Utilizing Contextually Relevant Terms in Bilingual Lexicon Extraction

Azniah Ismail
Department of Computer Science
University of York
York YO10 5DD UK
azniah@cs.york.ac.uk

Suresh Manandhar
Department of Computer Science
University of York
York YO10 5DD UK
suresh@cs.york.ac.uk

Abstract

This paper demonstrates one efficient technique in extracting bilingual word pairs from non-parallel but comparable corpora. Instead of using the common approach of taking high frequency words to build up the initial bilingual lexicon, we show contextually relevant terms that co-occur with cognate pairs can be efficiently utilized to build a bilingual dictionary. The result shows that our models using this technique have significant improvement over baseline models especially when highest-ranked translation candidate per word is considered.

1 Introduction

Bilingual lexicons or dictionaries are invaluable knowledge resources for language processing tasks. The compilation of such bilingual lexicons remains as a substantial issue to linguistic fields. In general practice, many linguists and translators spend huge amounts of money and effort to compile this type of knowledge resources either manually, semi-automatically or automatically. Thus, obtaining the data is expensive.

In this paper, we demonstrate a technique that utilizes contextually relevant terms that co-occur with cognate pairs to expand an initial bilingual lexicon. We use unannotated resources that are freely available such as English-Spanish Europarl corpus (Koehn, 2005) and another different set of cognate pairs as seed words.

We show that this technique is able to achieve high precision score for bilingual lexicon extracted from non-parallel but comparable corpora. Our model using this technique with spelling similarity approach obtains 85.4 percent precision at 50.0 percent recall. Precision of 79.0 percent at 50.0 percent recall is recorded when using this technique with context similarity approach. Furthermore, by using a string edit-distance vs. precision curve, we also reveal that the latter model is able to capture words efficiently compared to a baseline model.

Section 2 is dedicated to mention some of the related works. In Section 3, the technique that we used is explained. Section 4 describes our experimental setup followed by the evaluation results in Section 5. Discussion and conclusion are in Section 6 and 7 respectively.

2 Related Work

Koehn and Knight (2002) describe few potential clues that may help in extracting bilingual lexicon from two monolingual corpora such as identical words, similar spelling, and similar context features. In reporting our work, we treat both identical word pairs and similar spelling word pairs as *cognate pairs*.

Koehn and Knight (2002) map 976 identical word pairs that are found in their two monolingual German-English corpora and report that 88.0 percent of them are correct. They propose to restrict the word length, at least of length 6, to increase the accuracy of the collected word pairs. Koehn and Knight (2002) mention few related works that use different measurement to compute the similarity, such as longest common subsequence ratio (Melamed, 1995) and string edit distance (Mann

Proceedings of the NAACL HLT Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics, pages 10–17,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

and Yarowski, 2001). However, Koehn and Knight (2002) point out that majority of their word pairs do not show much resemblance at all since they use German-English language pair. Haghighi et al. (2008) mention one disadvantage of using edit distance, that is, precision quickly degrades with higher recall. Instead, they propose assigning a feature to each substring of length of three or less for each word.

For approaches based on contextual features or context similarity, we assume that for a word that occurs in a certain context, its translation equivalent also occurs in equivalent contexts. Contextual features are the frequency counts of context words occurring in the surrounding of target word *W*. A context vector for each *W* is then constructed, with only context words found in the seed lexicon. The context vectors are then translated into the target language before their similarity is measured.

Fung and Yee (1998) point out that not only the number of common words in context gives some similarity clue to a word and its translation, but the actual ranking of the context word frequencies also provides important clue to the similarity between a bilingual word pair. This fact has motivated Fung and Yee (1998) to use *tfidf* weighting to compute the vectors. This idea is similar to Rapp (1999) who proposed to transform all co-occurrence vectors using *log likelihood ratio* instead of just using the frequency counts of the co-occurrences. These values are used to define whether the context words are highly associated with the *W* or not.

