Word Translation Disambiguation without Parallel Texts[*]

Erwin Marsi André Lynum Lars Bungum Björn Gambäck
Department of Computer and Information Science
NTNU, Norwegian University of Science and Technology
Sem Sælands vei 7–9, NO–7491 Trondheim, Norway
{emarsi,andrely,larsbun,gamback}@idi.ntnu.no

Abstract

Word Translation Disambiguation means to select the best translation(s) given a source word in context and a set of target candidates. Two approaches to determining similarity between input and sample context are presented, using n-gram and vector space models with huge annotated monolingual corpora as main knowledge source, rather than relying on large parallel corpora. Experiments on SemEval's Cross-Lingual Word Sense Disambiguation task (2010 English→German part) show some models on average surpassing the baselines, suggesting that translation disambiguation without parallel texts is feasible.

Index Terms: word sense disambiguation, vector space models, n-gram language models

1 Introduction

One of the challenges in translating a word is that, according to a translation dictionary or some other translation model, a source language word normally has several translations in the target language. For instance, the English word *plant* may be translated as the German word *Fabrik* in the context of industry, but as *Pflanze* in the context of nature. Hence contextual information is required to resolve ambiguities in word translation. This task is known as Word Translation Disambiguation (WTD).

The currently predominant paradigm for data-driven machine translation is phrase-based statistical machine translation. In phrase-based MT the task of WTD is not explicitly addressed, but instead the influence of context on word translation probabilities is implicitly encoded in the model, both in the phrasal translation pairs learned from parallel text and stored in the phrase translation table (collocating words in the immediate context of an ambiguous source word are likely to end up together in a translation phrase, thus helping to disambiguate possible translations candidates) and in the target language model (usually n-gram models which tend to prefer collocations and other local dependencies).

One potential problem with this approach is that the amount of context taken into account is rather small. It is clear that word translation disambiguation often depends on cues from a wider textual context, for instance, elsewhere in the same sentence, paragraph or the document as a whole. This is beyond the scope of most phrase-based SMT approaches, which work with relatively small phrases. Another drawback of phrase-based MT (and of most data-driven MT approaches) is dependence on large aligned parallel text corpora for training purposes, a both scarce and expensive resource.

The work described here has been carried out in the context of the project PRESEMT (Pattern REcognition-based Statistically Enhanced MT; www.presemt.eu) which emphasises flexibility and adaptability towards new language pairs. A key part is to avoid relying on large and expensive parallel corpora, as such corpora are not available for the majority of language pairs; but to instead rely on very small purpose-built parallel corpora, widely available linguistic resources such as bilingual dic-

[*] This research has received funding from the European Community's 7th Framework Programme under contract nr 248307 (PRESEMT). Thanks to Els Lefever for responding to questions and request regarding the CL-WSD data sets.

tionaries, and huge monolingual corpora that can for example be easily mined from the web and automatically annotated with existing resources such as POS taggers. This combination of linguistically oriented resources and large corpora makes the system a hybrid MT system, combining data driven approaches and linguistic resources.

The next section details the word translation disambiguation task and introduces the data sets and evaluation measures used. Sections 3 and 4 then describe the n-gram and vector space modelling, respectively, followed by the experimental setup and ways to transform the vector space in Section 5. The actual experimental results are given in Section 6. Section 7 sets the work in context of efforts by others, before Section 8 discusses the results.

2 Task and data

The task addressed in this work is correctly translating a single word in context, or more formally:

Word Translation Disambiguation (WTD)
Given a source word in its context (e.g., a sentence) and a set of target word candidates (e.g., from a bilingual dictionary), the task of Word Translation Disambiguation is to select the best translation(s).

This is akin to word glossing or word-for-word translation provided that all translation candidates can be retrieved from a bilingual dictionary. WTD can thus be regarded as a ranking and filtering task. It is different, however, from full word translation, because it is assumed that all possible translations are given in advance, which is not the case in the more general task of full word translation. Full word translation can be regarded as a two-step process: (1) generation of word translation candidates, (2) word translation disambiguation. Any solution to WTD would partly solve full word translation and is therefore worthwhile to pursue.

This paper describes two approaches to WTD: First, n-gram language modelling where a surface representation of the Target Language (TL) sentence is constructed and the paths through these contexts are scored by the model. Second, vector space modelling using similarity based on the lexical semantics of the TL context to rank translation candidates according to semantic distance of the content.

