Bilingual lexicon extraction from comparable corpora using in-domain terms

Azniah Ismail
Department of Computer Science
University of York
azniah@cs.york.ac.uk

Suresh Manandhar
Department of Computer Science
University of York
suresh@cs.york.ac.uk

Abstract

Many existing methods for bilingual lexicon learning from comparable corpora are based on similarity of context vectors. These methods suffer from noisy vectors that greatly affect their accuracy. We introduce a method for filtering this noise allowing highly accurate learning of bilingual lexicons. Our method is based on the notion of *in-domain terms* which can be thought of as the most important contextually relevant words. We provide a method for identifying such terms. Our evaluation shows that the proposed method can learn highly accurate bilingual lexicons without using orthographic features or a large initial seed dictionary. In addition, we also introduce a method for measuring the similarity between two words in different languages without requiring any initial dictionary.

1 Introduction

In bilingual lexicon extraction, the context-based approach introduced by Rapp (1995) is widely used (Fung, 1995; Diab and Finch, 2000; among others). The focus has been on learning from comparable corpora since the late 1990s (Rapp, 1999; Koehn and Knight, 2002; among others). However, so far, the accuracy of bilingual lexicon extraction using comparable corpora is quite poor especially when orthographic features are not used. Moreover, when orthographic features are not used, a large initial seed dictionary is essential in order to acquire higher accuracy lexicon (Koehn and Knight, 2002). This means that current methods are not suitable when the language pairs are not closely related or when a large initial seed dictionary is unavailable.

When learning from comparable corpora, a large initial seed dictionary does not necessarily guarantee higher accuracy since the source and target texts are poorly correlated. Thus, inducing highly accurate bilingual lexicon from comparable corpora has so far been an open problem.

In this paper, we present a method that is able to improve the accuracy significantly without requiring a large initial bilingual dictionary. Our approach is based on utilising *highly associated terms* in the context vector of a source word. For example, the source word *powers* is highly associated with the context word *delegation*. We note that, firstly, both share context terms such as *parliament* and *affairs*. And, secondly, the translation equivalents of *powers* and *delegation* in the target language are not only highly associated but they also share context terms that are the translation equivalents of *parliament* and *affairs* (see Figure 1).

2 Related work

Most of the early work in bilingual lexicon extraction employ an initial seed dictionary. A large bilingual lexicon with 10k to 20k entries is necessary (Fung, 1995; Rapp, 1999).

Koehn and Knight (2002) introduce techniques for constructing the initial seed dictionary automatically. Their method is based on using identical spelling features. The accuracy of such initial bilingual lexicon is almost 90.0 percent and can be increased by restricting the word length (Koehn and Knight, 2002). Koehn and Knight found approximately 1000 identical words in their German