Earlier work relies on a large bilingual dictionary as their seed lexicon (Rapp, 1999; Fung and Yee, 1998; among others). Koehn and Knight (2002) present one interesting idea of using extracted cognate pairs from corpus as the seed words in order to alleviate the need of huge, initial bilingual lexicon. Haghighi et al. (2008), amongst a few others, propose using canonical correlation analysis to reduce the dimension. Haghighi et al (2008) only use a small-sized bilingual lexicon containing 100 word pairs as seed lexicon. They obtain 89.0 percent precision at 33.0 percent recall for their English-Spanish induction with best feature set, using topically similar but non-parallel corpora.

3 The Utilizing Technique

Most works in bilingual lexicon extraction use lists of high frequency words that are obtained from source and target language corpus to be their source and target word lists respectively. In our work, we aim to extract a high precision bilingual lexicon using different approach. Instead, we use list of contextually relevant terms that co-occur with cognate pairs.

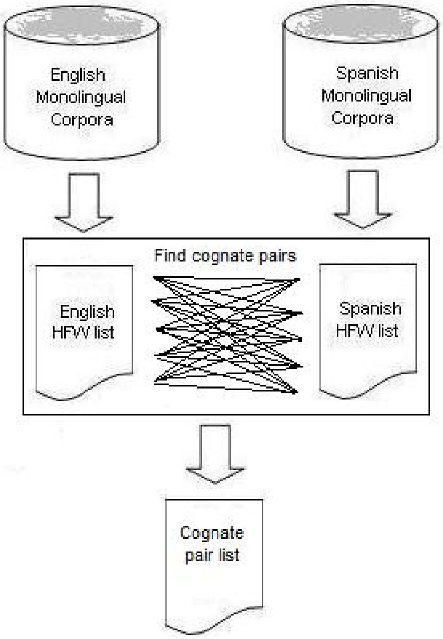


Figure 1: Cognate pair extraction

These cognate pairs can be derived automatically by mapping or finding identical words occur in two high frequency list of two monolingual corpora (see Figure 1). They are used to acquire list of source word W_s and target word W_t. W_s and W_t are contextually relevant terms that highly co-occur with the cognate pairs in the same context. Thus, log likelihood measure can be used to identify them.

Next, bilingual word pairs are extracted among words in these W_s and W_t list using either context similarity or spelling similarity. Figure 2 shows some examples of potential bilingual word pairs, of W_s and W_t, co-occurring with identical cognate pairs of word '*civil*'.

As we are working on English-Spanish language pair, we extract bilingual lexicon using string edit distance to identify spelling similarity between W_s

and W_t. Figure 3 outlines the algorithm using spelling similarity in more detail.

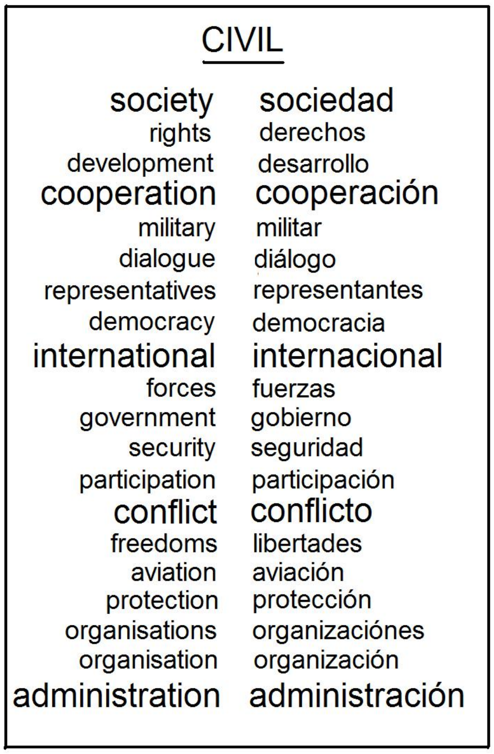
Using the same W_s and W_t lists, we extract bilingual lexicon by computing the context similarity between each $\{W_s,W_t\}$ pair. To identify the context similarity, the relation between each $\{W_s, W_t\}$ pair can be detected automatically using a vector similarity measure such as *cosine measure* as in (1). The *A* and *B* are the elements in the context vectors, containing either zero or non-zero seed word values for W_s and W_t, respectively.