AGREEMENT in the form of an exchange of letters between the European Economic Community and the Bank for International Settlements concerning the mobilization of claims held by the Member States under the medium-term financial assistance arrangements
{bank 4; bankengesellschaft 1; kreditinstitut 1; zentralbank 1; finanzinstitut 1}

1) The Office shall maintain an electronic data bank with the particulars of applications for registration of trade marks and entries in the Register. The Office may also make available the contents of this data **bank** *on CD-ROM or in any other machine-readable form.*
{datenbank 4; bank 3; datenbanksystem 1; daten 1}

(b) established as a band of 1 km in width from the banks of a river or the shores of a lake or coast for a length of at least 3 km.
{ufer 4; flussufer 3}

Table 1: Some contexts for the English word *bank* with possible German translations in the CL-WSD trial data

2.1 Data

There is a recent data set well suited for evaluating WTD systems. The 2010 exercises on Semantic Evaluation (SemEval-2) featured a Cross-Lingual Word Sense Disambiguation (CL-WSD) task (Lefever and Hoste, 2010) based on the English Lexical Substitution task from SemEval-2007. There systems had to find an alternative (synonym) substitute word or phrase for a target word in its context (McCarthy and Navigli, 2007). The CL-WSD task basically extends lexical substitution across languages, i.e., instead of finding substitutes for a word in the same language, its possible translations in another language have to be found. Although originally conceived in the context of word sense disambiguation, it is a word translation task.

While the source language in the CL-WSD data is English, there are five target languages: Dutch, French, Spanish, Italian and German. The trial set consists of 5 nouns (20 sentence contexts per noun, 100 instances in total per language), and the test set of 20 nouns (50 sentence contexts per noun, 1000 instances in total per language). Table 1 provides examples of contexts for the English word *bank* and its possible German translations from trial data.

The CL-WSD data sets were constructed in a two-step process. First, a “sense inventory” of all possi-

bank, bankanleihe, bankanstalt, bankdarlehen, bankengesellschaft, bankensektor, bankfeiertag, bankgesellschaft, bankinstitut, bankkonto, bankkredit, banknote, blutbank, daten, datenbank, datenbanksystem, euro-banknote, feiertag, finanzinstitut, flussufer, geheimkonto, geldschein, geschäftsbank, handelsbank, konto, kredit, kreditinstitut, nationalbank, notenbank, sparkasse, sparkassenverband, ufer, weltbank, weltbankgeber, west-bank, westbank, westjordanien, westjordanland, westjordanufer, westufer, zentralbank

Table 2: All German translation candidates for English *bank* as extracted from the CL-WSD trial gold standard

ble translations of a given source word was created, based on the Europarl corpus (Koehn, 2005), where alignments involving the relevant source words were manually checked. The corresponding target words were manually lemmatised and clustered into translations with a similar sense. Second, trial and test data were extracted from two independent corpora (JRC-ACQUIS and BNC). For each source word, four human translators picked the contextually appropriate sense cluster and chose up to three preferred translations it. Translations are thus restricted to those appearing in Europarl, probably introducing a slight domain bias. Each translation has an associated count indicating how many annotators considered it adequate in the given context. The spread of this count varies widely between different sentences, ranging from reasonably tight agreements on one or two candidates (with some other receiving a few votes) to sentences annotated with a long list of candidates (most receiving only one vote).

It is important to understand that the work in this paper addresses only part of the CL-WSD task: since the focus here is on WTD, it can be assumed that a perfect solution to finding translation candidates already exists. In practice this is accomplished by extracting all possible translations from the gold standard; e.g., for the English lemma *bank*, all translation candidates occurring in the trial gold standard for German are listed in Table 2.

2.2 Evaluation measures

The CL-WSD shared task employed two evaluation measures: the Best and Out-Of-Five scores (Lefever and Hoste, 2010). The Best criterion is intended to measure how well the system succeeds in delivering the best translation, i.e., the one preferred by the majority of annotators. The Out-Of-Five (OOF) criterion measures how well the top five candidates from the system match the top five translations in the gold standard. However, in WTD experiments, the Best measure has some deficiencies, most importantly that it is not normalized between 0 and 1. This results in a very uneven spread of scores, both among different target words and among the individual test sentences for each word, making it difficult — or not even meaningful — to judge differences in system performance by looking at average scores. Hence rather than using the original Best score, we adopt the normalized variant proposed by Jabbari et al. (2010), here referred to as Best$_{JHG}$.