Coling 2008: Poster Volume, pages 481–489,
Beijing, August 2010

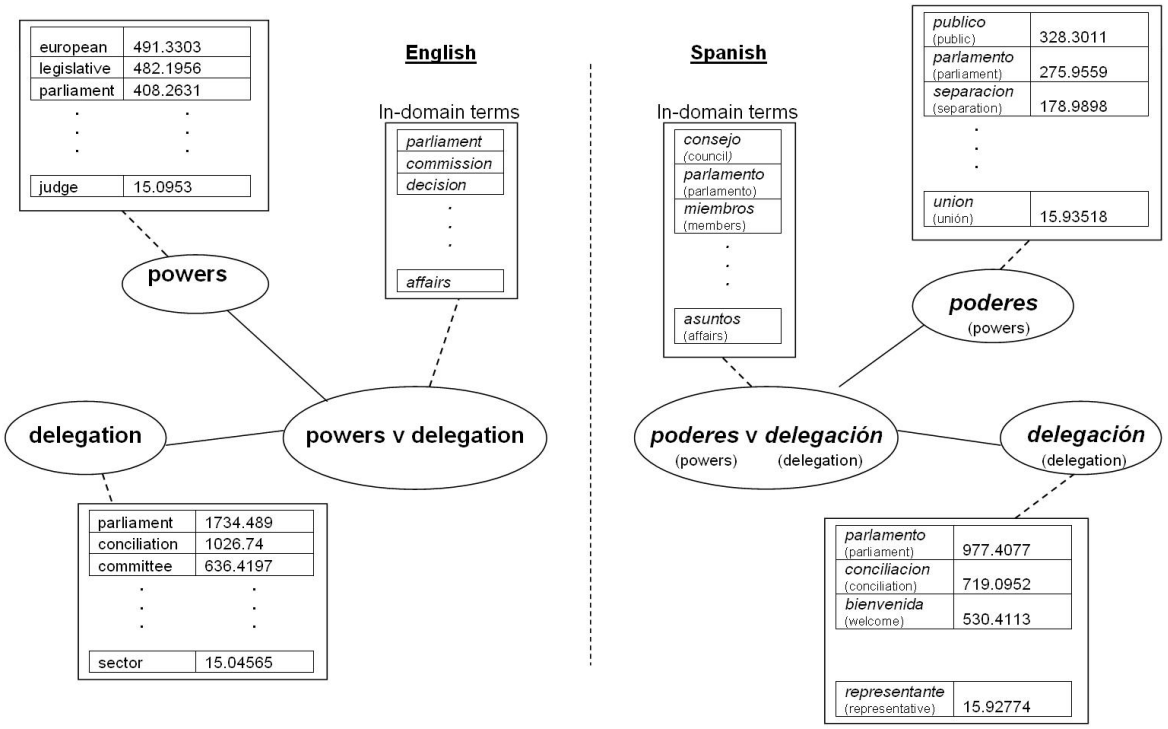

Figure 1: An example of in-domain terms that co-occur in English and Spanish. The source word is *powers* and the target word is *poderes*. The word *delegation* and *delegacion* are the highly associated words with the source word and the target word respectively. Their in-domain terms, as shown in the middle, can be used to map the source word in context of word *delegation* to its corresponding target word in context of *delegacion*.

and English monolingual corpora. They expanded the lexicon with the standard context-based approach and achieved about 25.0 percent accuracy (Koehn and Knight, 2002).

Similar techniques were used in Haghighi et al. (2008) who employ *dimension reduction* in the extraction method. They recorded 58.0 percent as their best F_1 score for the context vector approach on non-parallel comparable corpora containing *Wikipedia* articles. However, their method scores less on comparable corpora containing distinct sentences derived from the *Europarl English-Spanish* corpus.

3 Learning in-domain terms

In the standard context vector approach, we associate each source word and target word with their context vectors. The source and target context vectors are then compared using the initial seed dictionary and a similarity measure. Learning from comparable corpora is particularly problematic due to data sparsity, as important context terms may not occur in the training corpora while some may occur but with low frequency and can be missed. Some limitations may also be due to the size of the initial seed dictionary being small.

The initial seed dictionary can also contribute irrelevant or less relevant features that can mislead the similarity measure especially when the number of dimensions is large. The approach we adopt attempts to overcome this problem.

In Figure 1, for the source word *powers*, *delegation* is the highly associated word. Both *powers* and *delegation* share common contextual terms such as *parliament* and *affairs*. Now the translation equivalent of *delegation* is *delegacion*. For the potential translation equivalent *poderes*, we see that the common contextual terms shared by *powers* and *poderes* are terms *parlamento* (*parliament*) and *asuntos* (*affairs*).

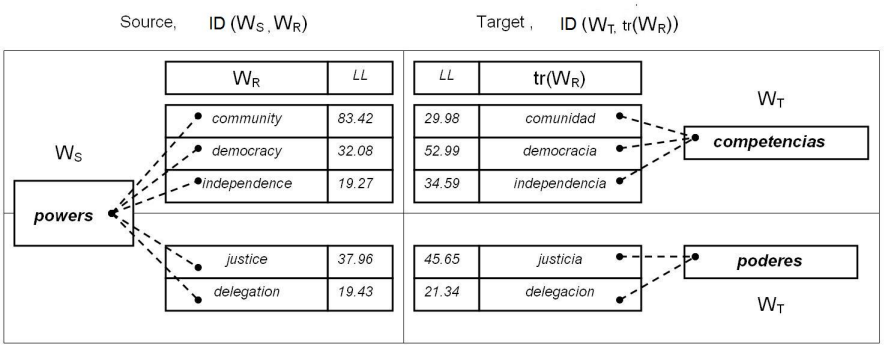

Figure 2: An example of English-Spanish lexicon learnt for the source word *powers*. On the top, the system suggested *competencias* and rejected *poderes* when *powers* is associated with *community*, *democracy* or *independence*. The word *poderes* is suggested when *powers* is associated with *justice* or *delegation*.