$$Cosine \text{ similarity} = cos(\theta) = \frac{A \times B}{||A|| \times ||B||} \quad (1)$$

The cosine measure favors $\{W_s,W_t\}$ pairs that share the most number of non-zero seed word values. However, one disadvantage of this measure is that the cosine value directly proportional to the actual W_s and W_t values. Even though W_s and W_t might not closely correlated with the same set of seed words, the matching score could be high if W_s or W_t has high seed word values everywhere. Thus, we transform the context vectors from real value into binary vectors before the similarity is computed. Figure 4 outlines the algorithm using context similarity in more detail.

In the algorithm, after the W_s and W_t lists are obtained, each W_s and W_t units is represented by their context vector containing log likelihood (LL) values of contextually relevant words, occurring in the seed lexicon, that highly co-occur with the W_s and W_t respectively. To get this context vector, for each W_s and W_t, all sentences in the English or Spanish corpora containing the respective word are extracted to form a particular sub corpus, e.g. sub corpus *society* is a collection of sentences containing the source word *society*.

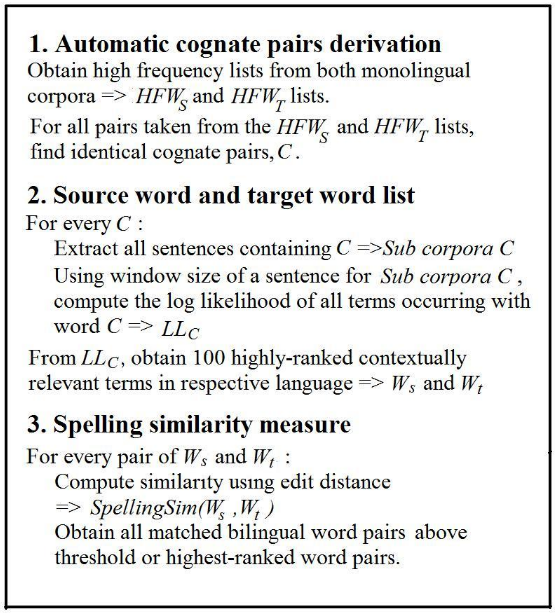
Using window size of a sentence, the LL value of term occurring with the word W_s or W_t in their respective sub corpora is computed. Term that is highly associated with the W_s or W_t is called contextually relevant term. However, we consider each term with LL value higher than certain threshold (e.g. *threshold* $\geq$ 15.0) to be contextually relevant. Contextually relevant terms occurring in the seed lexicon are used to build the context vector for the



CIVIL

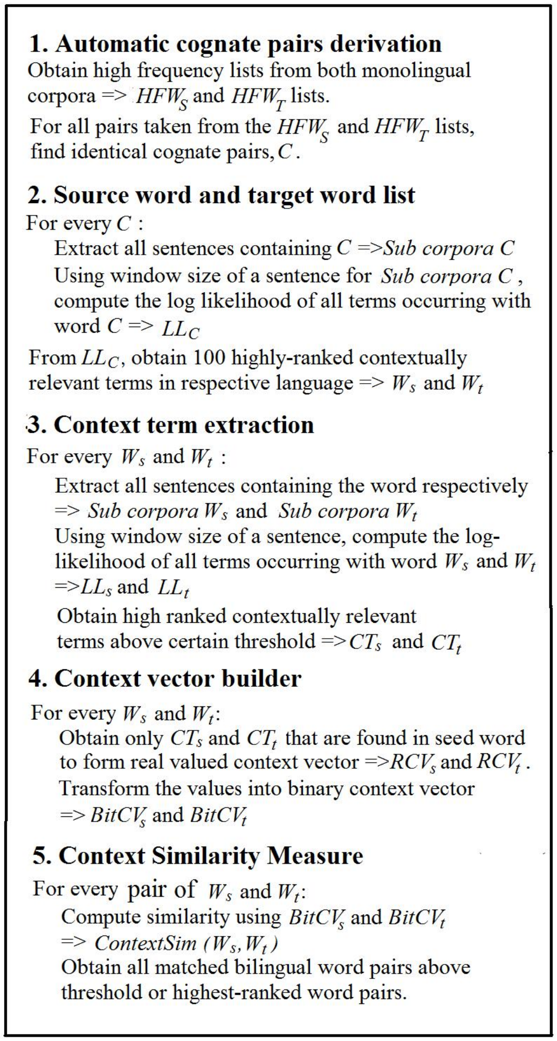
society	sociedad
rights	derechos
development	desarrollo
cooperation	cooperación
military	militar
dialogue	diálogo
representatives	representantes
democracy	democracia
international	internacional
forces	fuerzas
government	gobierno
security	seguridad
participation	participación
conflict	conflicto
freedoms	libertades
aviation	aviación
protection	protección
organisations	organizaciónes
organisation	organización
administration	administración

