Using Word Embeddings to Translate Named Entities

Octavia-Maria Şulea[1, 2, 3], Sergiu Nisioi[1, 2], Liviu P. Dinu[1, 2]
Faculty of Mathematics and Computer Science, University of Bucharest[1],
Center for Computational Linguistics, University of Bucharest[2],
Bitdefender Romania [3]
14 Academiei Street, Bucharest, Romania[1],
7 Edgar Quinet Street, Bucharest, Romania[2],
24 Delea Veche Street, Bucharest, Romania[3]
mary.octavia@gmail.com, sergiu.nisioi@gmail.com, liviu.p.dinu@gmail.com

Abstract

In this paper we investigate the usefulness of neural word embeddings in the process of translating Named Entities (NEs) from a resource-rich language to a language low on resources relevant to the task at hand, introducing a novel, yet simple way of obtaining bilingual word vectors. Inspired by observations in (Mikolov et al., 2013b), which show that training their word vector model on comparable corpora yields comparable vector space representations of those corpora, reducing the problem of translating words to finding a rotation matrix, and by results in (Zou et al., 2013), which showed that bilingual word embeddings can improve Chinese Named Entity Recognition (NER) and English to Chinese phrase translation, we use the sentence-aligned English-French EuroParl corpora and show that word embeddings extracted from a merged corpus (corpus resulted from the merger of the two aligned corpora) can be used to NE translation. We extrapolate that word embeddings trained on merged parallel corpora are useful in Named Entity Recognition and Translation tasks for resource-poor languages.

Keywords: Named Entity Translation, Corpus Aquisition, Word Embeddings

1. Introduction

Named Entity Recognition (NER) is a complex, Information Extraction subtask, requiring several preprocessing stages (i.e. part-of-speech tagger, tokenizer) which in turn involve dedicated tools. For resource-rich languages, such as English, NER is a highly researched area with the sate-of-the-art system achieving near-human performance: 93% F1 compared to the 97% F1 obtained by human annotators (Marsh and Perzanowski, 1998). For other languages having fewer language processing tools and especially task specific manually annotated data, NER is still a challenging task.

Word embeddings have been recently used as features to improve existing monolingual NER systems ((Katharina Sienčnik, 2015), (Demir and Ozgur, 2014)), or to aid the translation of NEs (Zirikly, 2015). Previous to this, (Shao and Ng, 2004) reported using word embeddings as part of a larger system that extracts named entities from comparable corpora. Others have used alignment models to extract this type of information from parallel datasets (see (Moore, 2003), (Ehrmann and Turchi, 2010)). In addition to parallel or comparable datasets, metadata information, when available, can also prove useful (Ling et al., 2011) for multilingual named entity extraction. Related to multilingual named entities, we note the *transliteration* of NEs given out of context, the decision on whether to transliterate or translate also having been investigated (Mahmoud Mahmoud Azab and Oflazer, 2013). The results of the 2015 ACL shared task on transliteration of named entities[1] revealed that further research is necessary to obtain satisfactory results in this direction.

The closest work to our own is represented by (Zou et al., 2013), which used monolingual and bilingual word embeddings for Chinese NER and English to Chinese phrase translation. Unlike the present study, but similar to other NE projection works (Ehrmann and Turchi, 2010), they required word-level pre-aligned parallel corpora. Our approach also takes hints from (Mikolov et al., 2013b), which showed that two word2vec models trained separately on comparable corpora (i.e. English and Spanish Wikipedia) will yield comparable vector spaces (i.e. there's a linear mapping between them), which in turn will aid in extending dictionaries.

In what follows we present a novel, yet simple approach to train word embeddings in order to extract entity-translation pairs. We focus on two types of entities - locations and organizations. We consider a parallel English-French corpus based on Europarl (Bojar et al., 2015) to train and evaluate our method. In Section 4. we present our results against a machine translation system and against a named entity recognizer trained on French. We show that this technique leads to quantitative improvements over the machine translated entities and it can be used to enhance the quality of the French named entity recognition system.

2. Dataset

We used the French-English set from the Europarl parallel corpus ((Koehn, 2005), (Koehn, 2012)), which was adapted for the 2015 Workshop of Machine Translation (Bojar et al., 2015). The set contains proceedings of the European Parliament (EP), from 1996 to 2011, aligned at the sentence level.

