Methods of Mathematical Analysis final project

Polina Stadnikova

February 28, 2018

Abstract

This project implements a basic idea of the inverted indexing approach for creating word embeddings where words are represented as one-dimensional arrays. The objective is to obtain cross-lingual word representations that capture similar or the same concepts in different languages and to investigate potential challenges and limitations. For the evaluation, different techniques are used, including measuring the distance between two vectors. The outcome of the project is a small tool that allows the user to perform some experiments with these embeddings.

1 Method

1.1 Cross-lingual embeddings

Cross-lingual embeddings are projections from different languages into the same semantic space. They can be used in many NLP tasks, such as machine translation, syntactic parsing, POS-tagging, and are an object of active research in distributional semantics.

Creating such word embeddings usually requires multilingual parallel corpora. My method follows the idea from [Søgaard et al.] of using Wikipedia as a source. In Wikipedia, there are a lot of articles that describe the same topic in different languages and are linked to the same node in the ontology. That means, for each concept there is a *cross-lingual* representation that consists of the words from all corresponding articles. Consequently, a word from each of the languages can be described by a set of the concepts[Søgaard et al.]. In information retrieval, listing documents or concepts per word is known as inverted index[1] and considered more efficient than listing words per document.

In my method, I simplify the definition of concept and treat each article as a single concept. That is, for example, the German article and the English articles about calculus correspond to the concept "calculus". To obtain the word embeddings, I use a count-based approach.

1.2 Count-based representations

The semantic space of a language can be composed by vectors that correspond to the words in this space. In count-based approaches vectors contain the information about word co-occurrences, represented by raw or weighted counts[Jurafsky & Martin].

[1] https://en.wikipedia.org/wiki/Inverted_index

Raw counts are not very discriminative, that is, they are not helpful if we want to distinguish between words occurring only in a particular context and words that are frequent in general[Jurafsky & Martin]. In this project, the informativeness of words is captured by *pointwise mutual information* (PMI). This is a measure of how much more often than expected by chance two words co-occur and computed as follows:

$$PMI(a,b) = log_2 \frac{p(a,b)}{p(a)p(b)}$$

PMI values can be negative, that means, that two terms co-occur less often than it would by expected by chance. Negative values are usually considered unreliable in distributional semantics[Jurafsky & Martin] and can be replaced by zero. In this case, we compute *positive pointwise mutual information* (PPMI):

$$PPMI(a,b) = max(log_2 \frac{p(a,b)}{p(a)p(b)}, 0)$$

Another problem with PMI is the bias towards very rare words[Jurafsky & Martin]. I try to avoid this by implementing a simple smoothing technique, *Laplace smoothing*: adding a small value k to each count. Preliminary tests showed that $k = 0.8$ fits best (larger values increase the probability of getting negative PMI and therefore zero values in vectors).

The method does not include dimensionality reduction because the vector space is relatively small.

The semantic space of a language is then a word-concept matrix, where a word w is represented by a vector matching a row in this matrix. Each value in the vector corresponds to $PPMI(w, c_i)$, with $c_i \in C$, a set of common concepts.

1.3 Corpus

To train my embeddings, I create a small parallel corpus including articles in English, German and Spanish. First, I extracted parallel article titles[2] and obtained a set of "common concepts". In order to simplify the approach and its evaluation, I restricted the number of concepts to 30, trying to approximate the "MoMA semantic space" where each concept corresponds to a (sub)topic from the course (s. Table 1). To download the articles, I make use of the MediaWiki API[3]. During the training process, an $m \times n$ matrix is created for each language, where m is the vocabulary size for the corresponding language and $n = 30$.

[2] https://github.com/clab/wikipedia-parallel-titles

[3] https://www.mediawiki.org/wiki/API:Main_page

de	en	es
Analysis	Mathematical analysis	Análisis matemático
Differentialrechnung	Differential calculus	Cálculo diferencial
Differentialrechnung	Derivative	Derivada
Integralrechnung	Integral	Integración
Fundamentalsatz der Analysis	Fundamental theorem of calculus	Teorema fundamental del cálculo
Grenzwert (Funktion)	Limit of a function	Límite de una función
Extremwert	Maxima and minima	Extremos de una función
Funktion (Mathematik)	Function (mathematics)	Función matemática
Mengenlehre	Set theory	Teoría de conjuntos
Taylor-Formel	Taylor's theorem	Teorema de Taylor
Mittelwertsatz der Differentialrechnung	Mean value theorem	Teorema del valor medio
Stetigkeit	Continuous function	Función continua
Potenz (Mathematik)	Exponentiation	Potenciación
Logarithmus	Logarithm	Logaritmo
Polynom	Polynomial	Polinomio
Taylorreihe	Taylor series	Serie de Taylor
Explizites Euler-Verfahren	Euler method	Método de Euler
Vektor	Euclidean vector	Vector
Vektorraum	Vector space	Espacio vectorial
Skalar (Mathematik)	Scalar (mathematics)	Escalar (matemática)
Norm (Mathematik)	Norm (mathematics)	Norma vectorial
Euklidischer Raum	Euclidean space	Espacio euclídeo
Matrix (Mathematik)	Matrix (mathematics)	Matriz (matemáticas)
Skalarprodukt	Dot product	Producto escalar
Lineare Gleichung	Linear equation	Ecuación de primer grado
Einheitsmatrix	Identity matrix	Matriz identidad
Orthogonale Matrix	Orthogonal matrix	Matriz ortogonal
Orthogonalität	Orthogonality	Ortogonalidad (matemáticas)
Einheitsvektor	Unit vector	Vector unitario
Determinante	Determinant	Determinante (matemática)

Table 1: mutual concepts, each row corresponds to a concept

1.4 Preprocessing

The training data is tokenized and lowercased. Stop words are removed from the data. I also tried to enforce projecting different forms of the same word into one single embedding (to make infomative words even more prominent) by stemming. But, according to the results of preliminary tests, such embeddings are harder to evaluate, so I do not include stemming in my implementation.

