Improving Word Translation Disambiguation by Capturing Multiword Expressions with Dictionaries

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

Abstract

The paper describes a method for identifying and translating multiword expressions using a bi-directional dictionary. While a dictionary-based approach suffers from limited recall, precision is high; hence it is best employed alongside an approach with complementing properties, such as an n-gram language model.

We evaluate the method on data from the English-German translation part of the cross-lingual word sense disambiguation task in the 2010 semantic evaluation exercise (SemEval). The output of a baseline disambiguation system based on n-grams was substantially improved by matching the target words and their immediate contexts against compound and collocational words in a dictionary.

1 Introduction

Multiword expressions (MWEs) cause particular lexical choice problems in machine translation (MT), but can also be seen as an opportunity to both generalize outside the bilingual corpora often used as training data in statistical machine translation approaches and as a method to adapt to specific domains. The identification of MWEs is in general important for many language processing tasks (Sag et al., 2002), but can be crucial in MT: since the semantics of many MWEs are non-compositional, a suitable translation cannot be constructed by translating the words in isolation. Identifying MWEs can help to identify idiomatic or otherwise fixed language usage, leading to more fluent translations, and potentially reduce the amount of lexical choice an MT system faces during target language generation.

In any translation effort, automatic or otherwise, the selection of target language lexical items to include in the translation is a crucial part of the final translation quality. In rule-based systems lexical choice is derived from the semantics of the source words, a process which often involves complex semantic composition. Data-driven systems on the other hand commonly base their translations nearly exclusively on cooccurrences of bare words or phrases in bilingual corpora, leaving the responsibility of selecting lexical items in the translation entirely to the local context found in phrase translation tables and language models with no explicit notion of the source or target language semantics. Still, systems of this type have been shown to produce reasonable translation quality without explicitly considering word translation disambiguation.

Bilingual corpora are scarce, however, and unavailable for most language pairs and target domains. An alternative approach is to build systems based on large monolingual knowledge sources and bilingual lexica, as in the hybrid MT system PRESEMT (Sofianopoulos et al., 2012). Since such a system explicitly uses a translation dictionary, it must at some point in the translation process decide which lexical entries to use; thus a separate word translation disambiguation module needs to be incorporated. To research available methods in such a module we have identified a task where we can use public datasets for measuring how well a method is able to select the optimal of many translation choices from a source language sentence.

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 21–30,
Atlanta, Georgia, 13-14 June 2013. ©2013 Association for Computational Linguistics

In phrase-based statistical MT systems, the translation of multiword expressions can be a notable source of errors, despite the fact that those systems explicitly recognize and use alignments of sequential chunks of words. Several researchers have approached this problem by adding MWE translation tables to the systems, either through expanding the phrase tables (Ren et al., 2009) or by injecting the MWE translations into the decoder (Bai et al., 2009). Furthermore, there has been some interest in automatic mining of MWE pairs from bilingual corpora as a task in itself: Caseli et al. (2010) used a dictionary for evaluation of an automatic MWE extraction procedure using bilingual corpora. They also argued for the filtering of stopwords, similarly to the procedure described in the present paper. Sharoff et al. (2006) showed how MWE pairs can be extracted from comparable monolingual corpora instead of from a parallel bilingual corpus.

The methodology introduced in this paper employs bilingual dictionaries as a source of multiword expressions. Relationships are induced between the source sentence and candidate translation lexical items based on their correspondence in the dictionary. Specifically, we use a deterministic multiword expression disambiguation procedure based on translation dictionaries in both directions (from source to target language and vice versa), and a baseline system that ranks target lexical items based on their immediate context and an n-gram language model. The n-gram model represents a high-coverage, low-precision companion to the dictionary approach (i.e., it has complementary properties). Results show that the MWE dictionary information substantially improves the baseline system.

The 2010 Semantic Evaluation exercise (SemEval'10) featured a shared task on Cross-Lingual Word Sense Disambiguation (CL-WSD), where the focus was on disambiguating the translation of a single noun in a sentence. The participating systems were given an English word in its context and asked to produce appropriate substitutes in another language (Lefever and Hoste, 2010b). The CL-WSD data covers Dutch, French, Spanish, Italian and German; however, since the purpose of the experiments in this paper just was to assess our method's ability to choose the right translation of a word given its context, we used the English-to-German part only.

