Multilingual and Interlingual Semantic Representations for Natural Language Processing: A Brief Introduction

Marta R. Costa-jussà
TALP Research Center, Universitat Politècnica de Catalunya
Marta.ruiz@upc.edu

Cristina España-Bonet
DFKI GmbH and Saarland University
cristinae@dfki.de

Pascale Fung
Hong Kong University of Science and Technology
pascale@ee.ust.hk

Noah A. Smith
University of Washington and Allen Institute for Artificial Intelligence
nasmith@cs.washington.edu

We introduce the Computational Linguistics *special issue on Multilingual and Interlingual Semantic Representations for Natural Language Processing. We situate the special issue's five articles in the context of our fast-changing field, explaining our motivation for this project. We offer a brief summary of the work in the issue, which includes developments on lexical and sentential semantic representations, from symbolic and neural perspectives.*

1. Motivation

This special issue arose from our observation of two trends in the fields of computational linguistics and natural language processing. The first trend is a matter of increasing *demand* for language technologies that serve diverse populations, particularly those whose languages have received little attention in the research community.

https://doi.org/10.1162/COLI_a_00373

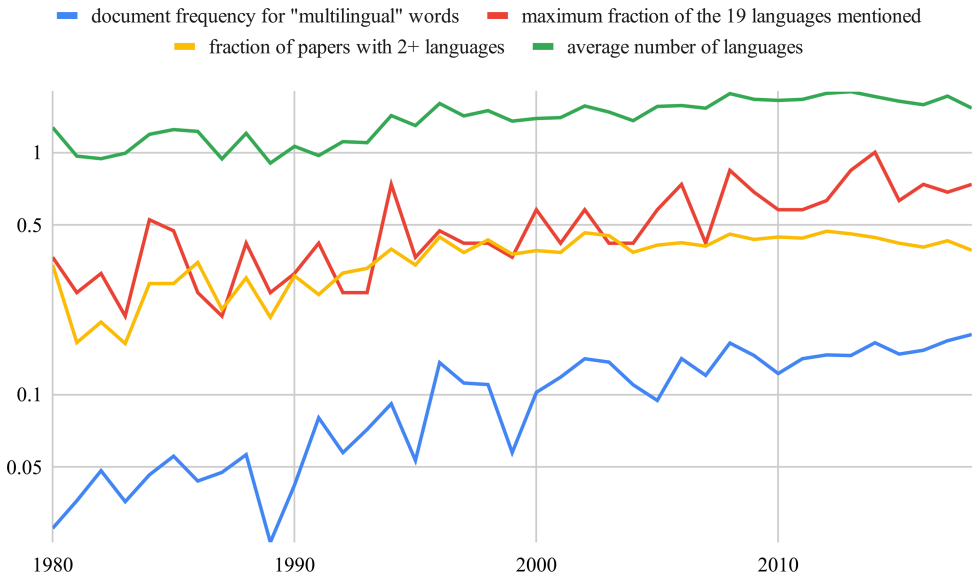

Figure 1
Trends in the ACL Anthology: Mulilinguality has increased in prominence over the past 40 years, with some signs of slowdown in the 2010s. Note that the graph uses a logarithmic scale for clarity.

As applications such as question answering, text summarization, speech recognition, and translation become useful, it is insufficient and inequitable in our globalized and connected world for these to serve only speakers of the *lingua franca*.

The growing importance of multilinguality is reflected in the community's research. Figure 1 quantifies this rise in the ACL Anthology.[1] Papers using either the term *multilingual* or *crosslingual* (or their hyphenated spellings) rose steadily by year (blue line). Mentions of specific languages are also increasing; starting from a list of 19 of the world's most spoken languages,[2] we find that the maximum observed fraction of these 19 in a given year is steadily increasing (from 4–10 in the 1980s to 11–19 in the 2010s; red line). The fraction of papers mentioning two or more languages (yellow line) and the average per year (green line) showed increases in the 1990s and 2000s, though these appear to have slowed recently.[3]

The other trend is a matter of increasing *supply*: The diversity of computational tools now available—from conceptual definitions of language meaning to operationalizations in downloadable models—has exploded in the past decade. The term "semantic representation" was, not long ago, one that referred to a range of linguistic abstractions.

