Resource-poor Translation of Multiword Expressions

Shiva Taslimipoor and Ruslan Mitkov

University of Wolverhampton, Wolverhampton, UK
shiva.taslimi@wlv.ac.uk, r.mitkov@wlv.ac.uk

BACKGROUND

- The correct identification, interpretation and translation of multiword expressions (MWE) is vital for the successful operation of most NLP applications and computer-aided tools that support various users including language learners, teachers, translators, interpreters and terminologists.
- The cross-lingual analysis of these expressions and automatic extraction of their translation equivalents is still an under-research topic.
- Previous corpus-based distributional similarity approaches to discover translation equivalents have not yet reported good results for MWEs.

Difficulties in translations of MWEs

pay attention:	*prestar/poner atención*
pay homage:	*rendir homenja*
pay a compliment:	*decir un cumplido*

OBJECTIVES

- To exploit from comparable corpora in order to find translations for a specific category of MWEs, entitled Verb + Noun expressions.
- To extract bilingual contexts from comparable corpora to find cross-lingual similarities between expressions.
- To propose a new distributional similarity approach based on word embedding in order to find translation-equivalents.

MOTIVATIONS

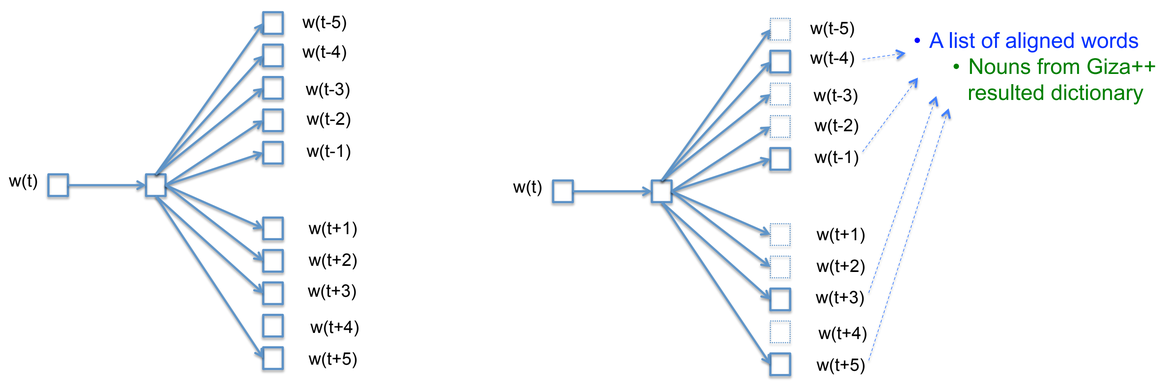
- Distributional similarity hypothesis

earthquake	riot	**take place**	…	problem
terremoto	disturbios	**tener lugar**	…	problema

METHODS

Word Embedding

Word2vec: representing each word as a dense vector derived by various training methods inspired by neural-network language modeling.



Word2vecf: a modified version which allows to define any context

We adapt the model to our task of finding translation equivalents for multi-word collocations, by:

(i) treating sequences of words as single units/terms

(ii) defining bilingual contexts by drawing on a core set of known translation pairs.

RESULTS

Corpora:

Using ACCURAT Toolkit

- Gathering news from the web
- Pairing the documents according to their similarity to have a highly comparable corpora

Experimental Expressions:

- Focusing on Verb+Noun(s) constructed from 9 highly-frequent verbs in English and 6 in Spanish.
- Reporting the results for expressions with frequencies higher than 9 in the paired comparable corpora

Extracting Translation Equivalents:

Given a candidate expression *s* from the source language, the goal is to find the best translation equivalent in the target language using the following algorithm:

For each document *D1* in the source language containing *s*:
For each target language document *D2* paired to *D1*:
Find the most similar expressions in *D2* with *s*

Evaluation:

- Baseline: A simple distributional similarity approach
  - Sets of context pairs
  - Jaccard similarity coefficient to compare the corresponding sets
- Using loosely comparable corpora
  - Computing and comparing the results by adding noisy pairs to our accurately-paired documents
- Human annotation
  - Finding a good translation among the top-5-ranked candidates

Spanish to English translations

	coverage	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
Paired CC	baseline	82%	55%	24%	22%	18%	16%	12%
	word2vec	50%	46%	40%	36%	34%	32%	33%
CC+Noise	baseline	78%	50%	24%	18%	14%	13%	8%
	word2vec	44%	45%	38%	**37%**	30%	**33%**	32%

English to Spanish translations

	coverage	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
Paired CC	baseline	79%	52%	46%	35%	26%	22%	18%
	word2vec	39%	37%	34%	36%	34%	29%	31%
CC+Noise	baseline	70%	50%	24%	22%	18%	12%	13%
	word2vec	38%	34%	31%	**39%**	**39%**	**32%**	31%

Translation accuracies for semantically coherent MWEs

	Accuracy (coverage: 80%)	
	Spanish	English
wodr2vec approach	48%	44%

- These results were presented in (Taslimipoor et al., 2016)

CONCLUSIONS

- A methodology is proposed for extracting cross-lingual contexts from comparable corpora.
- Cross-lingual contexts have been then used to build embedding-based vector representations for MWEs.
- The vectors have been successfully used to find translation equivalents for Verb+Noun combinations between Spanish and English.
- The results show that our approach outperforms a simple distributional similarity baseline.
- It has been also shown that, in contrast to the simple distributional similarity baseline, the word2vec approach is less vulnerable to noise in the corpus.

REFERENCES

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers) (pp. 302-308).

Rapp, R., & Sharoff, S. (2014). Extracting multiword translations from aligned comparable documents. In Proceedings of the 3rd workshop on hybrid approaches to translation (hytra) @ EACL 2014 (p. 83-91).

Taslimipoor, S., Mitkov, R., Corpas Pastor, G., Fazly, A. (2016). Bilingual contexts from comparable corpora to mine for translations of collocations. In proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016). To be published in the Springer LNCS series.