Because Europarl does not have gold standard annotated entities, we used the CoreNLP named entity recognizer

[1] http://www.colips.org/workshop/news2015

(Finkel et al., 2005) to extract locations and organizations. The choice of NE types was due to the domain the dataset belongs to, which most often will contain these two types and will have the person type, for instance, shared between source and target language. Regarding error rates with our NE acquisition strategy, we are aware that the entities discovered this way can also contain erroneous information, yet this is the only option and as such a typical first step in NE projection when manually annotated data lacks in both source and target language. By this approach we attempt to bring into discussion the extension of monolingual NER taggers from languages where it performs very well to languages where the performance is weaker. To compare the English entities extracted with CoreNLP against the French equivalent, we used NERC-fr (Azpeitia et al., 2014) - a named entity recognizer trained on the French ESTER corpus ((Galliano et al., 2009) (Galliano et al., 2014)), which contains annotated and transcribed news speeches. Its training domain suggests that locations and organizations are probably often encountered in the annotated version.

	English	**French**
Types	314,505	154,630
Tokens	50,263,003	59,040,195
Organizations	907,302	284,808
Locations	582,412	430,476
Sentences	2,007,723	

Table 1: Statistics on the English and French parallel corpus.

In Table 1, we render basic statistics of the French-English Europarl corpus[2]. Two important observations arise here. First, the number of types (unique words) within the French corpus is considerably smaller than the English equivalent, but, at the same time, there are 9 million more tokens in the French corpus. This fact is an indicator that the French version is less varied lexically. The second observation is related to the number of entities discovered by CoreNLP in English and by NERC-fr in French. While the number of locations is more or less comparable, the number of organizations is at least three times larger in English than French. A fact that can be attributed to the different standards (between English and French) of writing organizations with uppercase letters which can influence the quality of the French NER tool used.

3. Our Approach

The word embeddings are extracted using the skip-gram model, as introduced in (Mikolov et al., 2013a) and integrated in the gensim Python module (Řehůřek and Sojka, 2010). In addition, we use the Microsoft Bing Translator API[3] to obtain a machine translation of the entities identified on the English corpus.

In order to take advantage of the parallel aspect of our corpora (i.e. that we know apriori which sentence in English is the translation of which sentence in French), we forcefully introduced a high similarity between vectors of words appearing in the same sentence, but different language, by training the word2vec model on the corpus resulted from merging the two parallel corpora, sentence by sentence. More precisely, the merger was done so that, on each line of the resulting corpus, there was the English sentence followed by its French translation. This corpus was stripped of all punctuation marks with the exception of the apostrophe and the upper case letters. The upper case was maintained so that the embeddings model made a better distinction between a NE and its common noun counterpart such as *house* vs *House*, where the latter refers to the people assembled in the European Parliament. The apostrophe was maintained to keep the French articles as part of the words.

Once the merged corpus was obtained and preprocessed, we ran the gensim Phrases model[4], implemented after (Mikolov et al., 2013c), to extract from it word level bigrams, trigrams, and 4-grams. This was done in order to check whether the conclusions in (Passos et al., 2014) related to the usefulness in NER of embeddings trained over phrases instead of words applied to our corpora. The window size w of the training algorithm was decided by the following formula:

$$w = \bar{x} + 2\sigma(x)$$

where $\bar{x}$ is the mean sentence length in words and $\sigma(x)$ is the standard deviation. This gave us a window of approximately 100 words. We used an embedding size of 512 words and we did not restrict the size of the dictionary nor did we prune words that were below a certain frequency. By this, we allowed words that rarely appear (e.g. acronyms) to be taken into account.

The bigram and trigram models were also extracted from the monolingual corpora, where we noticed that multiword NEs that were identified by the Phrases module as frequently occuring bigrams in one corpus were also identified in the other corpus, although the training was done separately, which intuitively is as expected. These merged n-gram corpora were used to train several word2vec models. We then translated each English NE identified by the CoreNLP NER into French using Bing and used these translations as a baseline. Granted, since the EuroParl corpus is not manually annotated for NE, we cannot properly test the accuracy of our model, but some comparisons can be drawn as will be discussed in the following section.