We observe that these common contextual terms are simultaneously the *first-order* and *second-order* context terms of the target word. They are the *shared* context terms of the target word and its highly associated context term. We define these terms as *in-domain terms*. These in-domain terms can be used to map words to their corresponding translations. The highly associated context terms can be thought of as sense discriminators that differentiate the different uses of the target word. In Figure 2, we show how *delegation* helps in selecting between the "control or influence" sense of *powers* while rejecting the "ability or skill" sense.

In this paper, our focus is not on sense disambiguation and we follow current evaluation methods for bilingual lexicon extraction. However, it is clear that our method can be adapted for building sense disambiguated bilingual lexicons.

3.1 Identifying highly associated words

To identify the context terms $CT(W_S)$ of a source word W_S, as in (Rapp, 1999), we use *log-likelihood ratio* (LL) Dunning (1993). We choose all words with $LL > t_1$ where t_1 is a threshold.

The *highly associated words* then are the top k highest ranked context terms. In our experiments, we only choose the top 100 highest ranked context terms as our highly associated terms.

In order to compute the log-likelihood ratio of target word a to co-occur with context word b, we create a contingency table. The contingency table contains the observed values taken from a given corpus. An example of the contingency table is shown in Table 1.

C[i,j]	***community***	***¬ community***		
powers	124	1831	**1955**	*C(powers)*
¬ powers	11779	460218	**471997**	*C(¬ powers)*
	11903	**462049**		
	C(community)	*C(¬ community)*		

Here $C[i, j]$ denotes the count of the number of sentences in which i co-occurs with j.
Total corpus size: N = 473952 in the above

Table 1: Contingency table for observed values of target word *powers* and context word *community*.

The LL value of a target word a and context word b is given by:

$$LL(a, b) =$$

$$\sum_{i \in \{a, \neg a\}, j \in \{b, \neg b\}} 2C(i, j) \log \frac{C(i, j)N}{C(i)\, C(j)}$$

3.1.1 Identifying in-domain terms

In our work, to find the translation equivalent of a source word W_S, we do not use the context terms $CT(W_S)$. Instead, we use the *in-domain terms* $IDT(W_S, W_R)$. For each highly associated term

W_R, we get different in-domain terms. Furthermore, $IDT(W_S, W_R)$ is a subset of $CT(W_S)$.

The in-domain terms of W_S given the context terms W_R is given by:

$$ID(W_S, W_R) = CT(W_S) \cap CT(W_R)$$

Programme and *public* are some of the examples of in-domain terms of *powers* given *community* as the highly associated term.

3.1.2 Finding translations pairs

Note that $ID(W_S, W_R)$ is an in-domain term vector in the source language. Let W_T be a potential translation equivalent for W_S. Let, $tr(W_R)$ be a translation equivalent for W_R. Let $ID(W_T, tr(W_R))$ be an in-domain term vector in the target language.

We use $tr(W_S|W_R)$ to denote the translation proposed for W_S given the highly associated term W_R. We compute $tr(W_S|W_R)$ using:

$$tr(W_S|W_R) = \underset{W_T}{\operatorname{argmax}}\, sim(ID(W_S, W_R), ID(W_T, tr(W_R)))$$

Our method learns translation pairs that are conditioned on highly associated words (W_R). Table 2 provides a sample of English-Spanish lexicon learnt for the word *power* with different W_R.

