Book Review

Cross-Lingual Word Embeddings

Anders Søgaard, Ivan Vulić, Sebastian Ruder, Manaal Faruqui
(University of Copenhagen, University of Cambridge, DeepMind, Google Assistant)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 37), 2019, 132 pp; paperback, ISBN 9781681735725; ebook, ISBN 9781681730646; doi:10.2200/S00920ED2V01Y201904HLT042

Reviewed by
Eneko Agirre
University of the Basque Country (UPV/EHU)

The representation of words across languages is of interest since the early days of interlingual machine translation, as it allows us to connect the meaning of words in different languages and to generalize lexical semantic properties and relations across languages (Hutchins 2000). Structured representations such as multilingual lexical knowledge bases represent polysemy, as well as language internal and cross-lingual relations, but they require costly manual construction and maintenance (Vossen 1998). Alternatively, corpus-based methods have been used to automatically induce monolingual word representations like word embeddings with great success (Mikolov et al. 2013). Word embeddings represent the words in the vocabulary of a language as vectors in n-dimensional space, where words that are similar being located close to each other. Cross-lingual word embeddings (CLWE for short) extend the idea, and represent translation-equivalent words from two (or more) languages close to each other in a common, cross-lingual space.

The interest in cross-lingual word embeddings has grown in recent years. This is partly because of their success in cross-lingual transfer, where NLP tools trained in a resource-rich language such as English are transferred to another language with smaller or no annotated data. For instance, given training data for a text-classification task in English, a model using CLWE can classify foreign language documents. Beyond language pairs, CLWE allows us to represent words of several languages in a common space, and thus pave the way to build multilingual NLP tools that use the same model to process text in different languages.

This comprehensive and, at the same time, dense book has been written by Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui. It covers all key issues as well as the most relevant work in CLWE, including the most recent research (up to May 2019) in this vibrant research area. It does a great job of organizing different approaches in a typology, according to the kind of bilingual resources needed, and differentiating word-level, sentence-level, and document-level models. The book also covers extensions to CLWE that are able to represent multiple languages in the same space, as well as unsupervised learning, where the systems only use monolingual resources to build the cross-lingual space. The book is structured in 12 chapters.

Chapter 1 is a brief introduction, which includes an explanation of the notation used in the book. The book tries to establish a formal relation between several word-level

https://doi.org/10.1162/coli_r_00372

alignment models, and it does a thorough job of describing the methods using a consistent mathematical formalization. This consistency allows the authors to describe methods in a more compact way and helps to better see the common patterns across seemingly different approaches. In this sense, introducing the notation in the first chapter makes life easier for later. The formulas are quite dense and demanding, but although a mathematically naive reader might have a hard time following them, casual readers can also get value from a higher level read of the book.

Chapter 2 makes a brief introduction of the main monolingual word embedding models, focusing on the formalization of the loss function being optimized by each of the models.

Chapter 3 introduces a typology of supervised CLWE models, that is, methods that use some kind of bilingual signal. The typology leaves aside multilingual and unsupervised methods, which are covered in later chapters. The typology is based on data requirements along two dimensions: the type of bilingual signal (at the level of words, sentences, or documents), and whether the method requires parallel resources, or comparable resources suffice.

Chapter 4 introduces work on cross-lingual word representations that pre-dates the introduction of word embeddings. The chapter covers work on cross-lingual clusters, delexicalization strategies for cross-lingual transfer, earlier use of seed dictionaries, together with distributional vector space models, cross-lingual word alignment in machine translation, and latent cross-lingual concepts. This chapter draws connections with earlier work, and is a must-read for anyone wanting to take a step back from current techniques, to look at the big picture and draw inspiration in the larger picture of (non-embbeding-related) NLP methods.

