Bilingual Word Embeddings for Bilingual Terminology Extraction from Specialized Comparable Corpora

Amir Hazem and Emmanuel Morin

LS2N, Université de Nantes, France
The 8th International Joint Conference on Natural Language Processing (IJCNLP)

November 27- December 1, 2017 Taipei, Taiwan

Outline

1 Context

Outline

1. Context
2. State of art approaches
   - Distributional-Based Approaches
   - Distributed-Based Approaches

Outline

1. Context
2. State of art approaches
   - Distributional-Based Approaches
   - Distributed-Based Approaches
3. Data Combination Using Word Embeddings
   - Global Data Combination
   - Specific Data Combination
   - Combining Distributed Representations

Outline

1. Context
2. State of art approaches
   - Distributional-Based Approaches
   - Distributed-Based Approaches
3. Data Combination Using Word Embeddings
   - Global Data Combination
   - Specific Data Combination
   - Combining Distributed Representations
4. Data and Resources

Outline

1. Context
2. State of art approaches
   - Distributional-Based Approaches
   - Distributed-Based Approaches
3. Data Combination Using Word Embeddings
   - Global Data Combination
   - Specific Data Combination
   - Combining Distributed Representations
4. Data and Resources
5. Experiments and Results

Outline

1. Context
2. State of art approaches
   - Distributional-Based Approaches
   - Distributed-Based Approaches
3. Data Combination Using Word Embeddings
   - Global Data Combination
   - Specific Data Combination
   - Combining Distributed Representations
4. Data and Resources
5. Experiments and Results
6. Conclusion and Perspectives

Bilingual terminology extraction/comparable corpora

- specialized comparable corpora –> specialized texts sharing common features such as domain, genre, sampling period, etc.
- bilingual terminology extraction –> Single word terms translation

Bilingual terminology extraction/comparable corpora

drawbacks

- constrained by the small amount of data
- penalizes the performance of distributional-based approaches

Bilingual terminology extraction/comparable corpora

drawbacks

- constrained by the small amount of data
- penalizes the performance of distributional-based approaches

solutions

- associate external resources with the comparable corpus
- word embeddings for learning bilingual distributed representation of words

Contributions

- explore different word embedding models
- show how a general-domain comparable corpus can enrich a specialized comparable corpus via neural networks

Distributional-Based Approaches

The historical context-based projection approach, known as the standard approach, has been studied by a number of researchers:

- [Fung, 1998, Rapp, 1999, Chiao and Zweigenbaum, 2002]
- [Morin et al., 2007, Prochasson and Fung, 2011, Bouamor et al., 2013, Morin and Hazem, 2016]
- among others

Distributed-Based Approaches

Bilingual word embeddings has become a source of great interest in recent times:

- [Mikolov et al., 2013]
- [Vulić and Moens, 2013, Zou et al., 2013]
- [Chandar et al., 2014, Gouws et al., 2014, Artetxe et al., 2016]
- among others

Other approaches

- other work has focused on learning bilingual word representations without word-to-word alignments of comparable corpora
- sentence-aligned parallel data
  [Chandar et al., 2014, Gouws et al., 2014]
- document-aligned non-parallel data
  [Vulić and Moens, 2015, Vulić and Moens, 2016]

–> It is unlikely, not to say impossible, to find this type of alignment in a specialized comparable corpus

Data Combination Using Word Embeddings

- external data drastically improves the performance of the distributional-based approach [Hazem and Morin, 2016]
- distributed vector representations over large corpora capture many linguistic regularities and key aspects of words [Mikolov et al., 2013]

Data Combination Using Word Embeddings

we pursue the preceding works and propose different ways to combine specialized and general domain data using neural network models

- adapt the two data combination approaches proposed in [Hazem and Morin, 2016] using Skip-gram and CBOW models [Mikolov et al., 2013]
- propose different Skip-gram and CBOW models combinations [Garten et al., 2015] over specialized and general domain data

Global Data Combination Using Word Embeddings

- similar to the GSA approach [Hazem and Morin, 2016]
- we use the distributed Skip-gram or Continuous Bag-of-Words (CBOW) models [Mikolov et al., 2013]

Global Data Combination Using Word Embeddings

Bilingual terms extraction steps:

1. build a CBOW (or a Skip-gram) model for source and target languages
2. apply bilingual mapping [Artetxe et al., 2016] between the source and the target CBOW models (or the Skip-gram models)
3. mapping needs a bilingual dictionary (a dictionary subset of the 5,000 more frequent translation pairs)
4. compute a Cosine similarity between the mapped embedding vector and the embedding vectors of all the target words
5. rank the candidates according to their similarity score

Specific Data Combination Using Word Embeddings

- in the line of the SSA [Hazem and Morin, 2016] approach
- build two separate representations
- concatenate the distributed models while [Hazem and Morin, 2016] merge distributional context vectors
- our goal is to capture the two word characterisations thanks to CBOW/Skipgram models

Specific Data Combination Using Word Embeddings

Bilingual terms extraction steps:

1. build a CBOW (or a Skip-gram) model for both specialized and general domain data
2. concatenate source CBOW vectors (or Skip-gram vectors) of the specialized and the general domain data
3. apply bilingual mapping [Artetxe et al., 2016] between the source and target concatenated vectors
4. compute a Cosine similarity between the mapped embedding vectors and the embedding vectors of all the target words
5. rank the candidates according to their similarity score

Combining Distributed Representations

- vectors concatenation –> substantial improvements on a standard word analogy task [Garten et al., 2015]
- special advantages when training data is limited

Combining Distributed Representations

word embedding models lead to three different ways of concatenation:

- a CBOW model concatenation between the specialized and the general domain data
- a Skip-gram model concatenation
- a concatenation of both CBOW and Skip-gram models

Combining Distributed Representations

Concatenation

- 100 dim spec CBOW + 200 dim gen CBOW = 300 dim CBOW
- 100 dim spec Skip-gram + 200 dim gen Skip-gram = 300 dim Skip-gram
- concatenate CBOW and Skip-gram = 600 dimension combined model
  - allows to take advantage of both CBOW and Skip-gram models
  - learn a mapping matrix of the combined models

Data and Resources

Comparable corpus	# content words	
	FR	EN
BC	8,221	7,907
NC	5.7M	4.7M
EP7	61.8M	55.7M
JRC	70.3M	64.2M
CC	91.3M	81.1M

- specialized comparable corpus
  - breast cancer corpus (BC) of 500k tokens
- general domain corpus
  - News commentary corpus (NC) of 5.7M (Fr) and 4.7M (En)
  - Europarl corpus (EP7) of 61.8M (Fr) and 55.7M (En)
  - JRC acquis corpus (JRC) of 70.3M (Fr) and 64.2M (En)
  - Common Crawl corpus (CC) of 91.3M (Fr) and 81.1M (En)
- Terminology reference list of 248 single word pairs

Results

Corpus	*CBOW*	*SG*	*Concat*
BC	17.1	12.8	20.8
NC	33.9	31.2	33.6
EP7	42.3	40.8	43.1
JRC	40.3	40.5	43.4
CC	60.9	56.0	**61.0**

Results

Corpus	*CBOW*	*SG*	*Concat*
BC	17.1	12.8	20.8
NC	33.9	31.2	33.6
EP7	42.3	40.8	43.1
JRC	40.3	40.5	43.4
CC	60.9	56.0	**61.0**
BC ∪ NC	42.9	37.7	46.3
BC ∪ EP7	47.2	49.0	53.3
BC ∪ JRC	49.9	46.5	53.0
BC ∪ CC	67.7	63.2	**68.4**

Results

Corpus	*CBOW*	*SG*	*Concat*
BC	17.1	12.8	20.8
NC	33.9	31.2	33.6
EP7	42.3	40.8	43.1
JRC	40.3	40.5	43.4
CC	60.9	56.0	**61.0**
BC ∪ NC	42.9	37.7	46.3
BC ∪ EP7	47.2	49.0	53.3
BC ∪ JRC	49.9	46.5	53.0
BC ∪ CC	67.7	63.2	**68.4**
BC⌒(BC ∪ NC)	45.5	30.7	48.1
BC⌒(BC ∪ EP7)	51.6	35.7	53.8
BC⌒(BC ∪ JRC)	53.7	36.3	56.1
BC⌒(BC ∪ CC)	70.7	40.2	**70.9**

Table: Results (MAP %) of the Skip-gram model (noted SG), the Continuous Bag of Words model (noted CBOW) and their concatenation (noted Concat).