English		Spanish		Sim
W_S	W_R	$tr(W_R)$	W_T	
powers	community	comunidad	**competencias**	**0.9876**
			poderes	0.9744
			independiente	0.9501
	democracy	democracia	**competencias**	**0.9948**
			poderes	0.9915
			independiente	0.9483
	independence	independencia	**competencias**	**0.9939**
			poderes	0.9745
			independiente	0.9633
	justice	justicia	**poderes**	**0.9922**
			competencias	0.3450
			independiente	0.9296
	delegation	delegacion	**poderes**	**0.9568**
			competencias	0.9266
			independiente	0.8408

Table 2: A sample of translation equivalents learnt for *powers*.

In the next section, we introduce a similarity measure that operates on the context vectors in the source language and the target language without requiring a seed dictionary.

4 Rank-binning similarity measure

Most existing methods for computing similarity cannot be directly employed for measuring the similarity between in-domain term context vectors since each context vector is in a different language. A bilingual dictionary can be assumed but that greatly diminishes the practicality of the method.

We address this by making an assumption. We assume that the relative distributions of in-domain context terms of translation equivalent pairs are roughly comparable in the source language and in the target language. For example, consider the log-likelihood values of the in-domain terms for the translation pair *agreement-acuerdo* (conditioned on the highly associated term *association-associacion*) given in Figure 3. We note that the distribution of in-domain terms are comparable although not identical. Thus, the distribution can be used as a clue to derive translation pairs but we need a method to compute similarity of the vector of in-domain terms.

Rank-binning or rank histograms are usually used as a diagnostic tool to evaluate the spread of an ensemble rather than as a verification method. Wong (2009) use the method of rank-binning to roughly examine performance of a system on learning lightweight ontologies. We apply the rank-binning procedure for measuring the similarity of word pairs.

Pre-processing step:

1. Let W_S be a source language word and $x_1, x_2, ..., x_n$ be the set of n context terms ranked in descending log-likelihood values of W_S (see Table 3).

2. We transform the rank values of context terms x_k into the range [0,1] using:

$$z_k = \frac{rank(x_k) - 1}{n - 1}$$

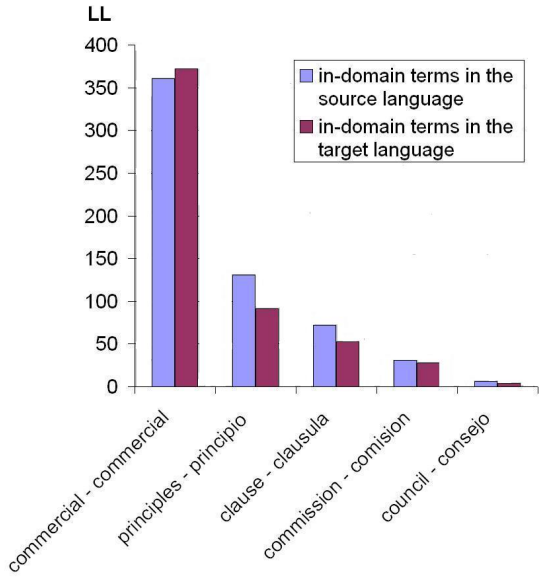

Figure 3: Similar distribution of in-domain terms for *agreement* with *association* and *acuerdo* with *asociacion*.

Binning procedure

We divide the interval $[0, 1]$ into g bins[1] of equal length. Let $b_1, \ldots, b_g$ denote the g bins. Then we map the in-domain terms vector $ID(W_S, W_R)$ into the binned vector $b_1, \ldots, b_g$. For each $x_k \in ID(W_S, W_R)$, this mapping is done by using the corresponding z_k from the pre-processing step. For each bin, we count the number of different in-domain terms that are mapped into this bin. Thus, if the range of the first bin b_1 is [0, 0.009] then *european, legislative, parliament* are mapped into b_1 i.e. $b_1 = 3$. The bins are normalised by dividing with $\mid ID(W_S, W_R) \mid$.

