from Hongkong), or other government sources. Works of literature (Orwell's 1984) or the Bible have also been used, although this is commonly hampered by either copyright concerns or outdated language use. There have been efforts to discover parallel texts on the World Wide Web (Resnik, 1999) or simply take the output of existing machine translation systems (Diab, 2000). Parallel corpora could also be especially created for the train-

[8] `http://www.clsp.jhu.edu/ws99/projects/mt/`

ing of machine translation systems, but this is a costly option – professional translation rates are roughly between 5 and 20 US cents per word.

But ultimately we have to face the fact that people do not naturally produce the same text in multiple language. Therefore, parallel corpora will always be a limited resource, often from unsuitable domains. The previous section highlighted that a word has to occur a sufficient number of times to be able to learn reasonable translation models. But even in the forty million sentence Hansard Corpus common words such as *directory*, *empathy*, *reflex*, *ant*, *filth*, *gangster*, and *fake* occur only once.

On the other hand, we can surely assume, that we will have a large monolingual corpus available in a language for which we want to build a machine translation system. After all, if this would not be the case, what would be the purpose of such a system? We have also good reason to believe that the information technology revolution will bring large monolingual text resources forward. The World Wide Wide alone currently consists of over one billion documents. According to the search engine Google[9], the words above occur extremely often – *directory* 42 million times, *empathy* 180,000 times, *reflex* 372,000 times, *ant* 859,000 times, and so on.

We are optimistic that much can be learned from this vast amount of data. This section will discuss methods how to learn translation models from monolingual data which require a bilingual lexicon. The next section will drop this requirement.

8.2 Experiment

If we have a corpus in the target language, we can apply two simple ideas:

Firstly, we could always choose the word in the lexicon that occurs most frequently in the target language corpus. This simple principle shows surprisingly good results.

Secondly, we can build a language model and choose the most likely sequence of words in the target language. This allows the use of context clues. Research in generation, such as the Nitrogen generation system (Langkilde and Knight, 1998) demonstrate the effectiveness of this method.

A more sophisticated approach is proposed by Koehn and Knight (2000). It uses a lexicon, a corpus in the target language, and a comparable corpus in the source language. The approach views the corpus in the source language (German) as actually being an English corpus, corrupted by a noisy channel.

Given word-level translation probabilities and a language model we can determine for each German word the most likely English word in its place. Also, given such a corpus of German-English word pairs, we can easily estimate word-level translation probabilities.

This set-up constitutes a chicken and egg problem: On the one hand we do not have word-level translation probabilities to estimate the best English word matches. On the other hand, we do not have these German-English word pairs to estimate word-level translation probabilities.

Fortunately this problem can be addressed using the Expectation Maximization (EM) algorithm. The algorithm alternatively scores the possible English words for each German word (the expectation step) and estimates translation probabilities based on this (the maximization step) until convergence.

The resulting word-level translation probabilities and the language model can be used combined with the target language model (trained on the target corpus) in the usual statistical machine translation setup to translate unseen German text. As before, we apply this method only to the nouns in the text.

8.3 Results

Using frequency counts results in 75.3% accuracy, the use of a language model 77.3%, and the EM method raises this to 79.0%. Still, when both a parallel corpus and a bilingual lexicon are used, the accuracy of noun translations is about ten percent higher (88.9%). However, these numbers are on par with the results for training from only a parallel corpus (76.9%). Thus, for our set of resources, a parallel corpus can be replaced with monolingual corpora and a bilingual lexicon.

9 Using only Monolingual Corpora

9.1 Background

We already argued that the most easily available knowledge source is monolingual text.

[9] `http://www.google.com/`

Some ideas have been investigated in the attempt to construct lexicons using only monolingual corpora. All these approaches try to create a one-to-one or one-to-many mapping. While this is not realistic (recall Figure 1), it is a good starting point.

Similarities between a word and its translation make this a feasible endeavor. Rapp (1995, 1999) bases his work on the notion that words that co-occur frequently in one language have translations that also co-occur frequently in another language. He uses this properties to fill gaps in an existing lexicon. Work by Fung (1995); Fung and Yee (1998) is based on the same principal: This allows her to add novel terms to a lexicon.

Diab and Finch (2000) make another interesting observation: Words that have a certain similarity in one language (say *dog* and *cat*) have translations that are similar in the other language. They measure similarity as occurring in a similar context (say *her new X is cute*). Similarity is measured by a context vector using 4-word window.

But also simpler clues may provide useful information: Words that are very frequent in one language have translations that may also be frequent in a comparable corpus in another language.