For each sentence t_i, let H_i denote the set of human translations. For each t_i there is a function freq_i returning the count of how many annotators chose it for each term in H_i and a value $\mathit{maxfreq}_i$ for the maximum count. The pairing of H_i and freq_i constitutes a multiset representation of the human answer set. Let $|S|^i$ denote the multiset cardinality of S according to freq_i, i.e., $\sum_{a \in S} \mathit{freq}_i(a)$, the sum of all counts in S. For the first example in Table 1: $H_1 = \{\text{bank, bankengesellschaft, kreditinstitut, zentralbank, finanzinstitut}\}$; $\mathit{freq}_1(\text{bankengesellschaft}) = 4$, $\mathit{freq}_1(\text{bank}) = 1$, etc; $\mathit{maxfreq}_1 = 4$; and $|H_1|^1 = 8$.

The Best$_{JHG}$ measure is defined as follows

$$\text{Best}_{JHG}(i) = \frac{\sum_{a \in A_i} \mathit{freq}_i(a)}{\mathit{maxfreq}_i \times |A_i|} \qquad (1)$$

where A_i is the set of translations for test item i produced by the system. The optimal score of 1.0 is achieved by returning a single translation whose count is $\mathit{maxfreq}_i$, with proportionally lesser credit given to answers in H_i with smaller counts. In principle a system can output several candidates in order to "hedge its bets", but there is a penalty for non-optimal translations, so the best strategy appears to be to output just one. The systems in our experiment always produced a single translation for the Best$_{JHG}$ score, so $|A_i| = 1$ always. In the first example of Table 1, the system output $A_1 = \{\text{bank}\}$ would give $\text{Best}_{JHG}(1) = 1.0$ whereas $A_1 = \{\text{bankengesellschaft}\}$ would give $\text{Best}_{JHG}(1) = 0.25$ and $A_1 = \{\text{ufer}\}$ would give $\text{Best}_{JHG}(1) = 0.0$.

The Out-Of-Five (OOF) criterion is defined as:

$$OOF(i) = \frac{\sum_{a \in A_i} freq_i(a)}{|H_i|^i} \qquad (2)$$

In this case systems are allowed to submit up to five candidates of equal rank. It is a recall-oriented measure with no additional penalty for precision errors, so there is no benefit in outputting less than five candidates. With respect to the previous example from Table 1, the maximum score is obtained by system output $A_1 = \{$bank, bankengesellschaft, kreditinstitut, zentralbank, finanzinstitut$\}$, which gives $OOF(1) = (4 + 1 + 1 + 1 + 1)/8 = 1$, whereas $A_1 = \{$bank, bankengesellschaft, nationalbank, notenbank, sparkasse$\}$ would give $OOF(1) = (4 + 1)/8 = 0.625$. One remaining problem with the OOF measure is that the maximum score is not always one, i.e. not normalized, because sometimes the gold standard contains more than five translation alternatives.

For assessing overall system performance, the average of Best$_{JHG}$ or OOF scores across all test items for a single source word is taken. In addition, the CL-WSD task employed a "mode" variant of both scores. These were not used in the evaluations for reasons explained by Jabbari et al. (2010). All experiments use TL context to rank translation candidates for a given word in the source sentence, but for the SemEval CL-WSD data the target language sentence is not given, which means that a suitable context has to be constructed in order to perform disambiguation. This is done by collecting all translation candidates for all words in the sentence. These translation candidates are put in a bag of words from which the words' appropriate feature vectors are constructed.

3 N-gram models

Utilising n-gram language models (LMs) to rank target contexts is motivated by their widespread use and that a naive approach to order translation candidates (TC) is a useful comparison for other models. The advantage of n-gram modelling is its conceptual simplicity and practical availability. Only one model is needed to process all trial and test words.