Rank binning similarity

We use Euclidean distance to compute similarity between bins. Given, bins $P = p_1, \ldots, p_g$ and $Q = q_1, \ldots, q_g$, the Euclidean distance is given by:

$$dist(P, Q) = \sqrt{\sum_{i=1}^{g} (p_i, q_i)^2}$$

[1]We used the following formula to estimate the number of bins:

$$g = 1 + 3.3 * log(\mid ID(W_S, W_R) \mid)$$

$CT(\textbf{\textit{powers}})$			
Context term	LL	$rank$	z_k
european	491.33	1	0.00000
legislative	482.19	2	0.00406
parliament	408.26	3	0.00813
:	:	:	:
:	:	:	:
:	:	:	:
public	16.96	245	0.99186
programme	15.40	246	0.99593
representatives	15.32	247	1.00000
$n = 247$			

Table 3: Some examples of transformed values of each term in $CT(powers)$.

In the next section, we describe the setup including the data, the lexicon and the evaluation used in our experiments.

5 Experimental setup

5.1 Data

For comparable text, we derive English and Spanish distinct sentences from the Europarl parallel corpora. We split the corpora into three parts according to year. We used about 500k sentences for each language in the experiments. This approach is further explained in Ismail and Manandhar (2009) and is similar to Koehn and Knight (2001) and Haghighi et al. (2008).

5.2 Pre-processing

For corpus pre-processing, we use sentence boundary detection and tokenization on the raw text before we clean the tags and filter stop words. We sort and rank words in the text according to their frequencies. For each of these words, we compute their context term log-likelihood values.

5.3 Lexicon

In the experiment, a bilingual lexicon is required for evaluation. We extract our evaluation lexicon from the `Word Reference`[2] free online dictionary. This extracted bilingual lexicon has low coverage.

[2]`http://wordreference.com`

5.4 Evaluation

In the experiments, we considered the task of building a bilingual English-Spanish lexicon between the 2000 high frequency source and target words, where we required each individual word to have at least a hundred highly associated context terms that are not part of the initial seed dictionary. Different highly associated W_R terms for a given W_T might derive similar (W_S, W_T) pairs. In this case, we only considered one of the (W_S, W_T) pairs. In future work, we would like to keep these for word sense discrimination purposes. Note that we only considered proposed translation pairs whose similarity values are above a threshold t_2.

We used the F_1 measure to evaluate the proposed lexicon against the evaluation lexicon. If either W_S or W_T in the proposed translation pairs is not in the evaluation lexicon, we considered the translation pairs as unknown, although the proposed translation pairs are correct. $Recall$ is defined as the proportion of the proposed lexicon divided by the size of the lexicon and $precision$ is given by the number of correct translation pairs at a certain recall value.

6 Experiments

In this section, we look into how the in-domain context vectors affect system performance. We also examine the potential of rank-binning similarity measure.

6.1 From standard context vector to in-domain context vector

Most research in bilingual lexicon extraction so far has employed the standard context vector approach. In order to explore the potential of the in-domain context vectors, we compare the systems that use in-domain approach against systems that use the standard approach. We also employ different sets of seed lexicon in each system to be used in the similarity measure:

- Lex_{700}: contains 700 cognate pairs from a few Learning Spanish Cognate websites[3].

- Lex_{100}: contains 100 bilingual entries of the most frequent words in the source corpus that have translation equivalents in the extracted evaluation lexicon. We select the top one hundred words in the source corpus, so that their translation equivalents is within the first 2000 high frequency words in the target corpus.

- Lex_{160}: contains words with similar spelling that occur in both corpora. We used 160 word pairs with an edit distance value less than 2, where each word is longer than 4 characters.

Models using the standard approach are denoted according to the size of the particular lexicon used in their context similarity measure, i.e. *CV-100* for using Lex_{100}, *CV-160* for using Lex_{160} and *CV-700* for using Lex_{700}. We use *IDT* to denote our model. We use lexicon sizes to distinguish the different variants, e.g. *IDT-CV100* for using Lex_{100}, *IDT-CV160* for using Lex_{160} and *IDT-CV700* for using Lex_{700}.

With *CV-700*, the system achieved 52.6 percent of the best F_1 score. Using the same seed dictionary, the best F_1 score has increased about 20 percent points with *IDT-CV700* recorded 73.1 percent. *IDT-CV100* recorded about 15.0 percent higher best F_1 score than *CV-100* with 80.9 and 66.4 percent respectively. Using an automatically derived seed dictionary, *IDT-CV160* yielded 70.0 percent of best F_1 score while *CV-160* achieved 62.4 percent. Results in Table 4 shows various precisions p_x at recall values *x*.

