Utilizing contextually relevant terms in bilingual lexicon extraction

Azniah Ismail & Suresh Manandhar

5 June 2009

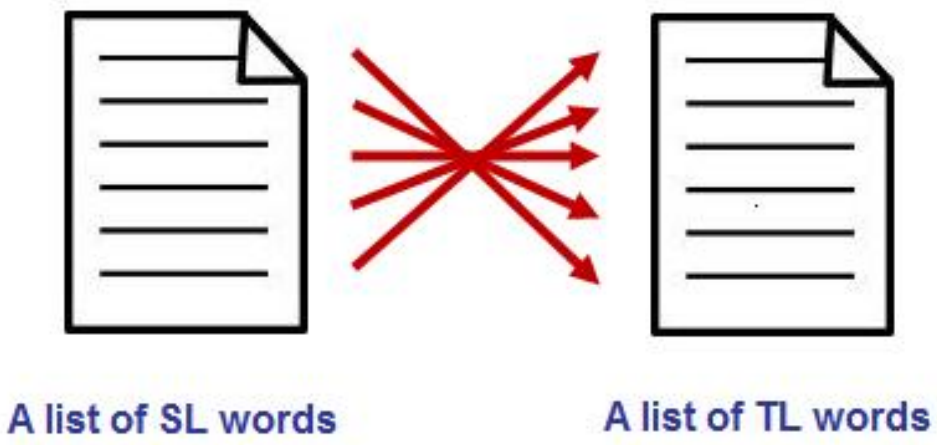
NAACL-2009 Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics

Boulder, Colorado, USA

Table of contents

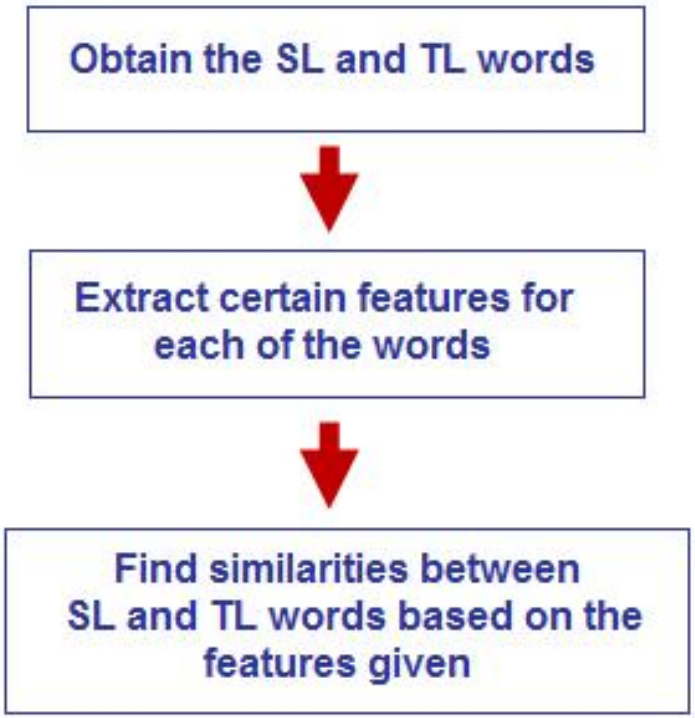
Introduction

Bilingual lexicon extraction (BLE) involves a matching process between a set of source language (SL) words with a set of target language (TL) words occurring in respective SL and TL corpora.



Introduction

General steps in a BLE method:



Introduction

Koehn and Knight (2002) describes a few clues that can be used for the features such as:

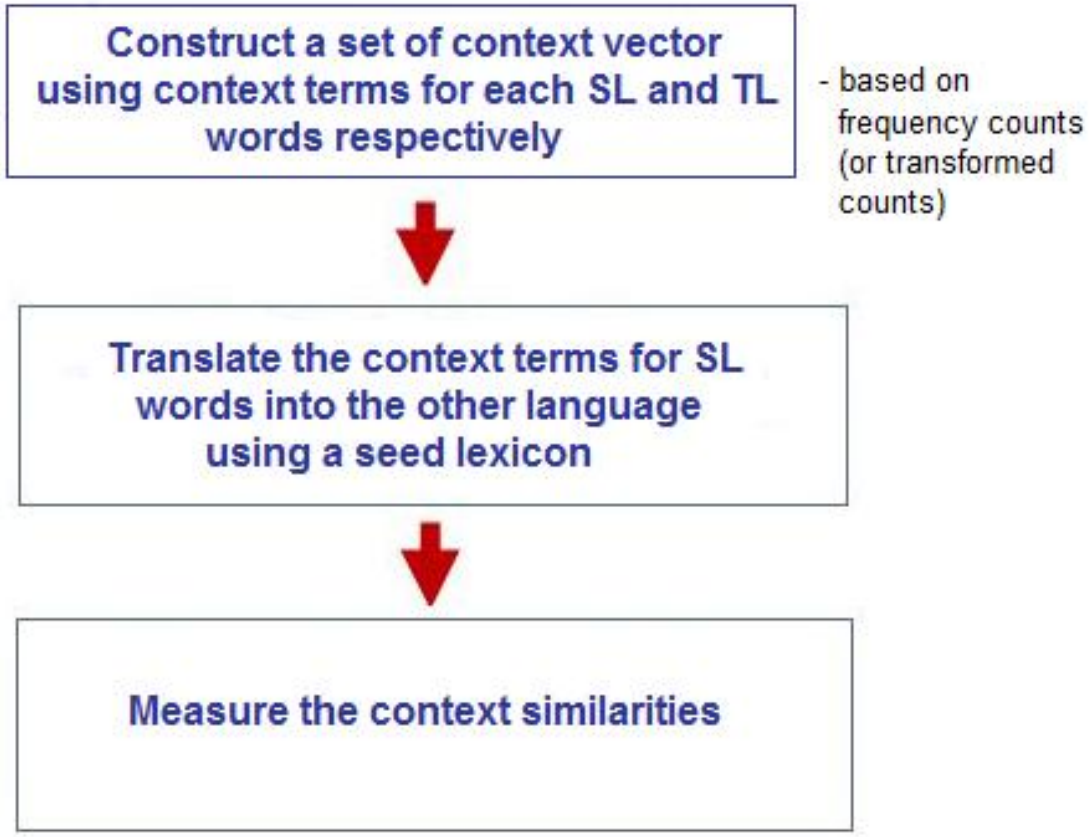
- context feature
- identical/ similar spelling feature

Related Work - Context feature approach

A hypothesis in machine translation:

Assume that for a word that occurs in a certain context, its translation equivalent also occurs in equivalent or similar contexts.

Related Work - Context feature approach



Related Work - Context feature approach

English:

..... various resolutions dealing with **civil** society that it has adopted during the current parliamentary term .....

....workshop held by the Committee on Constitutional Affairs with representatives of **civil** society organisations. ...

..... close cooperation between the EU institutions and Member States and **civil** society .....

Spanish:

..... el gobierno cubano y la sociedad **civil** durante las visitas .....

..... Grupo de la Sociedad **civil** reúne a más de 40 funcionarios .....

..... específico de justicia **civil** en el marco .....

Related Work - Context feature approach

- Clues that may show the similarity of a bilingual word pair:
  - common words occur in similar context
  - the actual ranking of the context word frequencies
    - Fung and Yee (1998) use *tfidf* weighting to compute the vectors.
    - Rapp (1999) proposed to transform all co-occurrence vectors using *log likelihood ratio*.
- These values are then used to define whether the context words are highly associated with the target word or not.
- The precision of the existing methods varies from 35.0 percent to 72.0 percent.
- The precision seems to improve when one requires the input words to have high occurrence frequencies in the corpus.

Related Work - Similar spelling feature approach

- Spelling similarity between word pairs can be computed by using:
  - string edit distance (Mann and Yarowski, 2001)
  - longest common subsequence ratio (Melamed, 1995)
- Koehn and Knight (2002) map 976 identical German-English word pairs - 88.0 percent correct.
  - propose to restrict the word length, at least of length 6, to increase the accuracy of the collected word pairs.
  - point out that majority of their German-English word pairs do not show much resemblance at all.

Related Work - Similar spelling feature approach

- Disadvantage of string edit distance: precision quickly degrades with higher recall.
  - Haghighi et al. propose assigning a feature to each substring of length of three or less for each word.
- Disadvantages of the approach:
  - Usually record higher accuracy only for related language pairs.
  - Sometimes a correct target is not always a cognate even though a cognate for it is available.

The issue, idea and technique

- When comparable corpora or non-parallel corpora are used, only a number of SL words might have their correspondence translations.
  - The matching process may be prone to errors.
  - Mapping all words in one corpus to the other may introduce lots of noise.
- Taking only high frequency words may seems to improve the precision however we might have missed some high precision word pairs of smaller frequency words.

The issue, idea and technique

Aim: To extract a high precision bilingual lexicon from comparable corpora.