Figure 2: Bilingual word pairs are found within context of cognate word *civil*



```
1. Automatic cognate pairs derivation
Obtain high frequency lists from both monolingual
corpora => $HFW_S$ and $HFW_T$ lists.
For all pairs taken from the $HFW_S$ and $HFW_T$ lists,
find identical cognate pairs, $C$.
2. Source word and target word list
For every $C$ :
    Extract all sentences containing $C$ =>Sub corpora $C$
    Using window size of a sentence for Sub corpora $C$ ,
    compute the log likelihood of all terms occurring with
    word $C$ => $LL_C$
From $LL_C$, obtain 100 highly-ranked contextually
relevant terms in respective language => $W_s$ and $W_t$
3. Spelling similarity measure
For every pair of $W_s$ and $W_t$ :
    Compute similarity using edit distance
    => SpellingSim($W_s$ ,$W_t$ )
    Obtain all matched bilingual word pairs above
    threshold or highest-ranked word pairs.
```

Figure 3: Utilizing technique with spelling similarity



```
1. Automatic cognate pairs derivation
Obtain high frequency lists from both monolingual
corpora => $HFW_S$ and $HFW_T$ lists.
For all pairs taken from the $HFW_S$ and $HFW_T$ lists,
find identical cognate pairs, $C$.
2. Source word and target word list
For every $C$ :
    Extract all sentences containing $C$ =>Sub corpora $C$
    Using window size of a sentence for Sub corpora $C$ ,
    compute the log likelihood of all terms occurring with
    word $C$ => $LL_C$
From $LL_C$, obtain 100 highly-ranked contextually
relevant terms in respective language => $W_s$ and $W_t$
3. Context term extraction
For every $W_s$ and $W_t$ :
    Extract all sentences containing the word respectively
    => Sub corpora $W_s$ and Sub corpora $W_t$
    Using window size of a sentence, compute the log-
    likelihood of all terms occurring with word $W_s$ and $W_t$
    =>$LL_s$ and $LL_t$
    Obtain high ranked contextually relevant
    terms above certain threshold =>$CT_s$ and $CT_t$
4. Context vector builder
For every $W_s$ and $W_t$:
    Obtain only $CT_s$ and $CT_t$ that are found in seed word
    to form real valued context vector =>$RCV_s$ and $RCV_t$ .
    Transform the values into binary context vector
    => $BitCV_s$ and $BitCV_t$
5. Context Similarity Measure
For every pair of $W_s$ and $W_t$:
    Compute similarity using $BitCV_s$ and $BitCV_t$
    => ContextSim ($W_s, W_t$)
    Obtain all matched bilingual word pairs above
    threshold or highest-ranked word pairs.
```

Figure 4: Utilizing technique with context similarity

W_s or W_t respectively. For example, word *participation* and *education* occurring in the seed lexicon are contextually relevant terms for source word *society*. Thus, they become elements of the context vector. Then, we transform the context vectors, from real value into binary, before we compute the similarity with cosine measure.

4 Experimental Setup

4.1 Data

For source and target monolingual corpus, we derive English and Spanish sentences from parallel Europarl corpora (Koehn, 2005).