Also, some words have similar spelling across languages. This may be due to same roots (English: *mother*, German: *Mutter*, Portuguese: *mãe*, Spanish: *madre*), or cultural exchange (English: *computer*, German: *Computer*, Portuguese: *computador*, Japanese: *konpyuta*).

Rapp (1999) points out the need for a seed for lexicon construction from monolingual corpora. Only the algorithm by Diab and Finch (2000) does not require a seed. However, although we successfully duplicated their work when applied to two comparable English corpora, the method failed to produce a useful lexicon for our comparable German and English corpora.

9.2 Experiments

Without a lexicon to start with, we could collect an initial lexicon from a small parallel corpus. But to stick to the spirit of this section, we used a different seed: words that have the exactly the same spelling in German and English.

From the 9,206 German nouns in our bilingual lexicon, we could find 1,016 words that occur also in English (such as *nation*, *computer*, *email*). The assumption that these words are in fact translations of each other is accurate for 88% of the words (exceptions are mostly shorter words, for example *ton, fee, kind, rat, art, rock, boot, gang, plane, taste, hut*). When relying solely on this same-word seed lexicon, we can achieve 11.9% accuracy on our evaluation metric.

Then we used four different methods to extend the lexicon, which exploit the fact that a word and its translation have similar:

- context (Rapp, 1999)
- spelling
- relationship to other words (Diab and Finch, 2000)
- frequency

9.3 Results

For computational reasons we focus on the 1000 most frequent German and English words according to the monolingual corpora. Only the context and the spelling property helped us to extend the lexicon. When adding lexicon entries based on similar spelling, we achieved 25.4% accuracy on our evaluation metric, with the context property 31.9%. When both properties are combined, we achieve 38.6%. This is quite significant, since the baseline – mapping words at random – is no better than the original score of 11.9% for identical words.

10 Conclusions

We established that a word-level translation model is a core element of machine translation systems – 95% of nouns can be translated within a conventional bilingual lexicon.

These models are usually trained on parallel texts. We investigated various methods to augment and replace the need for parallel corpora with monolingual corpora and bilingual lexicons. A common evaluation metric enabled a first quantitative comparison, as summarized in Figure 5.

A bilingual lexicon provides a clear benefit: While training on a parallel corpus alone we achieved 76.9% accuracy, the lexicon boosted this to 89.5%. Also, using both monolingual data and a lexicon allowed us to replace the parallel corpus and get similar performance (79.0%). Finally we showed how to acquire a

Knowledge sources	Method	Acc.
Parallel corpus, lexicon	most frequent	88.9%
Parallel corpus, lexicon	decision list	89.5%
Parallel corpus	Giza	76.9%
Monoling. corpus, lexicon	most frequent	75.3%
Monoling. corpus, lexicon	language model	77.3%
Monoling. corpus, lexicon	EM	79.0%
Monoling. corpus	identical	11.9%
Monoling. corpus	spelling	25.4%
Monoling. corpus	context	31.9%
Monoling. corpus	spell.+context	38.6%

Figure 5: Results for methods from Section 5-9

translation model purely from monolingual corpora.

The heart of the problem still seems to be finding the overall best translation for all the words, rather than advanced word sense disambiguation task of finding the right translation in a given context. This is even more true for low density languages, where less resources are available.

Clearly, there is still much room for improvement – many ideas that we touched upon here require further investigation.

Acknowledgment

This work was supported by DARPA-ITO grant N66001-00-1-9814.

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/`.

Brill, E. (1994). Some advances in rule-based part of speech tagging. In *Proceedings of AAAI*.

Brown, P., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., 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. (2000). An unsupervised method for multilingual word sense tagging using parallel corpora: A preliminary investigation. In *SIGLEX Workshop on Word Senses and Multi-Linguality*, pages 1–9.

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

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.

Gale, W. and Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational Lingustics*, 19(1).

Germann, U., Jahr, M., Knight, K., Marcu, D., and Yamada, K. (2001). Fast decoding and optimal decoding for machine translation. In *Proceedings of ACL 39*.

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.

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

Langkilde, I. and Knight, K. (1998). Generaton that exploits corpus-based statistical knowledge. In *Proceedings of ACL 36*, pages 704–710.

Melamed, I. D. (2000). Models of translational equivalence among words. *Computational Lingustics*, 26(2):221–249.

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*.

Rapp, R. (1995). Identifying word translations in nonparallel 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.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of ACL 37*, pages 527–534.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of ACL 39*.

Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *Proceedings of ACL 32*.

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