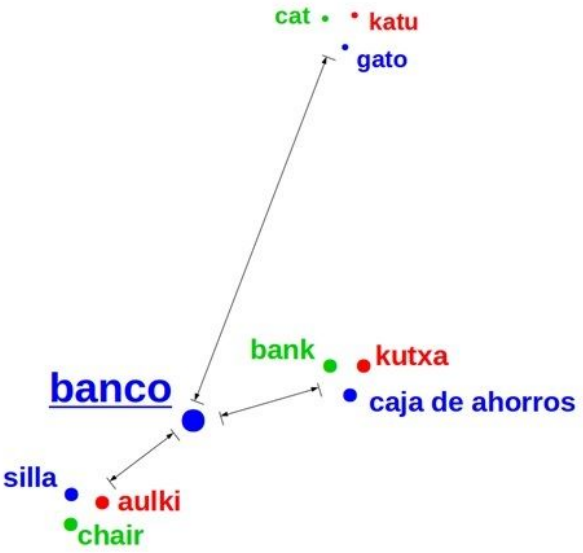
Unsupervised multilingual embeddings for deep learning - IXA group, Google Award

Proposers: Mikel Artetxe, Gorka Labaka, Eneko Agirre (IXA http://ixa.eus)

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

Description



Deep learning has recently changed the way in which Natural Language Processing (NLP) works. However, while deep learning models are inherently continuous, language has a discrete nature. This is not an issue for many areas in which deep learning has been successful (e.g. image processing), but poses a major challenge in NLP.

So as to overcome this problem, current deep learning models rely on word embeddings, abstract representations of words in a continuous vector space. Even if it is possible to learn these embeddings in an end-to-end setting, the scarcity of supervised data for many tasks has motivated numerous unsupervised methods to train them (Mikolov et al. 2013; Ruder et al. 2017). These methods, popularized by the word2vec toolkit, take a large body of text, and learn these vector representations of words in an abstract semantic space, which are a key component of many deep learning models.

However, most of these methods learn different representations for different languages, which is a major problem for building complex applications that work in multiple languages (e.g. chatbots like Siri), nor to say inherently cross-lingual tasks (e.g. machine translation). We have shown that this issue can be fixed by mapping word embeddings in two languages to a common semantic space with small or no dictionaries (Artetxe et al. 2016; 2017). However, the proposed method is designed for only two languages. This way, the aim of this project is to extend it to many languages, going from bilingual word embeddings to multilingual word embeddings.

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 a cross-lingual shared space for words, moving from 2 languages to several languages. The key insights are the following:

1) When representing two languages in a shared space (Artetxe et al. 2016), noise from the monolingual spaces affects the quality

2) When representing a third language, the third language can be used to cancel out or reduce the monolingual noise, increasing the quality of the shared space
3) The more languages we incorporate, the more the noise would be reduced.

This project will explore those key insights to build a shared embedding space for several languages.

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

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 (in particular in linear algebra) are also needed.