Model	$P_{0.10}$	$P_{0.25}$	$P_{0.33}$	$P_{0.50}$	$BestF_1score$
CV-700	58.3	61.2	64.8	55.2	52.6
CV-100	52.0	53.0	47.2	44.8	66.4
CV-160	68.5	56.8	48.8	48.8	62.4
IDT-CV700	83.3	90.2	82.0	66.7	73.1
IDT-CV100	80.0	75.8	66.7	69.4	80.9
IDT-CV160	90.0	80.6	73.9	69.2	70.0

Table 4: Performance of different models.

[3]such as `http://www.colorincolorado.org` and `http://www.language-learning-advisor.com`

6.2 Similarity measure using rank-binning

We use *RB* to denote our model based on the rank-binning approach. Running *RB* means that no seed dictionary is involved in the similarity measure. We also ran the similarity measure in the *IDT* (*IDT-RB160*) by employing the derived Lex_{160} for the in-domain steps.

We ran several tests using *IDT-RB160* with different numbers of bins. The results are illustrated in Figure 4. The *IDT-RB160* yielded 63.7 percent of best F_1 score with 4 bins. However, the F_1 score starts to drop from 61.1 to 53.0 percent with 6 and 8 bins respectively. With 3 and 2 bins the *IDT-RB160* yielded 63.7 and 62.0 percent of best F_1 score respectively. Using 1 bin is not be possible as all values fall under one bin. Thus, the rank-binning similarity measure for the rest of the experiments where *RB* is mentioned, refers to a 4 bins setting.

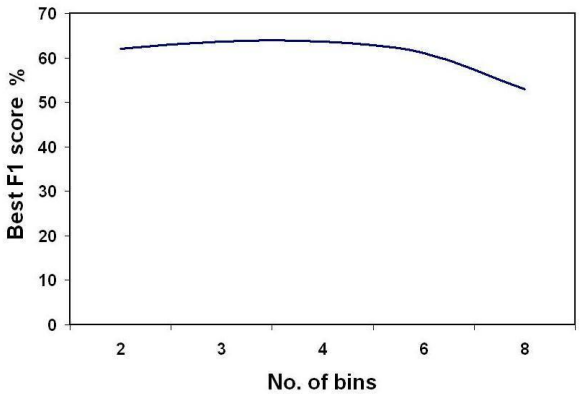

Figure 4: Performance of *IDT-RB160* with different numbers of bins.

While systems using the standard context similarity measure yielded scores higher than 50.0 percent of best F_1, the *RB* achieved only 39.2 percent. However, *RB* does not employ an initial dictionary and does not use orthographic features. As mentioned above, the system scored higher when the similarity measure was used in the *IDT* (i.e. *IDT-RB160*). Note that Lex_{160} is derived automatically so the approach can also be considered as unsupervised. The system performance is slightly lower compared to the conventional *CV-160*. However, *IDT-CV160* outperforms both of the systems (see Figure 5).

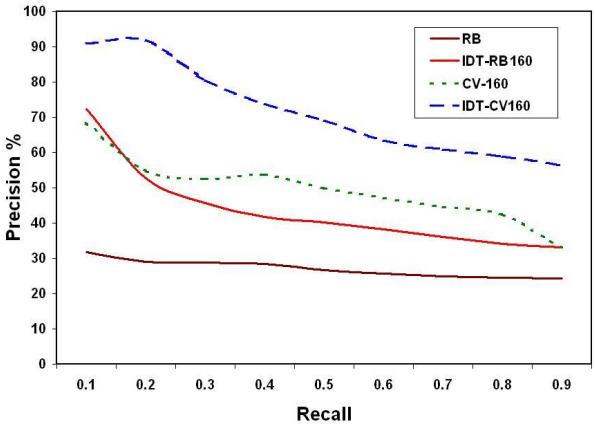

Figure 5: Performance of different unsupervised models.

Overall, systems that exploit in-domain terms yielded higher F_1 scores compared to the conventional context vector approach.