- We split each of them into three parts; year 1996 - 1999, year 2000 - 2003 and year 2004 - 2006.
- We only take the first part, about 400k sentences of Europarl Spanish (year 1996 - 1999) and 2nd part, also about 400k from Europarl English (year 2000 - 2003). We refer the particular part taken from the source language corpus as *S* and the other part of the target language corpus as *T*.

This approach is quite common in order to obtain non-parallel but comparable (or same domain) corpus. Examples can be found in Fung and Cheung (2004), followed by Haghighi et al. (2008). For corpus pre-processing, we only use sentence boundary detection and tokenization on raw text. We decided that large quantities of raw text requiring minimum processing could also be considered as minimal since they are inexpensive and not limited. These should contribute to low or medium density languages for which annotated resources are limited. We also clean all tags and filter out stop words from the corpus.

4.2 Evaluation

We extracted our evaluation lexicon from *Word Reference*[*] free online dictionary . For this work, the word types are not restricted but mostly are content words. We have two sets of evaluation. In one, we take high ranked candidate pairs where W_s could have multiple translations. In the other, we only consider highest-ranked W_t for each W_s. For evaluation purposes, we take only the top 2000 candidate ranked-pairs from the output. From that list, only candidate pairs with words found in the evaluation lexicon are proposed. We use F1-measure to evaluate proposed lexicon against the evaluation lexicon. The recall is defined as the proportion of the high ranked candidate pairs. The precision is given as the number of correct candidate pairs divided by the total number of proposed candidate pairs.

4.3 Other Setups

The following were also setup and used:

- *List of cognate pairs*

  We obtained 79 identical cognate pairs from the

[*] from website http://www.wordreference.com

top 2000 high frequency lists of our *S* and *T* but we chose 55 of these that have at least 100 contextually relevant terms that are highly associated with each of them.

- *Seed lexicon*
  We also take a set of cognate pairs to be our seed lexicon. We defined the size of a small seed lexicon ranges between 100 to 1k word pairs. Hence, our seed lexicon containing 700 cognate pairs are still considered as a small-sized seed lexicon. However, instead of acquiring this set of cognate pairs automatically, we compiled the cognate pairs from a few *Learning Spanish Cognates* websites [†]. This approach is a simple alternative to replace the 10-20k general dictionaries (Rapp, 1999; Fung and McKeown, 2004) or automatic seed words (Koehn and Knight, 2002; Haghighi et al., 2008). However, this approach can only be used if the source and target language are fairly related and both share lexically similar words that most likely have same meaning. Otherwise, we have to rely on general bilingual dictionaries.

- *Stop list*
  Previously (Rapp, 1999; Koehn and Knight, 2002; among others) suggested filtering out commonly occurring words that do not help in processing natural language data. This idea sometimes seem as a negative approach to the natural articles of language, however various studies have proven that it is sensible to do so.

- *Baseline system*
  We build baseline systems using basic context similarity and spelling similarity features.

5 Evaluation Results

For the first evaluation, candidate pairs are ranked after being measured either with cosine for context similarity or edit distance for spelling similarity. In this evaluation, we take the first 2000 of $\{W_s, W_t\}$ candidate pairs from the proposed lexicon where W_s may have multiple translations or multiple W_t. See Table 1.

Setting	**P**$_0$.1	**P**$_0$.25	**P**$_0$.33	**P**$_0$.5	**Best-F1**
ContextSim (***CS***)	42.9	69.6	60.7	58.7	49.6
SpellingSim (***SS***)	90.5	74.2	69.9	64.6	50.9

(a) from baseline models

Setting	**P**$_0$.1	**P**$_0$.25	**P**$_0$.33	**P**$_0$.5	**Best-F1**
E-ContextSim (***ECS***)	78.3	73.5	71.8	64.0	51.2
E-SpellingSim (***ESS***)	95.8	75.6	71.8	63.4	51.5

(b) from our proposed models

Table 1: Performance of baseline and our model for top 2000 candidates below certain threshold and ranked