The next section details the employed disambiguation methodology and describes the data sets used in the experiments. Section 3 then reports on the results of experiments applying the methodology to the SemEval datasets, particularly addressing the impact of the dictionary MWE correspondences. Finally, Section 4 sums up the discussion and points to issues that can be investigated further.

2 Methodology

The core of the disambiguation model introduced in this paper is dictionary-based multiword extraction. Multiword extraction is done in both a direct and indirect manner: *Direct extraction* uses adjacent words in the source language in combination with the word to be translated, if the combination has an entry in the source-to-target language (SL–TL) dictionary. *Indirect extraction* works in the reverse direction, by searching the target-to-source (TL–SL) dictionary and looking up translation candidates for the combined words. Using a dictionary to identify multiword expressions after translation has a low recall of target language MWEs, since often there either are no multiword expressions to be discovered, or the dictionary method is unable to find a translation for an MWE. Nevertheless, when an MWE really is identified by means of the dictionary-based method, the precision is high.

Due to the low recall, relying on multiword expressions from dictionaries would, however, not be sufficient. Hence this method is combined with an n-gram language model (LM) based on a large target language corpus. The LM is used to rank translation candidates according to the probability of the n-gram best matching the context around the translation candidate. This is a more robust but less precise approach, which servers as the foundation for the high-precision but low-recall dictionary approach.

In the actual implementation, the n-gram method thus first provides a list of its best suggestions (currently top-5), and the dictionary method then prepends its candidates to the top of this list. Consequently, n-gram matching is described before dictionary-based multiword extraction in the following section. First, however, we introduce the data sets used in the experiments.

(a) 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}

(b) 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}

(c) 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: Examples of contexts for the English word *bank* with possible German translations

2.1 The CL-WSD Datasets

The data sets used for the SemEval'10 Cross-Lingual Word Sense Disambiguation task were constructed by making a 'sense inventory' of all possible target language translations of a given source language word based on word-alignments in Europarl (Koehn, 2005), with alignments involving the relevant source words being manually checked. The retrieved target words were manually lemmatised and clustered into translations with a similar sense; see Lefever and Hoste (2010a) for details.

Trial and test instances were extracted from two other corpora, JRC-Acquis (Steinberger et al., 2006) and BNC (Burnard, 2007). The trial data for each language consists of five nouns (with 20 sentence contexts per noun), and the test data of twenty nouns (50 contexts each, so 1000 in total per language, with the CL-WSD data covering Dutch, French, Spanish, Italian and German). Table 1 provides examples from the trial data of contexts for the English word *bank* and its possible translations in German.

Gold standard translations were created by having four human translators picking the contextually appropriate sense for each source word, choosing 0–3 preferred target language translations for it. The 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 to be among their top-3 preferred translations in the given context.

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 *bank* as extracted from the gold standard

In this way, for the English lemma *bank*, for example, the CL-WSD trial gold standard for German contains the word *Bank* itself, together with 40 other translation candidates, as shown in Table 2. Eight of those are related to river banks (*Ufer*, but also, e.g., *Westbank* and *Westjordanland*), three concern databases (*Datenbank*), and one is for blood banks. The rest are connected to different types of financial institutions (such as *Handelsbank* and *Finanzinstitut*, but also by association *Konto, Weldbankgeber, Banknote, Geldschein, Kredit*, etc.).

2.2 N-Gram Context Matching

N-gram matching is used to produce a ranked list of translation candidates and their contexts, both in order to provide robustness and to give a baseline performance. The n-gram models were built using the IRSTLM toolkit (Federico et al., 2008; Bungum and Gambäck, 2012) on the DeWaC corpus (Baroni and Kilgarriff, 2006), using the stopword list from NLTK (Loper and Bird, 2002). The n-gram matching procedure consists of two steps:

1. An n^{th} order source context is extracted and the translations for each SL word in this context are retrieved from the dictionary. This includes stopword filtering of the context.

2. All relevant n-grams are inspected in order from left to right and from more specific (5-grams) to least specific (single words).

For each part of the context with matching n-grams in the target language model, the appropriate target translation candidates are extracted and ranked according to their language model probability. This results in an n-best list of translation candidates.