- How?

The issue, idea and technique

Aim: To extract a high precision bilingual lexicon from comparable corpora.

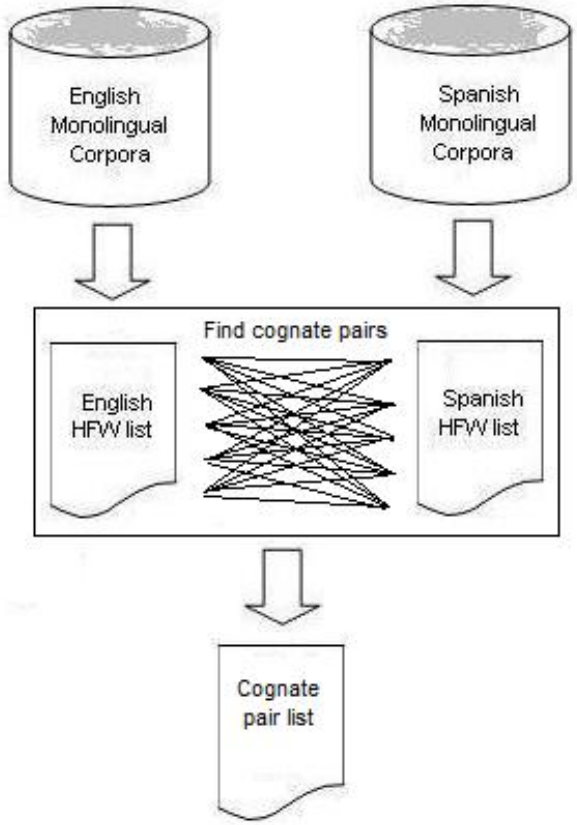
- How? By utilizing contextual relevant words.
- Method:
  - Select a bilingual word pair of translation equivalent.
  - Find context words that highly related to each of the words in their respective monolingual corpora.
  - Use the respective context words as the set of SL words and the set of TL words to be matched in the BLE.
- Possible advantages:
  - We may have higher precision bilingual lexicon extraction because the boundaries of source and target sets are restricted.
  - However, the idea is not only we could be selective with the SL and TL word and reduce the errors, but the words do not even have to be a high frequency word.

The issue, idea and technique

How do we get the initial bilingual word pairs that define the boundaries?

The issue, idea and technique

How do we get the initial bilingual word pairs that define the boundaries?
A set of word pairs can be derived automatically by mapping or finding identical words occur in two high frequency list of two monolingual corpora (Koehn and Knight, 2002)