1 We explored ACL Anthology papers in S2ORC (Lo et al. 2020) with publication years 1980–2019, a total of 40,402 papers.
2 The list is Ethnologue's list of the 20 most spoken languages in 2019, with Mandarin and Wu Chinese mapped to the string *chinese*. See `https://www.ethnologue.com/guides/ethnologue200`. Less dominant languages are, of course, also interesting, but also more sparse in the data.
3 The leveling off of these last two trends is, we speculate, due to the emergence of new representation learning methods that work best with very large data sets. We expect increasing multilinguality of the largest data sets and pretrained representations will enable a return to past upward trends.

Today, many of those have been transferred to annotated data sets, and many more have emerged through the application of representation *learning* methods to text corpora. These methods and the computational objects they produce (e.g., contextual word vectors) have reshaped the landscape of methods used to build applications, especially the scale and kinds of text data and other linguistic resources.

In the multilingual setting, semantic representations at word (Bojanowski et al. 2017; Lample et al. 2018) and sentence level (Artetxe and Schwenk 2019; Lample and Conneau 2019) are allowing the transfer of language technologies to dozens and even hundreds of languages for which the technologies are less evolved.

Beyond the listed trends, we believe there is a consensus in the computational linguistics community that the study of diverse natural languages is necessary for a full understanding of the phenomena, including universals and sources of variation. Though hegemonic languages have received greater research attention, methods, abstractions, and theories that explain evidence in many languages have obviously greater scientific value than those applicable to only one or a few. Noteworthy efforts in this area range from interlingual grammatical annotation schemes, such as the ones defined by the Universal Dependencies project,[4] which produce consistently multilingual annotated treebanks, to multilingual lexical databases such as multilingual WordNet (Bond and Paik 2012; Bond and Foster 2013) and BabelNet (Navigli and Ponzetto 2012).

Together, these conditions make 2020 an exciting time for natural language processing research, warranting a special issue to synthesize various lines of work that illustrate a range of creative advances exploring natural language meaning, specifically with a multilingual focus. In inviting submissions, we encouraged a broad reading of the term "representations," in granularity (words, sentences, paragraphs, etc.) and in theoretical assumptions (symbolic, neural, hybrid, etc.). We anticipated breadth as well in the set of motivating applications and evaluation methods. Our deliberate reference to *interlingual*—not only *multilingual*—representations evokes recent re-imaginings of interlingual machine translation, a classical approach (Richens 1958). We explicitly encouraged submissions that consider less-commonly studied languages and that go beyond mere projection of representations from text in one language to another.

Of particular interest to our editorial team is the potential for multilingual representations (of any kind) to help overcome challenges of polysemy in individual languages. It has been shown that translations into other languages can help at distinguishing senses monolingually (Resnik and Yarowsky 1999). But the complementary might also be true, and realizations in different languages of the same concept may help to obtain more robust embeddings at sense level as shown by one of the works presented here.

The contributions to this special issue are summarized in Table 1. The papers selected cover the different points we wanted to emphasize in our call. Three of the contributions refer to representations at word level and the others at sentence level, but the breadth of the field is reflected in the range of specific topics addressed. This issue presents novel work and reviews on interlingual representations (Ranta et al. 2020); semantic representations learned through translation at word (Mohiuddin and Joty 2020) and sentence level (Vázquez et al. 2020); senses, ambiguity, and polysemy (Colla, Mensa, and Radicioni 2020); and evaluation (Sahin 2020). Multilinguality is clearly the aim for all of them, with systems that cover from 4 up to 40 languages. Some systems also have the virtue to deal with text in low-resource languages such as Macedonian, Nepali, and Telugu.

4 http://universaldependencies.org.

Table 1
Summary of contributions to the special issue.

Granularity	Paper	Technique	Application	Languages
Word	(Mohiuddin and Joty 2020)	Unsupervised Adversarial	Translation	en, es, de, it, fi, ar, ms, he
	(Colla, Mensa, and Radicioni 2020)	Linked Data	Word Similarity	en, fr, de, it, fa, es, pt, eu, ru
	(Sahin 2020)	Intrinsic/extrinsic evaluation	POS, dependencies, SRL, NER, NLI	24 languages
Sentence	(Vázquez et al. 2020)	Attention	Translation	en, es, fr, cs
	(Ranta et al. 2020)	Grammars	Abstract language representation	40 languages

2. Lexical Representations

This special issue includes three papers that focus on different crosslingual challenges at the level of the lexical representation. The challenges addressed include learning unsupervised representations, introducing priors and linguistic knowledge to compute the representations, and evaluating the quality of these representations, taking into account linguistic features.