Adapted to the WTD task, an LM can predict the likelihood of a target context being part of the language. TC sentences are constructed by combining each TC with every possible translation of their context. The shortest TC sentence is the TC itself, and if the LM is queried for all TC candidates, the most frequent would turn out on top. For the English *bank*, the most likely German candidate is *Bank*. The n-gram model should rank TC sentences of the right sense higher, because co-located phrases like *the West Bank* and *Gaza Strip* are reflected in higher n-gram probabilities of their corresponding TC sentences. This applies when the n-gram model finds the TC with the content-bearing word in the right place (when word-to-word translation is correct), unlike for multi word expressions with different surface forms in German and English.

The LM was built from sentence-separated lemmatised parts of deWac, a large monolingual web corpus of German containing over 1,627M tokens (Baroni and Kilgarriff, 2006). For each TL context, a huge number of n-grams to query the model were compiled. With a 5-gram model, a possible 4 words preceding and succeeding the word to be translated could be tested. The results of various context lengths were kept in a 2-dimensional matrix, where each index represents words ahead of, and after the TC word. Results from different context lengths are extracted, until enough TC are found (often 5). If the [-4,1] entry (4 words before, 0 after) is ranked highest, the TC represented by these n-grams would be used exclusively in output, if the limit was reached. If not, the algorithm moves on to the next matrix entry. Because of the naive word-by-word translation, few n-gram candidates of higher order were found. Ranking by no surrounding context leads to the same answer for all instances of the word, with the most frequent TL sense first.

4 Vector space modelling

A simple idea underlies the approach to WTD: given a source word in context and a number of translation candidates, search in a large TL corpus for context samples exemplifying the translation candidates. Thus, given the English word *bank* and its possible German translations *Bank, Datenbank, Ufer, ...* retrieve sentences containing *Bank*, those containing *Datenbank*, those containing *Ufer*, etc. Next search these context samples for the one most similar to the given source word context. The best TC is the one associated with this context sample.

Two basic issues need to be addressed in this approach. First, matching a given context in the source language against any context samples in the TL is obviously complicated by the difference in language. We take the straight forward approach of carrying out a word-by-word translation of the source context by means of a translation dictionary. However, there are alternative solutions to this issue conceivable, e.g., by using an existing MT system for translating the source context, or by translating the TL contexts to the source language instead.

The second issue is how to measure similarity of textual contexts, a key issue in many language processing tasks. Numerous approaches have been proposed, ranging from simple measures for word overlap and approximate string matching (Navarro, 2001), through WordNet-based and corpus-based measures (Mihalcea et al., 2006), to elaborate combinations of deep semantic analysis, word nets, domains ontologies, background knowledge and inference (Androutsopoulos and Malakasiotis, 2010). The approach to similarity taken here is that of Vector Space Models (VSM) for words (Salton, 1989). These models are based on the assumption that the meaning of a word can be inferred from its usage, i.e., distribution in text (Harris, 1954): words with similar meaning tend to occur in similar contexts.

Vector space models for words are created as high-dimensional vector representations through a statistical analysis of the contexts in which words occur. Similarity between words is defined as similarity between their context vectors in terms of some vector similarity measure, e.g., cosine similarity. A major advantage of this approach is the balance of reasonably good results with a simple model. In addition, it does not require any external knowledge resources besides a large text corpus and is fully unsupervised (human annotations are not needed).

Vector space modelling is applied to disambiguation as follows: first training and test instances are converted to feature vectors in a common multi-dimensional vector space. Next this vector space is reshaped by applying one or more transformations. The motivation for a transformation can be, e.g., to reduce dimensionality, to reduce data sparseness, to promote generalization or to possibly induce latent dimensions. Finally, for each of the vectors in the test corpus, the N most similar vectors are retrieved from the training corpus using cosine similarity, and translation candidates are predicted from the target words associated with these vectors.

5 Experimental setup

The preliminary experiments in this paper cover the German part of the CL-WSD trial data, i.e., 5 nouns with 20 sentence contexts per noun, 100 instances. We intend to run experiments on the larger CL-WSD test data set, as well as on other language pairs, once our WTD approach has sufficiently stabilized on a couple of successful models. Since the CL-WSD task offers no training data, a *training corpus* was constructed in the following steps:

Context sampling: For each translation candidate of a source word, examples of its use in context were obtained. Up to 5000 contexts per translation candidate were sampled from deWac through the web API of the SketchEngine (Kilgarriff et al., 2004). Sentences containing more than 75 tokens were skipped.

Linguistic processing: Context sentences were tokenized, lemmatised and part-of-speech tagged using the TreeTagger for German (Schmid, 1994).

