Multilingual Vector Representations of Words, Sentences, and Documents

Gerard de Melo
Department of Computer Science
Rutgers University – New Brunswick
gdm@demelo.org

Abstract

Neural vector representations are now ubiquitous in all subfields of natural language processing and text mining. While methods such as word2vec and GloVe are well-known, *multilingual* and *cross-lingual* vector representations have also become important. In particular, such representations can not only describe words, but also of entire sentences and documents as well.

1 Introduction

Vector representations are ubiquitous in all subfields of natural language processing and text mining. Well-known neural methods such as word2vec and GloVe enable us to obtain distributed vector representations of words, overcoming some of the sparsity issues faced by traditional distributional semantics methods. Such representations are learnt from co-occurrence information drawn from large monolingual corpora.

Oftentimes, however, we are interested in representations that enable us to transition across language boundaries. Thus, it is useful to consider *multilingual* vector representations, covering multiple languages, and in particular *cross-lingual* vector representations, which capture the semantics of different items in the same vector space, even when said items stem from different source languages.

This is useful for representations of individual words (Section 2), but also of entire sentences (Section 3) and documents (Section 4) as well.

2 Multilingual Word Vectors

One can distinguish several broad classes of algorithms for inducing cross-lingual word vectors.

Projection Approaches. The first strategy is to train multiple separate vector spaces using standard methods and then align them cross-lingually. The latter can be achieved using techniques such as linear projections (Mikolov et al., 2013), Canonical Correlation Analysis (Faruqui and Dyer, 2014), or the approach by Gouws et al. (2015).

Parallel Corpora Approaches. The second strategy is to rely on parallel corpora and directly optimize a cross-lingual objective that considers sentence translations. Examples include the methods proposed by Klementiev et al. (2012), Kočiský et al. (2014), and Gouws and Søgaard (2015). Some of these simply use aligned sentences, while others require word alignments. Vulić and Moens (2015a) showed that comparable documents may suffice to learn bilingual embeddings.

External Supervision. Alternatively, a third strategy is to draw on supplementary sources of supervision. For this, one can extract more explicit semantic information from text and then incorporate the mined knowledge into the objective function (Chen and de Melo, 2015; Chen et al., 2016). Loza Mencía et al. (2016) propose exploiting document labels as a surrogate form of supervision for higher-quality embeddings.

Finally, one can also draw on lexical knowledge graphs such as WordNet and its multilingual extensions (de Melo and Weikum, 2010), or on Wikipedia (de Melo and Weikum, 2014). These resources provide a rich source of information to induce massively multilingual word vectors covering hundreds of languages in the same space (de Melo, 2015; de Melo, 2017), with the additional advantage of also yielding sense- or concept-specific representations.

3 Multilingual Sentence Vectors

Next, we turn to vector representations of sentences.

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 3–5,
Taipei, Taiwan, November 27 – December 1, 2017.

Word Vector-inspired Approaches. A widely used strategy is to simply average the word vectors of words in a given sentence. Despite its simplicity, this method often works surprisingly well (Wieting et al., 2015; Arora et al., 2017).

An early attempt to incorporate sentences more explicitly into the objective function was proposed with the Paragraph Vectors approach (Le and Mikolov, 2014). This method is also occasionally referred to as doc2vec, as it straightforwardly extends word2vec to additionally create representations of sentences or other longer units. Several authors have devised bilingual variants of the Paragraph Vector approach (Pham et al., 2015; Mogadala and Rettinger, 2016).

The Skip-Thought Vector approach (Kiros et al., 2015), while also inspired by the word2vec skip-gram method, instead draws on recurrent units to encode and decode sentence representations such that the resulting representations are optimized for predicting neighbouring sentences.

External Supervision. Wieting et al. (2015) explored using supervision from paraphrase information to obtain custom-tailored word vectors that give rise to high-quality sentence embeddings. The InferSent approach (Conneau et al., 2017) relies on supervision from the Stanford Natural Language Inference data as an auxiliary task to obtain sentence representations. In terms of cross-lingual methods, neural machine translation based on sequence-to-sequence learning can give rise to vector encodings of multilingual input sentences (Luong et al., 2015). These have been shown to be semantically meaningful (Schwenk and Douze, 2017).

4 Multilingual Document Vectors

Finally, we consider representations of entire text documents.

Word Vector-inspired Approaches. To obtain document representations, a common choice is again to simply take the average of word vectors, or a suitably weighted sum. In doing so, one can directly rely on multilingual word vectors to generate cross-lingual documents representations (Klementiev et al., 2012) that can be used for tasks such as cross-lingual text classification (de Melo and Siersdorfer, 2007).