Unsupervised Word Translation with Adversarial Encoder (Mohiuddin and Joty 2020). Crosslingual word embeddings are becoming crucial in multilingual natural language processing tasks and, recently, several authors claim that unsupervised methods even outperform the supervised ones (see for instance Lample et al. 2018, Artetxe, Labaka, and Agirre 2018, Xu et al. 2018), making them appealing also in the low-resource setting. This is not true in all cases, and specifically, adversarial techniques for dictionary induction show stability and convergence issues for some language pairs (Zhang et al. 2017; Lample et al. 2018). In general, unsupervised adversarial bilingual embeddings are learned in two phases: *(i)* induction of an initial seed dictionary using an adversarial network and *(ii)* refinement of the initial mapping, and therefore, dictionary, until convergence. This paper tries to address those limitations by extending adversarial autoencoders. One of the main contributions is training the adversarial mapping in a latent space, with the hope that this will minimize the effect of a lack of isomorphism between the two original embedding spaces. In addition, the authors combine several loss functions in the initial mapping of source-target embeddings and experiment with various refinement techniques for the second phase. Their deep analysis of the results shows that forcing cycle consistency (i.e., the source translated into the latent target space and then back-translated into the source original space must be the same) and symmetric re-weighting (i.e., re-weight the embedding components according to cross-correlation, to increase the relevance of those that best match across languages and select the top-*k* as dictionary) are the major contributions to the final performance of the method. These techniques have been used before, but the authors show that their combination with other variants is the main reason for improving the robustness of adversarial methods.

LessLex: Linking multilingual Embeddings to SenSe Representations of LEXical Items. Colla, Mensa, and Radicioni (2020) propose this lexical resource composed of a set of embeddings, multilingual by design, which are built by retrieving information from BabelNet

synsets and using ConceptNet Numberbatch (CNN) word embeddings (Havasi, Speer, and Alonso 2009). The approach is motivated by the proposition that anchoring lexical representations to multilingual senses should be beneficial for both word representations and final applications. Interestingly, a comprehensive evaluation of such vectors seems to support this hypothesis. By using both resources (BabelNet and CNN), the tool is proven to be more effective than most related approaches through a set of experiments focusing on conceptual, contextual, and semantic text similarity. From a novel perspective, given that the proposed word representations are based on senses, the authors design a new technique to evaluate word similarity taking into account the similarity between the target words and each of their senses; then they use it to scale the similarity between the two senses.

LINSPECTOR: Multilingual Probing Tasks for Word Representations. Sahin (2020) releases a ready-to-use tool to evaluate word representations (or neural model layers) on multiple languages together with a variety of linguistic features. This tool moves beyond the standard classification probing task by using case marking, possession word length, morphological tag count, and pseudoword identification. Such type-level probing tasks are relevant for the scientific community because it is useful to analyze the underlying linguistic properties captured by a word embedding, which is especially important for morphologically rich languages. In addition to the tool, the paper includes complete experiments both on probing and downstream tasks of a variety of word representations in 24 languages. The main contribution of these experiments is that results reveal mostly significant positive correlations between probing and downstream tasks. Further analysis shows that these correlations are higher for morphologically rich languages.

3. Sentence Representations

The level of sentence representations is covered in this special issue from two contrasting perspectives: neural and symbolic. While the paper on neural representation describes research on an encoder-decoder architecture that shares an attention bridge, the symbolic contribution is an overview of different frameworks.

A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation. Vásquez et al. (2020) analyze the performance of a particular multilingual translation model to build fixed-size sentence representations. The proposed architecture is based on using a shared attention bridge in between language independent encoders and decoders. Exhaustive experiments are reported in downstream tasks (from the SentEval toolkit) as well as in multilingual machine translation (on small and large data sets). The outcomes of the study show that higher-dimensional sentence representations improve translation quality and also the performance in classification tasks. However, shorter sentence representations increase the accuracy in non-trainable similarity tasks. Beyond these conclusions, the most revealing findings from the paper are that multilingual training leads to a better encoding for linguistic properties at the level of a sentence, meaning that using the proposed attention bridge layer is beneficial in extracting both semantic and syntactic information.