6.3 Comparison with CCA

Previous work in extracting bilingual lexicons from comparable corpora generally employ the conventional context vector approach. Haghighi et al. (2008) focused on applying *canonical correlation analysis* (*CCA*), a dimension reduction technique, to improve the method. They were using smaller comparable corpora, taken from the first 50k sentences of English Europarl and the second 50k of Spanish Europarl, and different initial seed dictionary. Hence, we tested *CCA* in our experimental setup. In *CV-700* setting, using *CCA* yields 57.5 percent of the best F_1 score compared to 73.1 percent of the best F_1 score with *IDT* that we reported in Section 6.2.

7 Discussion

7.1 Potential of in-domain terms

Our experiments clearly demonstrate that the use of in-domain terms achieves higher F_1 scores compared to conventional methods. It also shows that our method improves upon earlier reported dimension reduction methods. From our observation, the number of incorrect translation pairs

were further reduced when the context terms were filtered. Recall that the in-domain terms in the target language were actually the shared context terms of the target word and its highly associated context terms. Nevertheless, this approach actually depends on the initial bilingual lexicon in order to translate those highly associated context terms into the source language. Table 5 shows some examples of most confidence translation pairs proposed by the *IDT-CV100*.

English	Spanish	Sim score	Correct?
principle	principio	0.9999	Yes
government	estado	0.9999	No
government	gobierno	0.9999	Yes
resources	recursos	0.9999	Yes
difficult	dificil	0.9999	Yes
sector	competencia	0.9998	No
sector	sector	0.9998	Yes
programme	programa	0.9998	Yes
programme	comunidad	0.9998	No
agreement	acuerdo	0.9998	Yes

Table 5: Some examples of most confident translation pairs proposed by *IDT-CV100* ranked by similarity scores.

7.2 Seed dictionary variation

The initial seed dictionary plays a major role in extracting bilingual lexicon from comparable corpora. There are a few different ways for us to derive a seed dictionary. Recall that Lex_{700} and Lex_{100}, that are used in the experiments, are derived using different methods. The F_1 scores of the system using Lex_{100} were much higher compared to the system using Lex_{700}. Thus, extending Lex_{100} with additional high frequency words may provide higher accuracy.

One important reason is that all bilingual entries in Lex_{100} occur frequently in the corpora. Although the size of Lex_{700} is larger, it is not surprising that most of the words never occur in the corpora, such as *volleyball* and *romantic*. However, using Lex_{160} is more interesting since it is derived automatically from the corpora, though one should realize that the relationship between the language pair used in the respective monolingual corpora, English and Spanish, may have largely affect the results. Thus, for other systems involving unrelated language pairs, the rank-binning similarity measure might be a good option.

7.3 Word sense discrimination ability

As mentioned in Section 5.4, each source word may have more than one highly associated context term, W_R. Different W_R may suggest different target words for the same source word. For example, given the source word *powers* and the highly associated word *community*, *competencias* is proposed as the best translation equivalent. On the other hand, for same source word *powers*, when the highly associated word is *delegation*, the target word *poderes* is suggested.

8 Conclusion

We have developed a method to improve the F_1 score in extracting bilingual lexicon from comparable corpora by exploiting in-domain terms. This method also performs well without using an initial seed dictionary. More interestingly, our work reveals the potential of building word sense disambiguated lexicons.

References

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL 2008*, Columbus, Ohio.

Azniah Ismail and Suresh Manandhar. 2009. Utilizing contextually relevant terms in bilingual lexicon extraction In *Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, Boulder, Colorado.

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

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, Boston, Massachusetts, 173-183.

Philipp Koehn and Kevin Knight. 2001. Knowledge sources for word-level translation models. In *Proceedings of the Conference on empirical method in natural language processing (EMNLP).*

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL 2002*, Philadelphia, USA, 9-16.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the ACL 33*, 320-322.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the ACL 37*, 519-526.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistic*, volume 19(1), 61-74.

Wilson Yiksen Wong. 2009. *Learning lightweight ontologies from text across different domains using the web as background knowledge*. Ph.D. Thesis. University of Western Australia