Setting	**P**$_0$.1	**P**$_0$.25	**P**$_0$.33	**P**$_0$.5	**Best-F1**
ContextSim-Top1 (***CST***)	58.3	61.2	64.8	55.2	52.6
SpellingSim-Top1 (***SST***)	84.9	66.4	52.7	34.5	37.0

(a) from baseline models

Setting	**P**$_0$.1	**P**$_0$.25	**P**$_0$.33	**P**$_0$.5	**Best-F1**
E-ContextSim-Top1 (***ECST***)	85.0	81.1	79.7	79.0	57.1
E-SpellingSim-Top1 (***ESST***)	100.0	93.6	91.6	85.4	59.0

(b) from our proposed models

Table 2: Performance of baseline and our model for top 2000 candidates of top 1

Using either context or spelling similarity approach on *S* and *T* (labeled *ECS* and *ESS* respectively), our models achieved about 51.2 percent of best F1 measure. Those are not a significant improvement with only 1.0 to 2.0 percent error reduction over the baseline models (labeled *CS* and *SS*).

For the second evaluation, we take the first 2000 of $\{W_s, W_t\}$ pairs where W_s may only have the highest ranked W_t as translation candidates (See Table 2). This time, both of our models (with context similarity and spelling similarity, labeled *ECST* and *ESST* respectively) yielded almost 60.0 percent of best F1 measure. It is noted that using *ESST* alone recorded a significant improvement of 20.0 percent in the F1 score compared to *SST* baseline model. *ESST* obtained 85.4 percent precision at 50.0 percent recall. Precision of 79.0 percent at 50.0 percent recall is recorded when using *ECST*. However, the *ECST* has not recorded a significant difference over *CST* baseline model (57.1 and 52.6 percent respectively) in the second evaluation. The overall performances, represented by precision scores for different

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

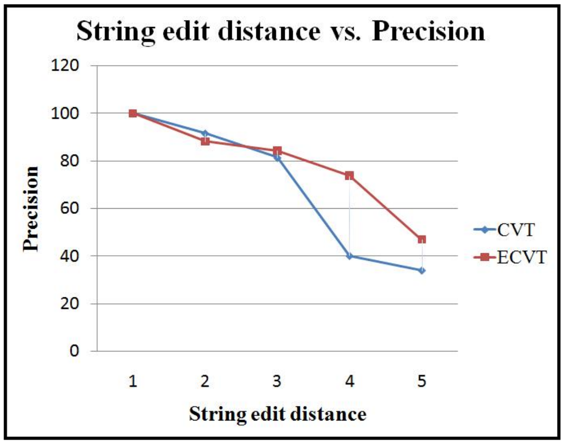


Figure 5: String Edit Distance vs. Precision curve

range of recalls, for these four models are illustrated in *Appendix A*.

It is important to see the inner performance of the *ECST* model with further analysis. We present a string edit distance value (*EDv*) vs. precision curve for *ECST* and *CST* in Figure 5 to measure the performance of the *ECST* model in capturing bilingual pairs with less similar orthographic features, those that may not be captured using spelling similarity.

The graph in Figure 5 shows that even though *CST* has higher precision score than *ECST* at *EDv* of 2, it is not significant (the difference is less than 5.0 percent) and the spelling is still similar. On the other hand, precision for proposed lexicon with *EDv* above 3 (where the W_s and the proposed translation equivalent W_t spelling becoming more dissimilar) using *ECST* is higher than *CST*. The most significant difference of the precision is almost 35.0 percent, where *ECST* achieved almost 75.0 percent precision compared to *CST* with 40.0 percent precision at *EDv* of 4. It is followed by *ECST* with almost 50.0 percent precision compared to *CST* with precision less than 35.0 percent, offering about 15.0 percent precision improvement at *EDv* of 5.

6 Discussion

As we are working on English-Spanish language pair, we could have focused on spelling similarity feature only. Performance of the model using this feature usually record higher accuracy otherwise they may not be commonly occurring in a corpus. Our models with this particular feature have recorded higher F1 scores especially when considering only the highest-ranked candidates.