Abstract Syntax as Interlingua: Scaling Up the Grammatical Framework from Controlled Languages to Robust Pipelines. Ranta et al. (2020) offer an overview of the linguistic principles of the Grammatical Framework (GF), a large-scale language resource that applies the abstract syntax idea to natural languages. Abstract syntax corresponds to

an interlingual representation of sentences in this case. From the linguistic perspective, the ambition of GF is to achieve an abstract representation that can accurately cover concrete linguistic phenomena such as inflectional and derivational morphology, segmentation and compounding, agreement, and semantic compositionality. Although GF is challenging by nature, the paper describes how NLP systems have successfully used the GF as well as how GF is related to other NLP semantic representations such as WordNet, FrameNET, Construction Grammar, and Abstract Meaning Representation. The relevance of GF is linked to the advantages of symbolic methods, including explainability, programmability, and data austerity, while the limitations remain in facing open domains. This paper gives the necessary background for future potential efforts, including the semi-automatic creation of a wide-coverage multilingual GF lexicon as well as hybrid approaches that combine GF with neural methods.

4. Outlook

As shown by the range of work showcased in this special issue, the area of multilingual natural language processing is active and developing rapidly. We expect continued advances and growth; our hope is that this special issue will spark new efforts and syntheses across subcommunities tackling this important agenda from different perspectives.

Acknowledgments

We thank Kyle Lo for assistance with the S2ORC data. MRC is supported in part by a Google Faculty Research Award 2018, Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund and the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal, the contract TEC2015-69266-P (MINECO/FEDER,EU) and the contract PCIN-2017-079 (AEI/MINECO). N. A. S. is supported by National Science Foundation grant IIS-1562364. C. E. B. is funded by the German Federal Ministry of Education and Research under the funding code 01IW17001 (Deeplee). Responsibility for the content of this publication is with the authors.

References

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia.

Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bond, Francis and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia.

Bond, Francis and Kyonghee Paik. 2012. A survey of WordNets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71, Matsue.

Colla, Davide, Enrico Mensa, and Daniele P. Radicioni. 2020. LESSLEX: Linking multilingual embeddings to SenSe linked representations of LEXical items. *Computational Linguistics: Special Issue Multilingual and Interlingual Semantic Representations for Natural Language Processing*. 46(2):289–333.

Havasi, Catherine, Robyn Speer, and Jason Alonso. 2009. *ConceptNet: A lexical resource for common sense knowledge*, MIT Media Lab.

Lample, Guillaume and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018,* April 30-May 3, 2018, *Conference Track Proceedings*, Vancouver.

Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the ACL*.

Mohiuddin, Tasnim and Shafiq Joty. 2020. Unsupervised word translation with adversarial autoencoder. *Computational Linguistics: Special Issue Multilingual and Interlingual Semantic Representations for Natural Language Processing*. 46:(2):257–288.

Navigli, Roberto and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Ranta, Arne, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina. 2020. Abstract syntax as interlingua: Scaling up the grammatical framework from controlled languages to robust pipelines. *Computational Linguistics: Special Issue Multilingual and Interlingual Semantic Representations for Natural Language Processing*. 46(2):425–486.

Resnik, Philip and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Richens, R. H. 1958. Interlingual machine translation. *Computer Journal*, (3).

Sahin, Gĺozde Gĺul. 2020. LINSPECTOR: Multilingual Probing Tasks for Word Representations. *Computational Linguistics: Special Issue Multilingual and Interlingual Semantic Representations for Natural Language Processing*. 46(2):335–385, TK.

Vázquez, Raúl, Alessandro Rafanato, Mathias Creutz, and Jĺorg Tiedemann. 2020. A systematic study of inner-attention-based sentence representations in multilingual neural machine translation. *Computational Linguistics: Special Issue Multilingual and Interlingual Semantic Representations for Natural Language Processing*. 46(2):387–424.

Xu, Ruochen, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels.

Zhang, Meng, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pages 1934–1945, Copenhagen.