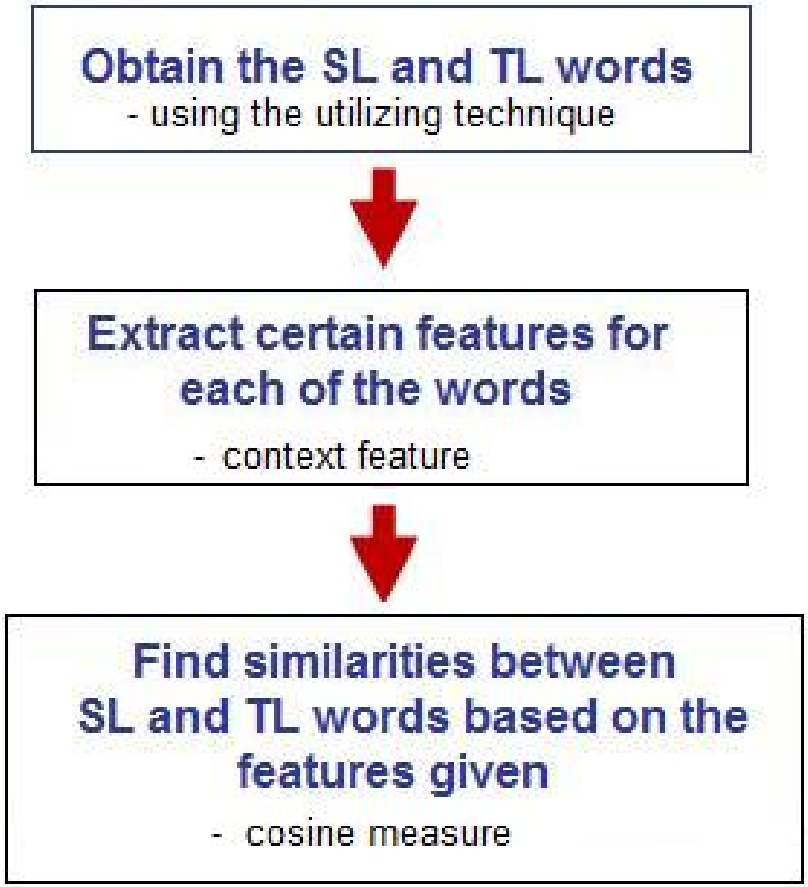


Cognate pair extraction

The issue, idea and technique

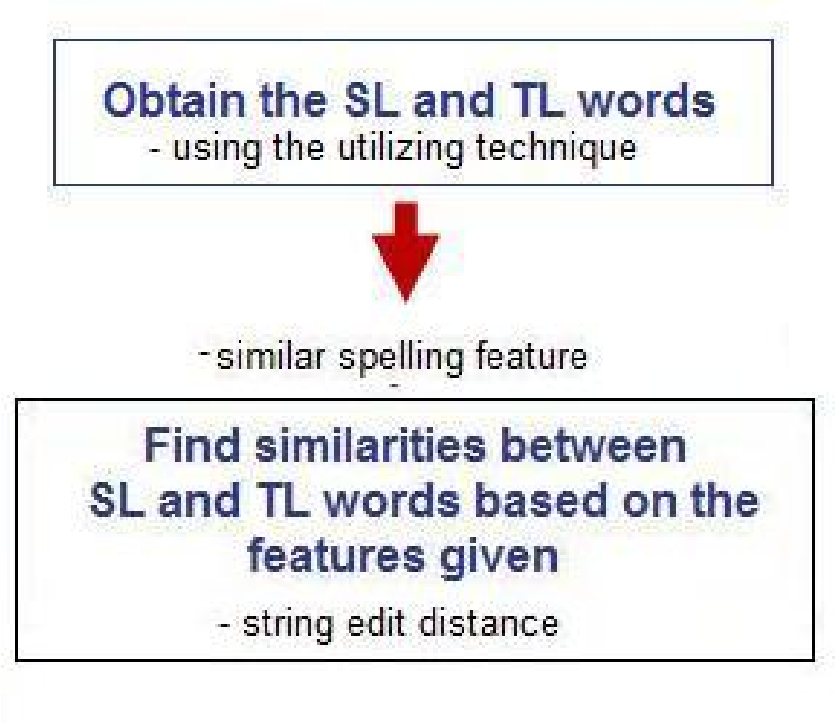
Some examples of English and Spanish words that are contextually relevant and highly co-occur with the cognate pair 'civil'-'civil'.