Vocabulary creation: A vocabulary of terms was created over all samples sentences for all translation candidates of a single source word. First, stop words were removed according to a list of 134 German stop words. Next, function words were removed based on the POS tag, leaving mostly content words. Regular expressions were used for removing ill-formed tokens. Finally, frequency-based filtering was applied, removing all terms occurring less than 10 times, and terms occurring in more than 5% of the samples.

Vector encoding: Each context sample was encoded as a labeled (sparse) feature vector, where the features are the vocabulary terms and the feature values are the counts of these terms in the context sample at hand. The vector was labeled with the translation candidate it is a sample of. All vectors for all translation candidates of a single source word were collected in a (sparse) matrix.

The CL-WSD trial data was processed in a similar way to obtain a *test corpus*, with preprocessing carried out by the TreeTagger for English (Schmid, 1994). The test sentences were then translated word-for-word by look-up of the lemma plus POS combination in an English-German dictionary with over 900K entries obtained by reversing an existing German-English dictionary. If multiple translations for an English word were found, all were included in the sentence translation. Finally, the test sentence translations were encoded as (sparse) feature vectors in the same way as the training contexts, using the same vocabulary. As a result all German translations outside of the vocabulary were effectively deleted.

The vector space models were implemented in Gensim (Řehůřek and Sojka, 2010), an efficient VSM framework in Python. It provides a number of models for transforming vector space. In addition we implemented the Summation and PMI models. The following transformations were evaluated:

Bare vector space model. Does not apply any transformation to the feature space.

Term Frequency*Inverse document frequency (Jones, 1972) effectively gives more weight to terms that are frequent in the context but do not occur in many other contexts.

Pointwise Mutual Information (Church and Hanks, 1990) measures the association between translations candidates and context terms, and should give higher weight to terms with more discriminative power.

Latent Semantic Indexing reduces the dimensionality of the vector space by applying a Singular Value Decompostion (Deerwester et al., 1990). It is claimed to model the latent semantic relations between terms and address problems of synonymy and polysemy, hence increasing similarity between conceptually similar context vectors, even if those vectors have few terms in common.

Random Projection (also called Random Indexing). Another way to reduce the dimensionality of the vector space by projecting the original vectors into a space of nearly orthogonal random vectors. RP is claimed to result in substantially smaller matrices and faster retrieval without significant loss in performance (Sahlgren and Karlgren, 2005).

Summation model. Sums all context vectors for the same translation candidate, resulting in a centroid vector for each translation candidate. It is attractive from a computational point of view because the resulting matrix is relatively small.

For each of the 20 vectors in the test corpus for a English word, the training corpus is searched for the most similar vectors and the associated labels provide the German translations. Cosine similarity is used to calculate vector similarity. For scoring on the Best$_{JHG}$ measure, we use the single best matching vector in the training corpus. For scoring OOF, first the n best matching vectors are retrieved ($n = 1000$ in the experiments). Next the cosine similarities of all vectors with the same label are summed and the five labels with the highest summed cosine similarity constitute the output.

6 Results

Two baselines were employed. The first baseline (MostFrequentBaseline) does not rely on parallel corpora. It consists of simply selecting the translation candidate whose lemma occurs most frequently in the deWaC corpus. It therefore completely ignores the context of the words. This results in low scores on the Best$_{JHG}$ measure, although the OOF scores for bank and occupation are high. The low scores may be due to differences between predominant translations in Europarl and in deWaC. Another factor which may reduce the efficiency of target side frequencies is that the word counts can be "polluted" because a certain German word is also the translation of another very frequent English word, a problem discussed by Koehn and Knight (2000).

The second baseline (MostFrequentlyAligned) does rely on parallel corpora and was also used in the CL-WSD shared task. It is constructed by taking the translation candidate most frequently aligned to the source word in the Europarl corpus with manually corrected source word alignments. As expected, the Best$_{JHG}$ scores are consistently much higher than those of the first baseline. However, this is not so with regard to the OOF scores, which are lower than the first baseline for bank and occupation.

The simple n-gram model was employed in three different orders, uni- tri and pentagram models, but without exploring all possible priorities of context lengths (skewing to before- or after context). On average the higher-order models performed better.