Results

	BC	NC	EP7	JRC	CC
SA	27.0	45.3	48.5	52.0	75.5
GSA	-	**58.9**	58.3	61.7	80.2
SSA	-	**58.9**	**60.8**	**66.6**	**82.3**

Results

	BC	NC	EP7	JRC	CC
SA	27.0	45.3	48.5	52.0	75.5
GSA	-	**58.9**	58.3	61.7	80.2
SSA	-	**58.9**	**60.8**	**66.6**	**82.3**
GCBOW	17.1	42.9	47.2	49.9	67.7
GSG	12.8	37.7	49.2	46.5	63.2
GCBOW + *GSG*	20.8	46.3	53.3	53.0	68.4

Results

	BC	NC	EP7	JRC	CC
SA	27.0	45.3	48.5	52.0	75.5
GSA	-	**58.9**	58.3	61.7	80.2
SSA	-	**58.9**	**60.8**	**66.6**	**82.3**
GCBOW	17.1	42.9	47.2	49.9	67.7
GSG	12.8	37.7	49.2	46.5	63.2
GCBOW + *GSG*	20.8	46.3	53.3	53.0	68.4
SCBOW	-	45.5	51.6	53.7	70.7
SSG	-	30.7	35.7	36.3	40.2
SCBOW + *SSG*	-	48.1	53.8	56.1	70.9

Table: Results (MAP %) of the *Standard Approach* (*SA*), the *Global Standard Approach* (*GSA*) and the *Selective Standard Approach* (*SSA*) for the breast cancer corpus (BC) using the different external data (the improvements indicate a significance at the 0.001 level using the Student t-test).

Conclusion

- we have proposed and contrasted different data combinations using neural networks
- We have shown under which conditions external resources as well as Skip-gram and CBOW models can be jointly used to improve the performance of bilingual terms extraction
- Although encouraging results, the performance are still below the distributional approach
- we concider this work as a starting point of applying word embeddings and the multiple proposed variants to specialized domains

Perspectives

- other specialized domains and external resources (wikipedia)
- domain adaptation techniques
- deep learning approaches (RNN, LSTM, CNN...)
- parallel aligned and document aligned approaches for external data
- syntactic dependencies
- distrubutional versus distributed

Thank you.

Thank you.

Artetxe, M., Labaka, G., and Agirre, E. (2016).
Learning principled bilingual mappings of word embeddings while preserving monolingual invariance.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 2289–2294, Austin, TX, USA.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2013).
Context vector disambiguation for bilingual lexicon extraction from comparable corpora.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 759–764, Sofia, Bulgaria.

Chandar, A. P. S., Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V. C., and Saha, A. (2014).
An autoencoder approach to learning bilingual word representations.

In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS'14)*, pages 1853–1861, Montreal, Quebec, Canada.


Chiao, Y.-C. and Zweigenbaum, P. (2002).
Looking for candidate translational equivalents in specialized, comparable corpora.
In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212, Tapei, Taiwan.


Fung, P. (1998).
A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora.
In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup (AMTA'98)*, pages 1–17, Langhorne, PA, USA.


Garten, J., Sagae, K., Ustun, V., and Dehghani, M. (2015).
Combining distributed vector representations for words.

In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 95–101, Denver, CO, USA.


Gouws, S., Bengio, Y., and Corrado, G. (2014).
Bilbowa: Fast bilingual distributed representations without word alignments.
CoRR, abs/1410.2455.


Hazem, A. and Morin, E. (2016).
Efficient data selection for bilingual terminology extraction from comparable corpora.
In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, pages 3401–3411, Osaka, Japan.


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


Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007).

Bilingual Terminology Mining – Using Brain, not brawn comparable corpora.
In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.


Morin, E. and Hazem, A. (2016).
Exploiting unbalanced specialized comparable corpora for bilingual lexicon extraction†.
Natural Language Engineering, 22(4):575—-601.


Prochasson, E. and Fung, P. (2011).
Rare Word Translation Extraction from Aligned Comparable Documents.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, pages 1327–1335, Portland, OR, USA.


Rapp, R. (1999).

Automatic Identification of Word Translations from Unrelated English and German Corpora.
In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.


Vulić, I. and Moens, M. (2013).
A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else).
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1613–1624, Seattle, WA, USA.


Vulić, I. and Moens, M. (2015).
Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL'15)*, pages 719–725, Beijing, China.


Vulić, I. and Moens, M. (2016).

Bilingual distributed word representations from document-aligned comparable data.
Journal of Artificial Intelligence Research (JAIR), 55(1):953–994.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1393–1398, Seattle, WA, USA.