The issue, idea and technique



Proposed technique with context similarity approach

The issue, idea and technique

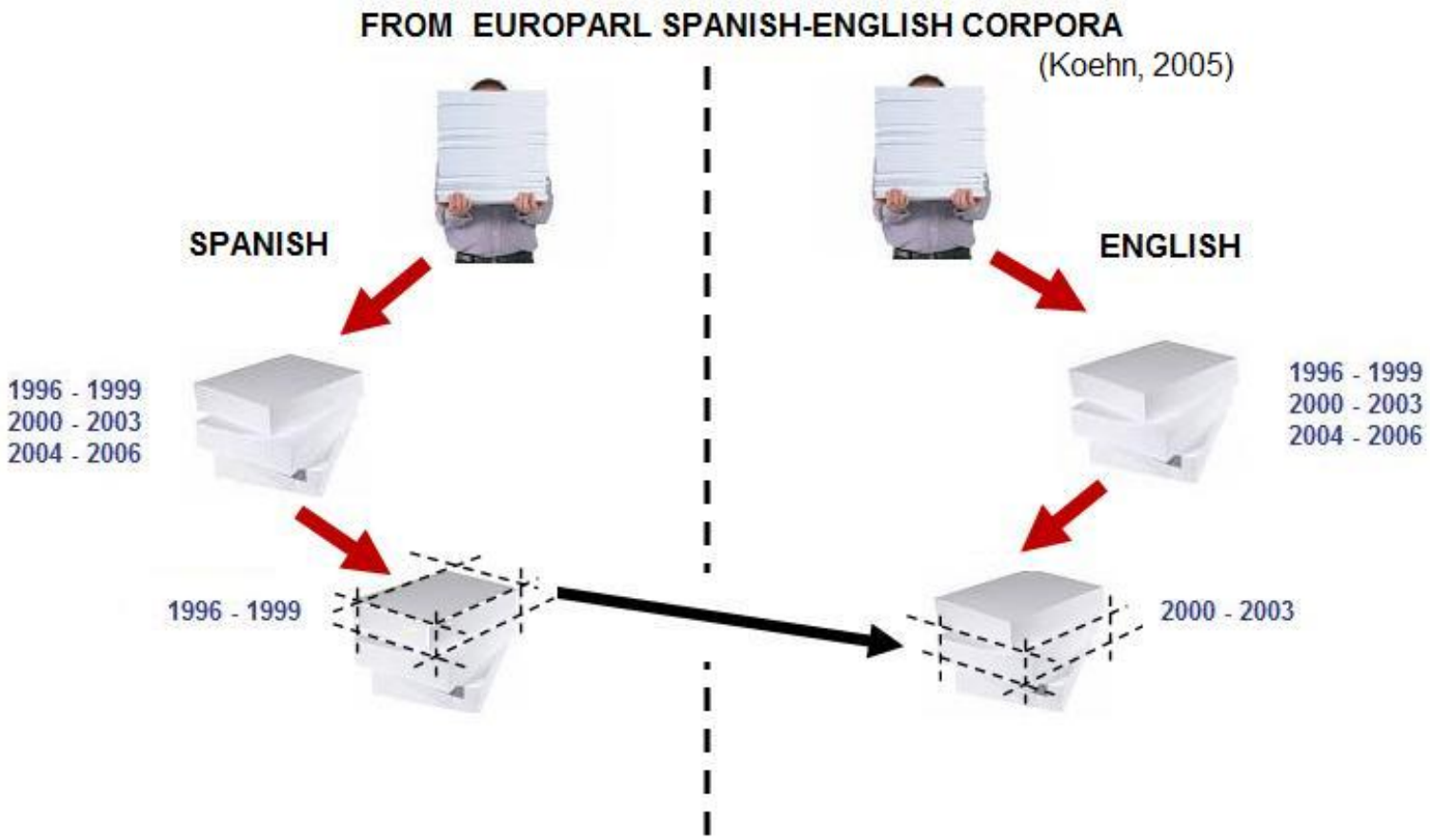


Proposed technique with spelling similarity approach

Experimental setup

- Data
- List of cognate pairs
- Seed lexicon
- Stop list
- Evaluation
- Baseline system

Experimental setup: Data



NOTE: This approach is quite common in order to obtain non-parallel but comparable corpus (Fung and Cheung, 2004; Haghighi et al., 2008).

Experimental setup: Data

- Corpus pre-processing includes the use of language processing tools on raw text such as:
  - a sentence detector
  - a tokenizer
- Further pre-processing also involves stop words and tags removal.

Experimental setup: List of cognate pairs

- Using the cognate pair extraction method, 79 identical cognate pairs are obtained from the top 2000 high frequency lists of respective S and T corpora.
- However, only 55 of them were chosen, that have at least 100 contextually relevant terms that are highly associated with each of them.

Experimental setup: Seed lexicon

- 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) only use a small-sized bilingual lexicon containing 100 word pairs as seed lexicon. They also propose using *canonical correlation analysis* to reduce the dimension.

Experimental setup: Seed lexicon

- *A set of cognate pairs as the seed lexicon*
  Instead of acquiring this set of cognate pairs automatically using cognate pair extraction method, the cognate pairs were compiled from a few *Learning Spanish Cognates* websites:
  - `http://www.colorincolorado.org`
  - `http://www.language-learning-advisor.com`.
- *Size of the seed lexicon*
  Using such approach, we easily compiled 700 cognate pairs. As we define the size of a small seed lexicon is to range between 100 to 1k word pairs, our seed lexicon containing the 700 cognate pairs are still considered as a small-sized seed lexicon.

Experimental setup: Seed lexicon

NOTE:

- This approach is a simple alternative to replace the 10-20k general dictionaries of Fung and McKeown (1997) and Rapp(1999) or automatic seed words as in Koehn and Knight (2002) and 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.

Experimental setup: Stop list

- Previously, Rapp (1999), and Koehn and Knight (2002) among others, suggest 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.

Experimental setup: Evaluation

For evaluation purposes, we only consider top 2000 candidate ranked-pairs from the output. From that list, only candidate pairs with words found in an evaluation lexicon are proposed.

- **F1-measure** is used 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.

Experimental setup: Evaluation

Two sets of evaluation:

- *Evaluation I*: Consider high-ranked candidate pairs where target word may have multiple translations.
- *Evaluation II*: Consider only highest-ranked candidate pairs where target word may only have single translation.