Since dictionary entries are lemma-based, lemmatization was necessary to use this approach in combination with the dictionary enhancements. The source context is formed by the lemmata in the sentence surrounding the focus word (the word to be disambiguated) by a window of up to four words in each direction, limited by a 5-gram maximum length. In order to extract the semantically most relevant content, stopwords are removed before constructing this source word window. For each of the 1–5 lemmata in the window, the relevant translation candidates are retrieved from the bilingual dictionary. The candidates form the ordered translation context for the source word window.

The following example illustrates how the translation context is created for the focus word 'bank'. First the relevant part of the source language sentence with the focus word in bold face:

(1) The BIS could conclude stand-by credit agreements with the creditor countries' central **bank** if they should so request.

For example, using a context of two words in front and two words after the focus word, the following source language context is obtained after a preprocessing involving lemmatization, stopword removal, and insertion of sentence start (<s>) and end markers (</s>):

(2) country central **bank** request </s>

From this the possible n-grams in the target side context are generated by assembling all ordered combinations of the translations of the source language words for each context length: the widest contexts (5-grams) are looked up first before moving on to narrower contexts, and ending up with looking up only the translation candidate in isolation.

Each of the n-grams is looked up in the language model and for each context part the n-grams are ordered according to their language model probability. Table 3 shows a few examples of such generated n-grams with their corresponding scores from the n-gram language model.[1] The target candidates (italics) are then extracted from the ordered list of target language n-grams. This gives an n-best list of translation candidates from which the top-1 or top-5 can be taken. Since multiple senses in the dictionary can render the same literal output, duplicate translation candidates are filtered out from the n-best list.

n	n-gram	LM score
5	land mittig *bank* nachsuchen </s>	Not found
4	mittig *bank* nachsuchen </s>	Not found
3	mittig *bank* nachsuchen	Not found
3	*kredit* anfragen </s>	-0.266291
2	mittig *bank*	-3.382560
2	zentral *blutbank*	-5.144870
1	*bank*	-3.673000

Table 3: Target language n-gram examples from lookups of stopword-filtered lemmata *country central bank request* reported in log scores. The first 3 n-grams were not found in the language model.

2.3 Dictionary-Based Context Matching

After creating the n-gram based list of translation candidates, additional candidates are produced by looking at multiword entries in a bilingual dictionary. The existence of multiword entries in the dictionary corresponding to adjacent lemmata in the source context or translation candidates in the target context is taken as a clear indicator for the suitability of a particular translation candidate. Such entries are added to the top of the n-best list, which represents a strong preference in the disambiguation system.

Dictionaries are used in all experiments to look up translation candidates and target language translations of the words in the context, but this approach is mining the dictionaries by using lookups of greater length. Thus is, for example, the dictionary entry *Community Bank* translated to the translation candidate *Commerzbank*; this translation candidate would be put on top of the list of prioritized answers.

Two separate procedures are used to find such indicators, a direct procedure based on the source context and an indirect procedure based on the weaker target language context. These are detailed in pseudocode in Algorithms 1 and 2, and work as follows:

Source Language (SL) Method (Algorithm 1)

If there is a dictionary entry for the source word and one of its adjacent words, search the set of translations for any of the translation candidates for the word alone. Specifically, transla-

[1] There are no scores for 4- and 5-grams; as expected when using direct translation to generate target language n-grams.

Algorithm 1 SL algorithm to rank translation candidates (tcands) for SL lemma b given list of $tcands$

```
procedure FindCand(list $rlist$,SL-lemma $b$, const $tcands$)    ▷ rlist is original ranking
    $comblemmas \leftarrow list(previouslemma(b) + b, b + nextlemma(b))$    ▷ Find adjacent lemmata
    for $lem \in comblemmas$ do
        $c \leftarrow sl\text{-}dictionary\text{-}lookup(lem)$    ▷ Look up lemma in SL→TL dict.
        if $c \in tcands$ then $rlist \leftarrow list(c + rlist)$    ▷ Push lookup result $c$ onto $rlist$ if in $tcands$
        end if
    end for
    return $rlist$    ▷ Return new list with lemmata whose translations were in $tcands$ on top
end procedure
```