A fallback strategy is to translate all documents to a single language and then consider monolingual document similarity metrics. This approach may be more costly in terms of the resources used, and may neglect language-specific subtleties. Still, in practice, it does appear to be a strong baseline (de Melo and Siersdorfer, 2007).

Modeling Document Semantics. While many methods treat sentences and documents as interchangeable, there are significant differences between the two. Methods that focus specifically on properties of documents have the potential to yield higher-quality document-level embeddings. Bag-of-words vectors can be rendered cross-lingual by translating individual words (Song et al., 2016), or by moving from original words to bag-of-concept representations (de Melo and Siersdorfer, 2007), optionally drawing on distributed vectors for concepts (de Melo, 2017). Representations may also account for the salience of different parts of the text (Yang et al., 2016, 2017). Hermann and Blunsom (2014) train a siamese-style network architecture on a parallel corpus such that it learns to compose sentence representations into document representations. Finally, when documents are to be compared against short queries, it is important to consider the peculiarities of relevance modeling (Vulić and Moens, 2015b; Hui et al., 2017), which differs from semantic similarity modeling.

5 Conclusion

In summary, vector representations have made it easier to target multilingual and cross-lingual semantics. This is possible both at the level of individual words as well as at the level of sentences or even entire documents.

Biography

Gerard de Melo is an Assistant Professor of Computer Science at Rutgers University, heading a team of researchers working on NLP, Big Data analytics, and web mining. He has published over 80 papers on these topics, with Best Paper/Demo awards at WWW 2011, CIKM 2010, ICGL 2008, the NAACL 2015 Workshop on Vector Space Modeling, as well as an ACL 2014 Best Paper Honorable Mention, a Best Student Paper Award nomination at ESWC 2015, and a thesis award for his work on graph algorithms for knowledge modeling. Notable research projects include UWN/MENTA, the first massively multilingual version of WordNet, and Lexvo.org, an important hub in the Web of Data. For further information, please refer to `http://gerard.demelo.org`.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR 2017*.

Jiaqiang Chen and Gerard de Melo. 2015. Semantic information extraction for improved word embeddings. In *Proceedings of the NAACL Workshop on Vector Space Modeling for NLP*.

Jiaqiang Chen, Niket Tandon, Charles Darwis Hariman, and Gerard de Melo. 2016. WebBrain: Joint neural learning of large-scale commonsense knowledge. In *Proceedings of ISWC 2016*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364.

Gerard de Melo. 2015. Wiktionary-based word embeddings. In *Proceedings of MT Summit XV*. AMTA.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL 2014*.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of ICML 2015*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of NAACL-HLT*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of ACL 2014*.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. A position-aware deep model for relevance matching in information retrieval. In *Proceedings of EMNLP 2017*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of NIPS 2015*. MIT Press.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of ACL 2014*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of ICML 2014*. PMLR.

Eneldo Loza Mencía, Gerard de Melo, and Jinseok Nam. 2016. Medical concept embeddings via labeled background corpora. In *Proceedings of LREC*.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *CoRR*, abs/1511.06114.

Gerard de Melo. 2017. Inducing conceptual embedding spaces from Wikipedia. In *Proceedings of WWW 2017*. ACM.

Gerard de Melo and Stefan Siersdorfer. 2007. Multilingual text classification using ontologies. In *Proceedings of ECIR 2007*, volume 4425 of *LNCS*. Springer.

Gerard de Melo and Gerhard Weikum. 2010. Towards universal multilingual knowledge bases. In *Proceedings of the 5th Global WordNet Conference*.

Gerard de Melo and Gerhard Weikum. 2014. Taxonomic data integration from multilingual Wikipedia editions. *Knowledge and Information Systems*, 39(1):1–39.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*.

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of NAACL 2016*.

Hieu Pham, Minh-Thang Luong, and Christopher D Manning. 2015. Learning distributed representations for multilingual text sequences. In *Proceedings of NAACL-HLT 2015*.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*.

Yangqiu Song, Shyam Upadhyay, Haoruo Peng, and Dan Roth. 2016. Cross-lingual dataless classification for many languages. In *Proceedings of IJCAI*.

Ivan Vulić and Marie-Francine Moens. 2015a. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of ACL-IJCNLP 2015*.

Ivan Vulić and Marie-Francine Moens. 2015b. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of SIGIR 2015*. ACM.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *CoRR*, abs/1511.08198.

Qian Yang, Yong Cheng, Sen Wang, and Gerard de Melo. 2017. HiText: Text reading with dynamic salience marking. In *Proceedings of WWW 2017*.

Qian Yang, Rebecca J. Passonneau, and Gerard de Melo. 2016. PEAK: Pyramid evaluation via automated knowledge extraction. In *Proceedings of AAAI*.