The evaluation lexicon is extracted from a free online dictionary website `http://www.wordreference.com`. For this work, the word types are not restricted but mostly are content words.

Experimental setup: Baseline system

The baseline systems are built based on two basic features:

- context similarity
  - Basic context similarity (CS) - for Evaluation I
  - Basic context similarity - top 1 (CST) - for Evaluation II
- spelling similarity
  - Basic spelling similarity (SS) - for Evaluation I
  - Basic spelling similarity - top 1 (SST) - for Evaluation II

NOTE: The only difference between baseline and our models is the way we obtain the SL and TL words.

Evaluation result

Evaluation I:

Setting	$\mathbf{P_0}.1$	$\mathbf{P_0}.25$	$\mathbf{P_0}.33$	$\mathbf{P_0}.5$	**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	$\mathbf{P_0}.1$	$\mathbf{P_0}.25$	$\mathbf{P_0}.33$	$\mathbf{P_0}.5$	**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 the proposed models

Performance of baseline and the proposed model for top 2000 candidates below certain threshold and ranked

Evaluation result

Evaluation II:

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

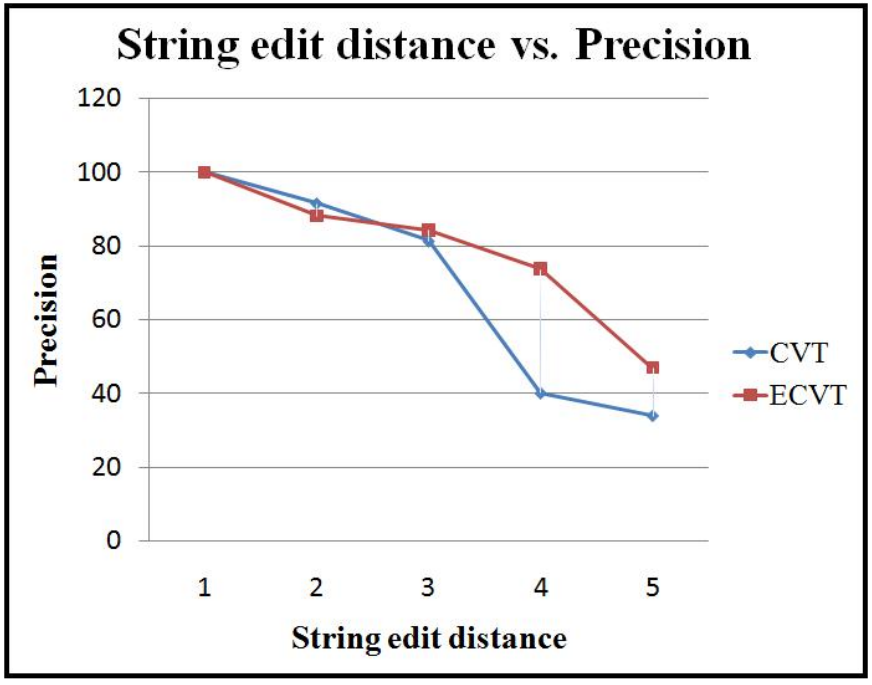
Setting	**P$_0$.1**	**P$_0$.25**	**P$_0$.33**	**P$_0$.5**	**Best-F1**
E-ContextSim-Top1 (ECST)	85.0	81.1	79.7	79.0	57.1
E-SpellingSim-Top1 (ESST)	100.0	93.6	91.6	85.4	59.0

(b) from the proposed models

Performance of the baseline model and the proposed model for top 2000 candidates of top 1

Evaluation result

Performance of our context similarity model (of top 1) in capturing bilingual pairs with less similar orthographic features:



- The baseline context similarity model (of top 1) has higher precision score than our proposed model at edit distance value of 2, but it is not significant and the spelling is still similar.
- On the other hand, the precision for proposed lexicon with the value above 3 using our model of top 1 is significantly higher than the baseline.

Some examples of output

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

Advantages & Disadvantages

- Reduced errors, hence able to improve precision scores.
- Extraction is more efficient in the contextual boundaries.
- Context similarity approach within the technique has a potential to add more to the candidate scores.

Yet, the 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.

Conclusion

We present a bilingual lexicon extraction technique that utilizes contextually relevant terms that co-occur with cognate pairs to expand an initial bilingual lexicon.

- We demonstrate this technique using unannotated resources that are freely available.
- The bilingual lexicon is extracted from non-parallel but comparable corpora.

Conclusion

- 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.
- 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 co-occur with cognate pairs can be efficiently utilized to build a bilingual dictionary.