Algorithm 2 TL algorithm to rank translation candidates (tcands) for SL lemma b given list of $tcands$
[The ready-made TL $tcands$ from the dataset are looked up in TL-SL direction. It is necessary to keep a list of the reverse-translation of the individual $tcand$ as well as the original $tcand$ itself, in order to monitor which $tcand$ it was. If the SL context is found in either of these reverse lookups the matching $tcand$ is ranked high.]

```
procedure FindCand(list $rlist$,SL-lemma $b$, const $tcands$)    ▷ rlist is original ranking
    for $cand \in tcands$ do    ▷ Assemble list of TL translations
        $translist \leftarrow list(cand, tl\text{-}dictionary\text{-}lookup(cand)) + translist$
                    ▷ Append TL→SL lookup results of $tcands$ with $cand$ as id
    end for
    for $cand, trans \in translist$ do
        if $previouslemma(b) \| nextlemma(b) \in trans$ then    ▷ If $trans$ contains either SL lemma
            $rlist \leftarrow list(cand) + rlist$    ▷ append this $cand$ onto $rlist$
        end if
    end for
    return $rlist$
        ▷ Return $tcands$ list; top-ranking $tcands$ whose SL-neighbours were found in TL→SL lookup
end procedure
```

tions of the combination of the source word and an adjacent word in the context are matched against translation candidates for the word.

Target Language (TL) Method (Algorithm 2)
If a translation candidate looked up in the reverse direction matches the source word along with one or more adjacent words, it is a good translation candidate. TL candidates are looked up in a TL–SL dictionary and multiword results are matched against SL combinations of disambiguation words and their immediate contexts.

For both methods the dictionary entry for the target word or translation candidate is matched against the immediate context. Thus both methods result in two different lookups for each focus word, combining it with the previous and next terms, respectively. This is done exhaustively for all combinations of translations of the words in the context window. Only one adjacent word was used, since very few of the candidates were able to match the context even with one word. Hence, virtually none would be found with more context, making it very unlikely that larger contexts would contribute to the disambiguation procedure, as wider matches would also match the one-word contexts.

Also for both methods, translation candidates are only added once, in case the same translation candidate generates hits with either (or both) of the methods. Looking at the running example, stopword filtered and with lemmatized context:

(3) country central **bank** request

This example generates two source language multiword expressions, *central bank* and *bank request.* In the source language method, these word combinations are looked up in the dictionary where the *zentralbank* entry is found for *central bank*, which is also found as a translation candidate for *bank*.

The target language method works in the reverse order, looking up the translation candidates in the TL–SL direction and checking if the combined lemmata are among the candidates' translations into the source language. In the example, the entry *zentralbank:central bank* is found in the dictionary, matching the source language context, so *zentralbank* is assumed to be a correct translation.

2.4 Dictionaries

Two English-German dictionaries were used in the experiments, both with close to 1 million entries (translations). One is a free on-line resource, while the other was obtained by reversing an existing proprietary German-English dictionary made available to the authors by its owners:

- The GFAI dictionary (called 'D1' in Section 3 below) is a proprietary and substantially extended version of the Chemnitz dictionary, with 549k EN entries including 433k MWEs, and 552k DE entries (79k MWEs). The Chemnitz electronic German-English dictionary[2] itself contains over 470,000 word translations and is available under a GPL license.

- The freely available CC dictionary[3] ('D2' below) is an internet-based German-English and English-German dictionary built through user generated word definitions. It has 565k/440k (total/MWE) EN and 548k/210k DE entries.

Note that the actual dictionaries are irrelevant to the discussion at hand, and that we do not aim to point out strengths or weaknesses of either dictionary, nor to indicate a bias towards a specific resource.

3 Results

Experiments were carried out both on the trial and test data described in Section 2.1 (5 trial and 20 test words; with 20 resp. 50 instances for each word; in total 1100 instances in need of disambiguation). The results show that the dictionaries yield answers with high precision, although they are robust enough to solve the SemEval WSD challenge on their own.