The book then follows with three chapters organized according to typology. Chapter 5 covers the most popular family of models, those based on word-level information. The models that require parallel data in the form of bilingual dictionaries (or word alignments induced from parallel corpora) are further classified into those that learn separate monolingual spaces for each language to then learn a mapping, or those that learn the cross-lingual space for the two languages jointly, as well as mixed approaches. The authors make an effort to show that some methods coming from mapping, joint, and mixed approaches are very similar. In addition, the chapter also covers methods that ground words into images or image features. Most of the space in the chapter is taken by mapping methods, as they take the bulk of recent publications.

Chapter 6 introduces sentence-level information, usually in the form of sentence-aligned parallel translations. The additional supervision is used to learn either shared sentence representations, bilingual encoders, or a bilingual version of the monolingual skip-gram loss.

Chapter 7 introduces methods that use information from comparable documents only, which offers less supervision compared with the methods in the previous two chapters. These methods typically use Wikipedia articles from different languages as comparable documents.

Chapter 8 introduces multilingual CLWE, where more than two languages are involved. Apart from the practical interest, some works show that the use of multiple languages improves the quality of word embeddings. Most works use bilingual CLWE learning methods taking a pivot language (e.g., English).

Chapter 9 is devoted to unsupervised methods, that is, those that learn the CLWE space without any bilingual information. At the core, unsupervised methods apply one of the supervised methods (e.g., a word-level mapping method from Chapter 5).

An initial small or low-quality seed lexicon is produced using some method, and iteratively, better dictionaries are obtained and used as seed lexicons.

Chapter 10 gathers a wide array of applications and intrinsic evaluation tasks that have been used across the literature. Arguably, bilingual dictionary induction is the reference evaluation task for word-level mapping algorithms, but other methods have chosen to evaluate on other tasks, making comparison across methods in the same family difficult, and comparison across types of methods unfeasible. Note that the book does not provide information about the experimental performance of the methods, which is understandable, given the lack of an agreed-upon evaluation task or data set that covers all methods in the typology. That said, some experimental evaluation information, although limited, would have been of interest to the reader, as it would allow one to have a grasp of the relative standing of some relevant methods introduced in the book.

On the practical side, Chapter 11 introduces a comprehensive list of monolingual corpora and embeddings, bilingual dictionaries, parallel corpora, and CLWE open source models. It also lists some relevant evaluation data sets and applications.

The book finishes in Chapter 12 with general challenges and future direction, where the authors outline some of the current challenges in this field, alongside some specific open problems.

In summary, this book provides a comprehensive and in-depth overview to cross-lingual word embeddings, covering the breadth of techniques and resources used. This book is recommended not only to researchers, students, and practitioners who work on the area of cross-lingual and multilingual word embeddings, but also to a wider range of readers who have interest in cross-lingual and multilingual processing.

Note that the book is very similar to a contemporaneous journal survey (Ruder, Vulić, and Søgaard 2019), written by three of the authors. The book is more detailed and contains a separate section on unsupervised methods and another section with useful data and software. On the other hand, the journal survey contains a handful of newer references, up through ACL 2019. The field is moving forward fast, and seeing the latest developments (at the time of writing this review), it seems the authors chose a very good time to publish the book. The irruption of contextual word embedding models, where the representation of a word depends on the context of occurrence, has put on the table a new family of alternative methods that learn cross-lingual word representations. It is good to see that one of the authors has already started to cover some of the newer methods in an excellent blog.[1]

References

Hutchins, W. John. 2000. *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*, volume 97. John Benjamins Publishing.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 3111–3119, Lake Tahoe, NV.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1):569–630.

Vossen, Piek. 1998. *EUROWORDNET: A Multilingual Database with Lexical Semantic Networks*. Springer.

1 https://ruder.io/unsupervised-cross-lingual-learning/.

Eneko Agirre is a professor at the Computer Science Faculty of the University of the Basque Country and director of the HiTZ research center. His research interests are mainly in natural language understanding, including (lexical) semantics, inference, and multilinguality. His email address is e.agirre@ehu.eus.