We also experiment with context similarity approach. We would like to see how far this approach helps to add to the candidate scores from our corpus *S* and *T*. The other reason is sometimes a correct target is not always a cognate even though a cognate for it is available. Our *ECST* model has not recorded significant improvement over *CST* baseline model in the F1-measure. However, we were able to show that by utilizing contextually relevant terms, *ECST* gathers more correct candidate pairs especially when it comes to words with dissimilar spelling. This means that *ECST* is able to add more to the candidate scores compared to *CST*. Thus, more correct translation pairs can be expected with a good combination of *ECST* and *ESST*.

The following are the advantages of our utilizing technique:

- Reduced errors, hence able to improve precision scores.

- Extraction is more efficient in the contextual boundaries (see Appendix B for examples).

- Context similarity approach within our technique has a potential to add more to the candidate scores.

Yet, our attempt using cognate pairs as seed words is more appropriate for language pairs that share large number of cognates or similar spelling words with same meaning. Otherwise, one may have to rely on bilingual dictionaries.

There may be some possible supporting strategies, which we could use to help improve further the precision score within the utilizing technique. For example, dimension reduction using canonical correlation analysis (CCA), resemblance detection, measure of dispersion, reference corpus and further noise reduction. However, we do not include a reranking method, as we are using collection of cognate pairs instead of a general bilingual dictionary. Since our corpus *S* and *T* is in similar domain, we might still not have seen the potential of this technique in its entirety. One may want to test the technique with different type of corpora for future works.

Nevertheless, we are still concerned that many spurious translation equivalents were proposed because the words actually have higher correlation with the input source word compared to the real target word. Otherwise, the translation equivalents may not be in the boundaries or in the corpus from which translation equivalents are to be extracted. Haghighi et al (2008) have reported that the most common errors detected in their analysis on top 100 errors were from semantically related words, which had strong context feature correlations. Thus, the issue remains. We leave all these for further discussion in future works.

7 Conclusion

We present a bilingual lexicon extraction technique that utilizes contextually relevant terms that cooccur with cognate pairs to expand an initial bilingual lexicon. We show that this utilizing technique is able to achieve high precision score for bilingual lexicon extracted from non-parallel but comparable corpora. We demonstrate this technique using unannotated resources that are freely available.

Our model using this technique with spelling similarity obtains 85.4 percent precision at 50.0 percent recall. Precision of 79.0 percent at 50.0 percent recall is recorded when using this technique with context similarity approach. We also reveal that the latter model with context similarity is able to capture words efficiently compared to a baseline model. Thus, we show contextually relevant terms that cooccur with cognate pairs can be efficiently utilized to build a bilingual dictionary.

References

Cranias, L., Papageorgiou, H, and Piperidis, S. 1994. A matching technique in Example-Based Machine Translation. *In International Conference On Computational Linguistics Proceedings*, 15th conference on Computational linguistics, Kyoto, Japan.

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

Fung, P., and Cheung, P. 2004. Mining very non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. *In Proceedings of the 2004 Conference on Empirical Method in Natural Language Processing (EMNLP)*, Barcelona, Spain.

Fung, P., and Yee, L.Y. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. *In Proceedings of COLING-ACL98*, Montreal, Canada, 1998.

Fung, P., and McKeown, K. 1997. Finding Terminology Translations from Non-parallel Corpora. *In The 5th Annual Workshop on Very Large Corpora*, Hong Kong, Aug 1997.

Haghighi, A., Liang, P., Berg-Krikpatrick, T., and Klein, D. 2008. Learning bilingual lexicons from monolingual corpora. *In Proceedings of The ACL 2008*, June 15 -20 2008, Columbus, Ohio

Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. *In MT Summit*

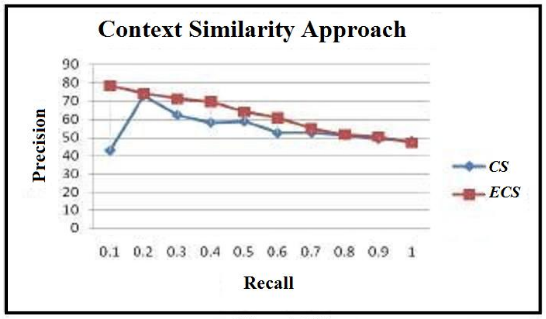
Koehn, P., and Knight , K. 2001. Knowledge sources for word-level translation models. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Koehn, P., and Knight , K. 2002. Learning a translation lexicon from monolingual corpora. *In Proceedings of ACL 2002*, July 2002, Philadelphia, USA, pp. 9-16.

