Deep learning of word sense meaning - IXA group, Google Award

Proposers: Mikel Artetxe, Josu Goikoetxea, Eneko Agirre (IXA http://ixa.eus)

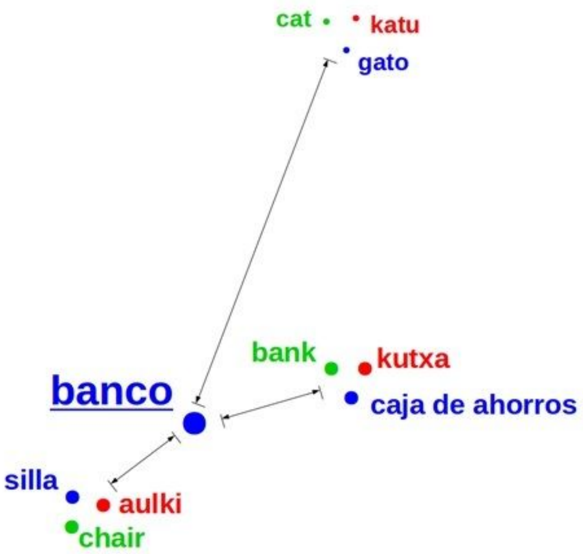
Contact: e.agirre@ehu.eus http://ixa2.si.ehu.eus/eneko

Description

Deep learning has changed the way in which NLP works, specially in the representations of words (Mikolov et al. 2013). Words are represented as vectors in semantic space, where the representations have been learned automatically from large bodies of text. In a way, the machine is able to build the semantic space just reading large collections of text, and this semantic space is useful for applications such machine translation, question answering, search engines and chatbots.

Word representations have some shortcomings, though:

- The representation is different for each language, while the intuition is that people have language-independent semantic spaces, that is, a semantic space which is useful for different languages.
- The representations are built for words, and thus confuse different meanings, e.g. the meaning representation of banco conflates both the bench meaning and the bank meanings.



The figure shows the meaning of banco in a cross-lingual embedding space (Ruder et al. 2017), where the meanings of Basque, English and Spanish words are also represented. As banco is polysemous, it’s representation mixes the bench and the bank meanings.

We have previously shown that the first shortcoming can be fixed with small or no dictionaries (Artetxe et al. 2016; 2017). The second is partially fixed combining word representations with knowledge bases (Goikoetxea et al. 2015; 2016), but no one has found a satisfactory method to build word representations that distinguish different meanings. We will explore novel research ideas to disentangle such meanings

This project is in the context of the Google Faculty Research Award received by Eneko Agirre.

Goals

The student will apply deep learning in order to learn the meaning of words, moving from the representation from words to the representation of senses of words. The key insights are the following:

1) Translation is useful to disentangle meanings of words (e.g. banco when translated as bank has a different meaning from banco when translated as bench)
2) Multilingual knowledge bases (e.g. Wordnet, Wikipedia) explicitly represent the concepts and their lexicalizations in each language (e.g. concept 428724 is lexicalized as banco and bank in Spanish and English, respectively, and concept 567234 is lexicalized as banco and bench in Spanish and English, respectively)

This project will explore those key insights to build concept embeddings in cross-lingual embedding space.

Requirements

English. Machine learning. Good programming skills, basic math skills.

Although it is not a requirement, taking the course "**Seminar on language technologies. Deep Learning**" (see below) will allow the student to accomplish more ambitious goals. Contact us for further details.

The dissertation can be written in Basque, English or Spanish.

Tasks and plan

Dec-Jan: Study literature

Feb: Attend course "Seminar on language technologies. Deep Learning" (see below)

Mar-May: Development and experiments

June: Write down and presentation

References

Artetxe, M., Labaka, G., & Agirre, E. (2016) Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In Proceedings of EMNLP.

Artetxe, M., Labaka, G., & Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In Proceedings of the ACL

Goikoetxea, J., Soroa, A., and Agirre, E. (2015). Random Walks and Neural Network Language Models on Knowledge Bases. In Proceedings of NAACL-HLT.

Goikoetxea, J., Soroa, A., and Agirre, E. (2016). Single or Multiple? Combining Word Representations Independently Learned from Text and WordNet. In Proceedings of AAAI.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of NIPS.

Ruder, S., Vulić, I. and Søgaard A. A Survey Of Cross-lingual Word Embedding Models. https://arxiv.org/abs/1706.04902

Seminar on Language Technologies. Deep learning. (LAP18)

Deep Learning neural network models have been successfully applied to natural language processing. These models are able to infer a continuous representation for words and sentences, instead of using hand-engineered features as in other machine learning approaches. The seminar will introduce the main deep learning models used in natural language processing, allowing the students to gain hands-on understanding and implementation of them in Tensorflow .

Topics

- Introduction to machine learning and NLP with Tensorflow
- Deep learning
- Word embeddings
- Language modeling and recurrent neural networks
- Convolutional neural networks
- Attention mechanisms

Prerequisite. Basic programming experience, a university-level course in computer science and experience in Python. Basic math skills (algebra or pre-calculus) are also needed.