	Bank	*Movement*	*Occupation*	*Passage*	*Plant*	Mean
RP (300)	15.83	17.50	11.25	5.42	20.00	14.00
LSI (200)	30.42	11.25	21.25	9.17	20.42	18.50
SumModel	**43.75**	17.50	**37.92**	7.92	**43.75**	**30.17**
PMI	32.08	**21.25**	26.67	2.92	38.33	24.25
TF*IDF	20.00	11.67	35.83	3.33	23.33	18.83
BareVSM	28.33	10.00	37.08	9.58	17.08	20.42
5-gram model	25.00	12.92	27.08	14.17	15.42	18.92
3-gram-model	10.00	16.67	24.17	11.67	6.67	13.84
1-gram-model	42.50	5.00	2.50	1.67	3.33	11.00
MostFreqAlignBaseline	6.25	19.17	35.83	**15.00**	40.00	23.25
MostFreqBaseline	1.25	5.00	2.50	1.67	10.26	4.14

Table 3: Best$_{JHG}$ scores for different models (underlined=above both baselines; bold=highest)

Results for different models in terms of the Best$_{JHG}$ score and Out-of-five scores are listed in Table 3 and Table 4. Regarding system scores, several general observations can be made. To begin with, the scores on passage tend to be lower than those on bank, occupation and plant. To a lesser extent, the same holds for scores on movement, keeping in mind that max OOF score on movement is also lower. Seemingly no correlation with the number of translation candidates though, as passage has 42 whereas bank and plant have 40 and 60 respectively. Furthermore, even though most models often outperform both baselines on some words, there is no model that consistently outperforms both baselines on all five words, although the SumModel comes close, it has a problem with passage. Looking at the mean scores over all five words, however, the SumModel outperforms both baselines. This is a promising result considering that model is smallest and does not rely on parallel text.

In a similar vein, no model consistently outperforms all others. For instance, even though SumModel yields high OOF scores on four out of five words, PMI scores higher on plant. LSI seems to provide no improvements over the BareVSM. RP performed badly, which may be related to implementation issues. TF*IDF seems to give slightly worse results in comparison to BareVSM. A possible explanation is that its feature weighting is unrelated to vector labels, so it may actually reduce the weight of discriminative context words. PMI, which does take the vector label into account, gives a slight improvement over BareVSM on the Best$_{JHG}$ score.

7 Related work

Koehn and Knight compare different methods to train word-level translation models for German-to-English translation of nouns, three of which also rely on a translation dictionary in combination with monolingual corpora (Koehn and Knight, 2000; Koehn and Knight, 2001). The first is identical to our MostFrequent baseline, the second uses a target LM to pick the most probable word sequence, and the third relies on monolingual source and target language corpora in combination with the Expectation Maximization (EM) algorithm to learn word translation probabilities. Performance of the latter two is reported to be comparable to that of using a standard SMT model trained on a parallel corpus. Our SVM approach is different in that it models a much larger contexts, i.e., full sentences. Similarly, Monz and Dorr (2005) employ an iterative procedure based on EM to estimate word translation probabilities. However, rather than relying on an n-gram LM, they measure association strength between pairs of target words, which they claim is less sensitive to word order and adjacency, and therefore data sparseness, than higher n-gram models. Their evaluation is only indirect as application of the method in a cross-lingual IR setting.

Rapp proposes methods for extracting word translations from unrelated monolingual corpora, based on the idea that words that frequently co-occur in the source language also have translations that frequently co-occur in the target language (Rapp, 1995; Rapp, 1999). His use of distributional similarity between translations in the form of a vector space is

	Bank	*Movement*	*Occupation*	*Passage*	*Plant*	Mean
MaxScore	95.60	82.62	93.58	89.57	83.22	88.92
RP (300)	24.80	12.65	22.70	8.82	21.63	18.12
LSI (200)	47.07	12.61	35.40	17.03	35.61	29.54
SumModel	**52.59**	**28.01**	42.03	17.72	32.54	**34.58**
PMI	41.00	16.33	38.41	15.47	**38.52**	29.95
TF*IDF	37.76	12.31	27.72	12.16	25.00	22.99
BareVSM	47.88	13.86	40.83	14.60	28.33	29.10
5-gram model	31.75	23.01	**37.73**	15.06	26.55	26.82
3-gram model	27.14	23.01	36.81	17.70	22.16	25.42
1-gram-model	22,92	14.17	24.39	6.63	20.04	17.63
MostFreqAlignBaseline	23.23	20.34	32.78	**27.25**	21.06	24.93
MostFreqBaseline	31.69	14.17	40.02	6.63	20.04	22.51