For measuring the success rate of the developed models, we adopt the 'Out-Of-Five' (OOF) score (Lefever and Hoste, 2010b) from the SemEval'10 Cross-Lingual Word Sense Disambiguation task. The Out-Of-Five criterion measures how well the top five candidates from the system match the top five translations in the gold standard:

$$OOF(i) = \frac{\sum_{a \in A_i} \mathit{freq}_i(a)}{|H_i|}$$

where H_i denotes the multiset of translations proposed by humans for the focus word in each source sentence s_i ($1 \leq i \leq N$, N being the number of test items). A_i is the set of translations produced by the system for source term i. Since each translation has an associated count of how many annotators chose it, there is for each s_i a function freq_i returning this count for each term in H_i (0 for all other terms), and $max\ \mathit{freq}_i$ gives the maximal count for any term in H_i. For the first example in Table 1:

$$\begin{cases} H_1 = \{\text{bank, bank, bank, bank, zentralbank,} \\ \quad \text{bankengesellschaft, kreditinstitut, finanzinstitut}\} \\ \mathit{freq}_1(\text{bank}) = 4 \\ \ldots \\ \mathit{freq}_1(\text{finanzinstitut}) = 1 \\ max\mathit{freq}_1 = 4 \end{cases}$$

and the cardinality of the multiset is: $|H_1| = 8$. This equates to the sum of all top-3 preferences given to the translation candidates by all annotators.

For the Out-Of-Five evaluation, the CL-WSD systems were allowed to submit up to five candidates of equal rank. OOF 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 $A_2 = \{$bank, bankengesellschaft, nationalbank, notenbank, sparkasse$\}$ would give $OOF(1) = (4+1)/8 = 0.625$.[4]

[2] http://dict.tu-chemnitz.de/

[3] http://www.dict.cc/

[4] Note that the maximum OOF score is not always 1 (i.e., it is not normalized), since the gold standard sometimes contains more than five translation alternatives.

	Source language			Target language			All
Dictionary	D1	D2	comb	D1	D2	comb	comb
Top	*8.89*	6.99	*8.89*	22.71	24.43	***25.34***	24.67
Low	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean	2.71	0.99	*3.04*	8.35	7.10	*9.24*	**10.13**

Table 4: F_1-score results for individual dictionaries

	Source language			Target language			All
Dictionary	D1	D2	comb	D1	D2	comb	comb
coach	1.00	0.00	1.00	0.21	0.00	0.21	0.21
education	0.83	0.67	0.83	0.47	0.62	0.54	0.53
execution	0.00	0.00	0.00	0.17	0.22	0.17	0.17
figure	1.00	0.00	1.00	0.51	0.57	0.55	0.55
job	0.88	0.80	0.94	0.45	0.78	0.46	0.44
letter	1.00	0.00	1.00	0.66	0.75	0.62	0.66
match	1.00	1.00	1.00	0.80	0.50	0.80	0.80
mission	0.71	0.33	0.71	0.46	0.37	0.36	0.36
mood	0.00	0.00	0.00	0.00	0.00	0.00	0.00
paper	0.68	0.17	0.68	0.53	0.35	0.55	0.55
post	1.00	1.00	1.00	0.39	0.48	0.45	0.48
pot	0.00	0.00	0.00	1.00	1.00	1.00	1.00
range	1.00	1.00	1.00	0.28	0.37	0.30	0.30
rest	1.00	0.67	1.00	0.60	0.56	0.56	0.58
ring	0.09	0.00	0.09	0.37	0.93	0.38	0.38
scene	1.00	0.00	1.00	0.50	0.42	0.44	0.50
side	1.00	0.00	1.00	0.21	0.16	0.23	0.27
soil	1.00	0.00	1.00	0.72	0.58	0.66	0.69
strain	0.00	0.00	0.00	0.51	0.88	0.55	0.55
test	1.00	1.00	1.00	0.62	0.52	0.57	0.61
Mean	0.84	0.74	0.84	0.50	0.56	0.49	0.51

Table 5: Precision scores for all terms filtering out those instances for which no candidates were suggested

For assessing overall system performance in the experiments, we take the best ('Top'), worst ('Low'), and average ('Mean') of the OOF scores for all the SL focus words, with F_1-score reported as the harmonic mean of the precision and recall of the OOF scores. Table 4 shows results for each dictionary approach on the test set, with 'D1' being the GFAI dictionary, 'D2' the CC dictionary, and 'comb' the combination of both. Target language look-up contributes more to providing good translation candidates than the source language methodology, and also outperforms a strategy combining all dictionaries in both directions ('All comb').

Filtering out the instances for which no candidate translation was produced, and taking the average precision scores only over these, gives the results shown in Table 5. Markedly different precision scores can be noticed, but the source language method again has higher precision on the suggestions it makes than the target language counterpart.