4. Results. Discussion

Table 4. shows a few examples for the first and second results, along with their scores, obtained when applying the *most_similar* function (implemented using cosine distance) to the addition of the vectors for each word in an English NE (as identified by the CoreNLP NER). The vectors here were trained on the unigram corpus (i.e. collocations were not treated as a single word).

[2] The annotated corpora together with the experiments in this paper, available at `http://nlp.unibuc.ro/resources.html`

[3] `http://www.microsoft.com/en-us/translator/translatorapi.aspx`

[4] http://radimrehurek.com/gensim/models/phrases.html

English	1st results	Score	2nd results	Score
Member States	États	0.86	Membres	0.85
Scotland	Écosse	0.87	Wales	0.70
New York	Zealand	0.53	Londres	0.51
London	Londres	0.89	Paris	0.64
Romania	Roumanie	0.88	Bulgaria	0.78

Table 2: Some English to French NE translations using the 1st and 2nd most similar word vectors in the word2vec model and their corresponding similarity scores

We immediately notice very high similarity scores between an English NE and the French words pertaining to the French translation of that NE. The downside to this approach is that NEs that are written identically in both languages (e.g. *New York*) will not have two separate vectors and, therefore, the result of the *most_similar* function will not lead to its correct "translation" (in this case, the result should be equal to the vector given as argument to the function) and, instead, will return the next most similar, which usually is another English word. However, one can argue that this sort of NEs don't actually require a translation and we can thus easily find a solution to this problem by either checking whether the English NE appears in the French corpus, or by using the information that the first and second most similar results are not French words and that their scores are much lower than in the case of French translations of English NEs. We therefore ignored this class and proceeded to look only at the NEs which shared no common word in the two languages.

Since the EuroParl corpus is not annotated for NEs and therefore a clearcut metric such as BLEU cannot be computed, we proceeded to compute how many of the Bing translations were actually in the French corpus and compare that percentage with how many of the NEs translated with our word2vec models were also in. In doing so, we assumed it unlikely for names of organisations or locations, which often times are multiword expressions, would appear in the French text in exactly the order Bing translated them in and as common nouns (i.e. not representing NEs).

With the Bing translation, only 63% of the translated organization NEs and 71% of the translated location NEs actually appeared in the French EuroParl corpus. However, the location dataset was considerably smaller than the oganization one (CoreNLP identified 9188 unique locations and 31537 unique organizations). We also noticed that the organization dataset extracted with CoreNLP incorrectly contained locations such as Barents Sea. Out of all the NEs identified by CoreNLP and translated with Bing, 67% of the locations and 54% of the organizations contained one or more words which were common to both languages. We proceeded to look at unigram and bigram English NEs which did not share any word with their translations (e.g. *European Parliament* vs. *Parlament Européen*).

For the locations dataset, out of the 3029 (43% of the entire dataset) English NEs which did not share any word with their Bing translation, 1251 (41% of unshared NEs and 13% of the entire dataset) were unigrams (e.g. *America* vs. *Amerique*) and 1212 were bigrams (e.g. *United States* vs. *États Unis*). Since these NEs covered roughly 81% of the entire dataset, we focused on them.

Model	# correct 1gram NEs	# correct 2gram NEs
Bing	<1%	1.2%
word2vec	19%	6%

Table 3: Comparison between word2vec trained on the merged corpus and Bing results for Location NEs

Model	# correct 1gram NEs	# correct 2gram NEs
Bing	<1%	1.2%
word2vec	23%	11%

Table 4: Comparison between word2vec trained on the merged corpus and Bing results for Organization NEs

As shown in Table 4., out of the Bing translations of the unshared unigram Location NEs less than 1% were found correct (i.e. were contained in the French corpus) and out of the Bing unshared bigrams, only 1.2% were correct. When using only the first two results from the most_similar function of word2vec trained on the unigram merged corpus, out of the 1251 unigram unshared NEs, 19% of the translations were present in the French corpus in that exact form and, out of the 1212 bigram unshared NEs, 6% of the translations were present.

In the case of the organization NEs (see Table 4.), out of a total of 14654 unshared NEs, 10% were unigram NEs and 4143 (28%) were bigram NEs. Out of the 1499 unshared unigram organization NEs, less than 1% translated by Bing were also present in the French corpus and, out of the 4143 unshared bigram organization NEs, 1.2% translated by Bing were present in the corpus. In the case of word2vec, out of the 1499 unshared unigrams, 23% matched multiword expressions in the French corpus, and, out fo the 4143 unshared bigrams, 11% of the translations resulted in using either the first or second result from most_similar were present in French. In both cases, we see an important improvement when using word2vec.