In this approach, the main objective is to make the keywords for each topic most prominent. Some POS in the corpus, like verbs, that are not stopwords and therefore are taken into account when computing embeddings, do not really fit my experiments. To enforce more interesting results, I add a POS tagger[4] and keep only adjectives and nouns (including proper names). For Spanish, I also include past participle forms to preserve some important terms like “derivada” (derivative).

2 Evaluation

Comparing cross-lingual word embeddings is not a trivial task, especially without integrating complex NLP tasks (cannot be implemented within this project). Here I provide only a basic evaluation.

[4]http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/ I intentionally do not use NLTK taggers here.

2.1 Experiment 1

The aim of this experiment is to project onto one concept per iteration. That is, each concept is assigned to 35 words from each language that are supposed to represent this concept (these are the word with the highest PPMI). The words *across* the languages are expected to be roughly related to each other (ideally, to be translations, like "derivative" - "Ableitung" - "derivada" for the concept "function"). I also expect less related topic to be represented by different words.

Results without filtering POS tags seem to contain much noise and are hard to interpret. Filtering helps to get better results, even if they are still far from what was expected. Smoothing does not seem to significantly influence the result. One can see the same personal names across the languages, e.g Baire, Taylor, Euler, or related terms within one concept, e.g for the concept "dot product": "Kosinuswerte" for German, Θ for English, "cos"/"coseno" and "radianes" for Spanish, or "Gewichts" vs. "weights". As expected, there are also differences between the concepts: for example, the word "matrix" or "matriz" occurs in concepts that are more related to linear algebra than to calculus.

I provide some files with full tables for all the concepts.

2.2 Experiment 2

This experiment compares words of the *same* language in terms of the distance between their vectors. To measure the distance I use the *cosine similarity*[Jurafsky & Martin] and Euclidean distance[5] . The cos similarity corresponds to the cosine of the angle between two vectors: the larger the cosine value, the smaller the angle and therefore the closer the word representations. Euclidean metric captures the straight-line distance between two vectors, that means, the smaller the distance, the smaller the difference.

Here I expect "similar" words to have larger cosine values and smaller Euclidean distance. This is verified by most of the test runs. For example, in German "Matrix" and "Determinante"(*cos = 0.8, E.d. = 3.4*) have more similar representations than "Funktion" and "Matrix"(*cos = 0, E.d. = 6.4*), in English: "function" and "maximum"(*cos = 0.4, E.d. = 5.8*) vs "function" and "matrix"(*cos = 0, E.d. = 6.3*). Smoothing and POS filtering do not have a significant impact on the results.

2.3 Experiment 3

This experiment attempts to compare words in the same way as experiment 2, but *across* the languages. In this case, I also expect larger cosine values and smaller Euclidean distance between "related" words, ecpecially between the translations of the same word. The results also contain shared concepts.

Again, my expectations are verified by the tests: "Funktion"(de) and "function"(en) are considered similar (*cos = 0.8, E.d. = 2.4*), as well as "Ableitung"(de) and "derivada"(es) (*cos = 0.9, E.d. = 2*). Words "function/Funktion/función" and "matrix/Matrix/matriz" have the zero cosine value since they do not have concepts in common. Here I obtain slightly better results without both smoothing and filtering or with filtering and without smoothing.

[5]https://en.wikipedia.org/wiki/Euclidean$_distance$

2.4 Conclusion

This simple approach allows to build an approximation of the "multilingual MoMA semantic space", as verified by experiments 2 and 3. With this tool users can measure the similarity between words within one language or across different languages. It also displays shared concepts for multilingual embeddings.

Cross-lingual word embeddings are hard to compare (s. experiment 1), more sophisticated experiments are necessary for more informative evaluation.

The size and the content of corpora have a large impact on the quality of word embeddings. Wikipedia seems to be a good source for multilingual parallel corpora.

Laplace smoothing discounts the non-zero values, but is not crucial for the performance. Filtering POS tags is only important for experiment 1, its application and relevance should be explored more carefully.

3 User Guide

- to start experiments run *process.py*
- *process.py* has to be in the same directory as *invert.py*
- you need internet connection to load the Wikipedia articles
- training the embeddings takes about 1-2 minutes
- you can choose whether to use smoothing and filtering
- you can choose the experiment (numbers correspond to the experiments in this report)
- you can perform experiments as long as you want :)

References

[Søgaard et al.] Anders Søgaard, Zeljko Agic, Hector Martinez Alonso, Barbara Plank, and Bernd Bohnet (2015). Inverted indexing for cross-lingual NLP *In ACL, Vol. 1*, 1713–1722.

[Jurafsky & Martin] Daniel Jurafsky and James H. Martin (2017). Speech and Language Processing. *Third Edition draft*, 275–291.