	Source language		Target language	
Dictionary	D1	D2	D1	D2
Mean	*3.25*	1.5	**12.65**	11.45
Total	223	256	**1,164**	880

Table 6: Number of instances with a translation candidate ('Mean') and the total number of suggested candidates

	Most Freq	Most Freq Aligned	5-gram	5-gram + Dict	All Dict Comb	VSM Model
Top	51.77	68.71	52.02	52.74	24.67	**55.92**
Low	1.76	9.93	14.09	**15.40**	0.00	10.73
Mean	21.18	34.61	30.36	**36.38**	10.13	30.30

Table 7: Overview of results (F_1-scores) on SemEval data

As shown in Table 6, this higher precision is offset by lower coverage, with far fewer instances actually producing a translation candidate with the dictionary lookup methods. There is a notable difference in the precision of the SL and TL approaches, coinciding with more candidates produced by the latter. Several words in Table 5 give 100% precision scores for at least one dictionary, while a few give 0% precision for some dictionaries. The word 'mood' even has 0% precision for both dictionaries in both directions.

Table 7 gives an overview of different approaches to word translation disambiguation on the dataset. For each method, the three lines again give both the best and worst scoring terms, and the mean value for all test words. The maximum attainable score for each of those would be 99.28, 90.48 and 95.47, respectively, but those are perfect scores not reachable for all items, as described above (OOF-scoring). Instead the columns *Most Freq* and *Most Freq aligned* give the baseline scores for the SemEval dataset: the translation most frequently seen in the corpus and the translation most frequently aligned in a word-aligned parallel corpus (Europarl), respectively. Then follows the results when using only a stopword-filtered *5-gram* model built with the IRSTLM language modeling kit (Federico and Cettolo, 2007), and when combining the 5-gram model with the dictionary approach (*5-gram + Dict*).

The next column (*All Dict Comb*) shows how the dictionary methods fared on their own. The combined dictionary approach has low recall (see Table 6) and does not alone provide a good solution to the overall problem. Due to high precision, however, the approach is able to enhance the n-gram method that already produces acceptable results. Finally, the column *VSM Model* as comparison gives the results obtained when using a Vector Space Model for word translation disambiguation (Marsi et al., 2011).

Comparing the dictionary approach to state-of-the-art monolingual solutions to the WTD problem on this dataset shows that the approach performs better for the Lowest and Mean scores of the terms, but not for the Top scores (Lynum et al., 2012). As can be seen in Table 7, the vector space model produced the overall best score for a single term. However, the method combining a 5-gram language model with the dictionary approach was best both at avoiding really low scores for any single term and when comparing the mean scores for all the terms.

4 Discussion and Conclusion

The paper has presented a method for using dictionary lookups based on the adjacent words in both the source language text and target language candidate translation texts to disambiguate word translation candidates. By composing lookup words by using both neighbouring words, improved disambiguation performance was obtained on the data from the SemEval'10 English-German Cross-Lingual Word Sense Disambiguation task. The extended use of dictionaries proves a valuable source of information for disambiguation, and can introduce low-cost phrase-level translation to quantitative Word Sense Disambiguation approaches such as N-gram or Vector Space Model methods, often lacking the phrases-based dimension.

The results show clear differences between the source and target language methods of using dictionary lookups, where the former has very high precision (0.84) but low coverage, while the TL method compensates lower precision (0.51) with markedly better coverage. The SL dictionary method provided answers to only between 1.5 and 3.25 of 50 instances per word on average, depending on the dictionary. This owes largely to the differences in algorithms, where the TL method matches any adjacent lemma to the focus word with the translation of the pre-defined translation candidates, whereas the SL method matches dictionaries of the combined lemmata of the focus word and its adjacent words to the same list of translation candidates. False positives are expected with lower constraints such as these. On the SemEval data, the contribution of the dictionary methods to the n-grams is mostly in improving the average score.

The idea of acquiring lexical information from corpora is of course not new in itself. So did, e.g., Rapp (1999) use vector-space models for the purpose of extracting ranked lists of translation candidates for extending a dictionary for word translation disambiguation. Chiao and Zweigenbaum (2002) tried to identify translational equivalences by investigating the relations between target and source language word distributions in a restricted domain, and also applied reverse-translation filtering for improved performance, while Sadat et al. (2003) utilised non-aligned, comparable corpora to induce a bilingual lexicon, using a bidirectional method (SL→TL, TL→SL, and a combination of both).