Table 4: Out-of-five (OOF) scores for different models (underlined=above both baselines; bold=highest)

similar to our approach. However, his goal is to bootstrap a bilingual lexicon, whereas our goal is to disambiguate. As a result, Rapp's input consists of a source word in isolation for which contexts are retrieved from a source language corpus, while our input consists of a source word in a particular context. Other work on lexical bootstrapping from monolingual corpora inspired by Rapp's work include Fung and Yee (1998) and Fung and McKeown (1997).

The submissions to the SemEval 2010 CL-WSD workshop presented a number of relevant approaches to the WTD task (van Gompel, 2010; Silberer and Ponzetto, 2010; Vilariño Ayala et al., 2010). All submitted systems, however, relied on using parallel text. Still most systems were unable to outperform the MostFrequentlyAligned baseline. Something our systems do, but a direct comparison is not fair because we only address the subtask of disambiguation and not the task of finding translation candidates.

8 Discussion and conclusion

While it is hard to draw a general conclusion on the basis of these preliminary experiments, it is our experience that it is difficult to find an approach that generalises well over any word or context for the WTD task. In our experiments, increases in performance for one set of target words were generally accompanied by reduction in performance for other words. This leads one to speculate that there are hidden variables governing the disambiguation behaviour of words such that a classification of words according to such hidden variables yield a more evenly distributed performance increase. For n-gram models the expected improvement in performance with higher-order models is observed.

In sentence space we have explored re-sampling subsets of the sentences and combining all sentences by summing all the matrix rows (sum). Attempts to cluster the sentences through for k-means and within-between cluster distances have largely been unsuccessful. Plans for future work include evaluation of the best models on the CL-WSD test data set and in the context of the full PRESEMT system.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187, May.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 87–90, Trento, Italy, April. ACL.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

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

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 414–420, Morristown, NJ, USA. ACL.

Zellig Harris. 1954. Distributional structure. *Word*, 10:146–162. Reprinted in Z. Harris, *Papers in Structural and Transformational Linguistics*, Reidel, Dordrecht, Holland 1970.

Sanaz Jabbari, Mark Hepple, and Louise Guthrie. 2010. Evaluation metrics for the lexical substitution task. In *Proceedings of the 2010 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 289–292, Los Angeles, California, June. ACL.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of Euralex*, pages 105–116, Lorient, France, July.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proceedings of the National Conference on Artificial Intelligence*, pages 711–715. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Philipp Koehn and Kevin Knight. 2001. Knowledge sources for word-level translation models. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 27–35.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand, September.

Els Lefever and Véronique Hoste. 2010. SemEval-2010 Task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden, July. ACL.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. ACL.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21th National Conference on Artifical Intelligence*, Boston, Massachusetts, July. AAAI.

Christof Monz and Bonnie J. Dorr. 2005. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th International Conference on Research and Development in Information Retrieval*, pages 520–527, Salvador, Brazil, August. ACM SIGIR.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, March.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, MIT, Cambridge, Massachusetts, June. ACL.

Reinhard 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*, pages 519–526, Madrid, Spain, July. ACL.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 45–50, Valetta, Malta, May. ELRA. Workshop on New Challenges for NLP Frameworks.

Magnus Sahlgren and Jussi Karlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(2), June. Special Issue on Parallel Texts.

Gerard Salton. 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, Reading, Massachusetts.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the 1st International Conference on New Methods in Natural Language Processing*, pages 44–49, University of Manchester Institute of Science and Technology, Manchester, England, September.

Carina Silberer and Simone Paolo Ponzetto. 2010. UHD: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 134–137, Uppsala, Sweden, July. ACL.

Maarten van Gompel. 2010. UvT-WSD1: A cross-lingual word sense disambiguation system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 238–241, Uppsala, Sweden, July. ACL.

Darnes Vilariño Ayala, Carlos Balderas Posada, David Eduardo Pinto Avendaño, Miguel Rodríguez Hernández, and Saul León Silverio. 2010. FCC: Modeling probabilities with GIZA++ for Task 2 and 3 of SemEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 112–116, Uppsala, Sweden, July. ACL.