5. Conclusions and Future Work

In this paper, we have introduced a novel approach of employing word embeddings to create named entity resources for low resourced languages. Our unsupervised machine learning approach requires a parallel corpus consisting of one or more languages and a named entity tagger for one of the languages. This approach can potentially be useful to translate or find multilingual equivalents not only for entities, but also for collocations or multi-word units. Our results show that the approach is more stable under context specific particularities (e.g. House as Assemblée or acronyms) and it can potentially improve tools that already exist for the target language. Overall, the number of entities discovered by word embeddings + CoreNLP for French is larger than the one obtained with NERC-fr trained on a French corpus and the one obtained using a state-of-the-art commercial machine translation system. As future work, we plan to extend our experiments to other less resourced languages and verify these hypotheses further. We are also in the process of developing a manual annotated corpus in order to provide consistent comparisons and additional evaluation.

6. Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback. The work of Octavia-Maria Şulea was initiated while affiliated with Bitdefender Romania and being funded by the strategic grant POSDRU/187/1.5/S/155559. The work of Liviu P. Dinu was supported by UEFISCDI, PNII-ID-PCE-2011-3-0959.

7. Bibliographic References

Azpeitia, A., Cuadros, M., Gaines, S., and Rigau, G. (2014). Nerc-fr: Supervised named entity recognition for french. In Petr Sojka, et al., editors, *Text, Speech and Dialogue - 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings*, volume 8655 of *Lecture Notes in Computer Science*, pages 158–165. Springer.

Ondřej Bojar, et al., editors. (2015). *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, September.

Demir, H. and Ozgur, A. (2014). Improving named entity recognition for morphologically rich languages using word embeddings. In *13th International Conference on Machine Learning and Applications, ICMLA 2014, Detroit, MI, USA, December 3-6, 2014*, pages 117–122. IEEE.

Ehrmann, M. and Turchi, M. (2010). Building multilingual named entity annotated corpora exploiting parallel corpora. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora AEPC 2010*. Northern European Association for Language Technology (NEALT).

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *In ACL*, pages 363–370.

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech*, volume 9, pages 2583–2586.

Katharina Sienčnik, S. (2015). Adapting word2vec to named entity recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 239–243, Vilnius, Lithuania, May. Linköping University Electronic Press, Sweden.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Ling, W., Calado, P., Martins, B., Trancoso, I., Black, A. W., and Coheur, L. (2011). Named entity translation using anchor texts. In *2011 International Workshop on Spoken Language Translation, IWSLT 2011, San Francisco, CA, USA, December 8-9, 2011*, pages 206–213. ISCA.

Mahmoud Mahmoud Azab, Houda Bouamor, B. M. and Oflazer, K. (2013). Dudley North visits North London: Learning When to Transliterate to Arabic. In *Proceedings of NAACL 2013*. ACL, January.

Marsh, E. and Perzanowski, D. (1998). Muc-7 evaluation of ie technology: Overview of results. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998.*

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Moore, R. C. (2003). Learning translations of named-entity phrases from parallel corpora. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 259–266.

Passos, A., Kumar, V., and McCallum, A. (2014). Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 78–86, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Shao, L. and Ng, H. T. (2004). Mining new word translations from comparable corpora. In *Proceedings of the 20th international conference on Computational Linguistics*, page 618. Association for Computational Linguistics.

Zirikly, A. (2015). Cross-lingual transfer of named entity recognizers without parallel corpora. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 390–396. The Association for Computer Linguistics.

Zou, W. Y., Socher, R., Cer, D. M., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Con-*

ference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1393–1398. ACL.

8. Language Resource References

Galliano, Sylvain and Gravier, Guillaume and Chaubard, Laura. (2014). *Evaluation of Broadcast News enriched transcription systems (ESTER)*. ELRA, ESTER Evaluation Package, 1.0, ISLRN 110-079-844-983-7.

Koehn, Philipp. (2012). *EuroParl*. EuroMatrixPlus project funded by the European Comission, Digital Corpus of the European Parliament, English-French, v7, ISLRN 823-807-024-162-2.