Extending the method to use an arbitrary size window around all words in the context of each focus word (not just the word itself) could identify more multiword expressions and generate a more accurate bag-of-words for a data-driven approach. Differences between dictionaries could also be explored, giving more weight to translations found in two or more dictionaries. Furthermore, the differences between the SL and TL methods could explored further, investigating in detail the consequences of using a symmetrical dictionary, in order to study the effect that increased coverage has on results. Testing the idea on more languages will help verify the validity of these findings.

Acknowledgements

This research has received funding from NTNU and from the European Community's 7th Framework Programme under contract nr 248307 (PRESEMT). Thanks to the other project participants and the anonymous reviewers for several very useful comments.

References

Bai, M.-H., You, J.-M., Chen, K.-J., and Chang, J. S. (2009). Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 478–486, Singapore. ACL.

Baroni, M. and Kilgarriff, A. (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. ACL.

Bungum, L. and Gambäck, B. (2012). Efficient n-gram language modeling for billion word web-corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 6–12, Istanbul, Turkey. ELRA. Workshop on Challenges in the Management of Large Corpora.

Burnard, L., editor (2007). *Reference Guide for the British National Corpus (XML Edition)*. BNC Consortium, Oxford, England. `http://www.natcorp.ox.ac.uk/XMLedition/URG`.

Caseli, H. d. M., Ramisch, C., das Graças Volpe Nunes, M., and Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77. Special Issue on Multiword expression: hard going or plain sailing.

Chiao, Y.-C. and Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized comparable corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 1–5, Philadelphia, Pennsylvania. ACL. Also published in *AMIA Annual Symposium 2002*, pp. 150–154.

Federico, M., Bertoldi, N., and Cettolo, M. (2008). Irstlm: an open source toolkit for handling large scale language models. In *INTERSPEECH*, pages 1618–1621. ISCA.

Federico, M. and Cettolo, M. (2007). Efficient handling of n-gram language models for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 88–95, Prague, Czech Republic. ACL. 2nd Workshop on Statistical Machine Translation.

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

Lefever, E. and Hoste, V. (2010a). Construction of a benchmark data set for cross-lingual word sense disambiguation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 1584–1590, Valetta, Malta. ELRA.

Lefever, E. and Hoste, V. (2010b). SemEval-2010 Task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 15–20, Uppsala, Sweden. ACL. 5th International Workshop on Semantic Evaluation.

Loper, E. and Bird, S. (2002). NLTK: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

LREC06 (2006). *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genova, Italy. ELRA.

Lynum, A., Marsi, E., Bungum, L., and Gambäck, B. (2012). Disambiguating word translations with target language models. In *Proceedings of the 15th International Conference on Text, Speech and Dialogue*, pages 378–385, Brno, Czech Republic. Springer.

Marsi, E., Lynum, A., Bungum, L., and Gambäck, B. (2011). Word translation disambiguation without parallel texts. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation*, pages 66–74, Barcelona, Spain.

Rapp, R. (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. ACL.

Ren, Z., Lü, Y., Cao, J., Liu, Q., and Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 47–54, Singapore. ACL. Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications.

Sadat, F., Yoshikawa, M., and Uemura, S. (2003). Learning bilingual translations from comparable corpora to cross-language information retrieval: Hybrid statistics-based and linguistics-based approach. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 57–64, Sapporo, Japan. ACL. 6th International Workshop on Information Retrieval with Asian languages; a shorter version published in *ACL Annual Meeting 2003*, pp. 141–144.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing: Proceedings of the 3rd International Conference*, number 2276 in Lecture Notes in Computer Science, pages 189–206, Mexico City, Mexico. Springer-Verlag.

Sharoff, S., Babych, B., and Hartley, A. (2006). Using collocations from comparable corpora to find translation equivalents. In LREC06 (2006), pages 465–470.

Sofianopoulos, S., Vassiliou, M., and Tambouratzis, G. (2012). Implementing a language-independent MT methodology. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 1–10, Jeju, Korea. ACL. First Workshop on Multilingual Modeling.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In LREC06 (2006), pages 2142–2147.