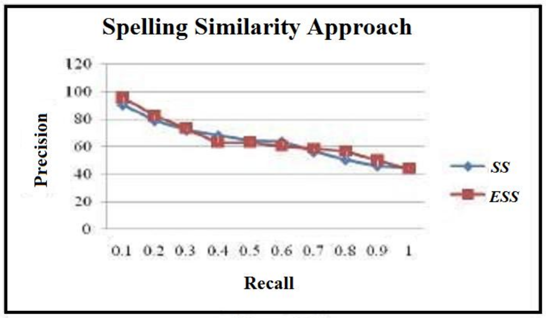
Rapp, R. 1995. Identifying word translations in non-parallel texts. *In Proceedings of ACL 33*, pages 320-322.

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

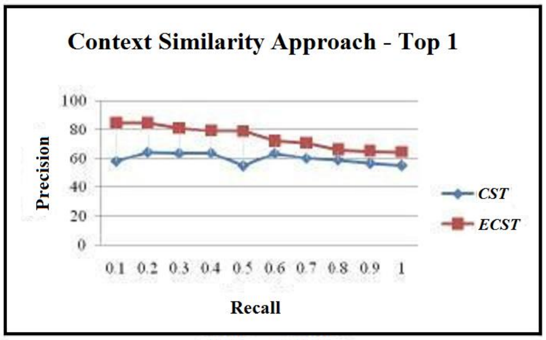
Appendix A. Precision scores with different recalls



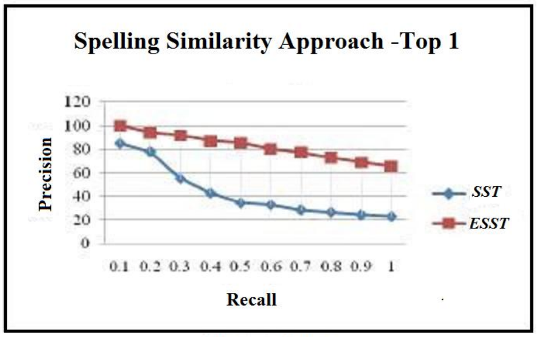
CS vs. ECS



SS vs. ESS



CST vs. ECST



SST vs. ESST

Appendix B. Some examples of effective extraction via utilizing technique

Source	Target	ECVT		CV		Rank
		Candidate found	Sim. value	Candidate found	Sim. value	
clause	*clausula*	***clausula***	***0.402015126***	*autentica* *fortalecimiento* *economico* *respeto* *vigor* ***clausula***	*0.447213595* *0.430331483* *0.412478956* *0.40824829* *0.402015126* ***0.402015126***	1 2 < > < > < > < >
pillar	*pilar*	***pilar***	***0.547722558***	*daramente* ***pilar*** *basada* *comercial* *iniciado* *exterior* *agricola*	*0.632455532* ***0.547722558*** *0.53935989* *0.516397779* *0.516397779* *0.478091444* *0.447213595*	1 2 3 4 4 5 6
state	*estado*	***estado***	***0.433012702***	*derecho* ***estado*** *respeto*	*0.43519414* ***0.433012702*** *0.412478956*	1 2 < >
confidence	*confianza*	***confianza***	***0.424264069***	*errores* *desarrollo* *haberse* *demuestran* *deficiencias* ***confianza***	*0.447213595* *0.447213595* *0.447213595* *0.447213595* *0.447213595* ***0.424264069***	1 1 1 1 1 2
welfare	*bienestar*	***bienestar***	***0.40824829***	*hubiera* ***bienestar***	*0.500000000* ***0.40824829***	1 < >