The Problems of Evaluation of Distributional Semantic Models

Amir Bakarov and Andrey Kutuzov

Higher School of Economics, Moscow, Russia
University of Oslo, Oslo, Norway

aabakarov@edu.hse.ru, andreku@ifi.uio.no

2017 Ivannikov ISPRAS Open Conference

30.11.2017

Overview

Introduction

Distributional semantics

Distributional semantic models are frameworks that can represent words of natural language through real-valued vectors of fixed dimensions.

The word "distributional" here is a reference to a **distributional hypothesis** that says that word semantics is distributed along all of its contexts.

Zelig S. Harris (1954)
Distributional Structure
Word, 10(23): 146-162..

Real-valued representations of words are called **word embeddings**.

History of evaluation of distributional semantics

Common approaches to evaluation

- **Word similarity.** Given a dataset of word pairs (`word1,word2, similarity`) where `similarity` reports human judgements about degree of similarity of two considered words, the task is to evaluate correlation for two vectors of similarity labels: $X = x_1, ..., X_n$ and $Y = y_1, ..., y_n$, where X is a dataset of human judgements and Y is a dataset of similarity metrics for the same word pairs produced by the word embeddings models (for example, cosine similarities between word vectors).
- **Downstream tasks.** Word embeddings are used as feature vectors of classifiers dedicated to resolve more complex tasks like POS-tagging or detection of semantic relatedness between two sentences.

Critique of common approaches to evaluation

1. Word similarity
   - The notion of semantics (hence, notion of any semantic relation) is obscure; the annotation task is unclear;
   - Human annotations tend to be subjective;
   - It is unclear if the human representations of semantics absolutely correct;
   - The model is considered as "good" if it represents one type of semantic relations well; but what if such models dedicated to represents another type of semantic relations?
2. Downstream tasks
   - Performance in different tasks don't correlate between each other, therefore the evaluation score is not absolute.

Novel and experimental approaches

Novel methods

- Extrinsic (downstream tasks, *in vitro* evaluation)
- Intrinsic (absolute evaluation, *in vivo* evaluation)
    - **Conscious** (offline methods in terms of psycholinguistic research);
    - **Unconscious** (online methods in terms of psycholinguistic research);
    - **Knowledge-based** (comparison with manually constructed knowledge bases);
    - **Linguistic-driven** (using empirical information about language).

Extrinsic methods of evaluation

- Noun Phrase Chunking;
- Named Entity Recognition;
- Semantic Role Labeling;
- Paraphrase Detection;
- ...etc.

Intrinsic conscious methods of evaluation

- Word Similarity;
- Word Analogy;
- Word Categorization;
- Thematic fit;
- ...etc.

Intrinsic unconscious (experimental) methods of evaluation

- Semantic Priming;
- Measuring brain activity (electroencephalography, functional magnetic resonance imaging)

Other methods of intrinsic evaluation

- Knowledge-based:
  - Semantic Networks (e.g. WordNet);
  - Explicit Semantic Analysis;
  - Dictionaries.
- Linguistic-driven:
  - Bigram frequency;
  - Phonosemantic word representations.

Our experiments

Motivation of our experiments

- Are results of different distributional semantic models on the same dataset **different**?
- Do results of different models on different tasks **correlate** with each other?

Explored models

1. **Word2Vec** (2013): computation of the prediction loss of the target words from the context words.
2. **GloVe** (2014): dimensionality reduction on the co-occurrence matrix.
3. **Word2Vec-f** (2014): extension of Word2Vec with the use of arbitrary context features of dependency parsing.
4. **Wang2Vec** (2015): extension of Word2Vec with the sensitivity to the word order.
5. **AdaGram** (2015): extension of Word2Vec learning multiple word representations with capturing different word meanings.
6. **FastText** (2015): extension of Word2Vec which represents words as bags of character n-grams.
7. **Swivel** (2016): capturing unobserved (word, context) pairs in sub-matrices of a co-occurrence matrix.

Datasets

1. Word Similarity (Russian Datasets):
   - **HJ: Human Judgements of Word Pairs**, 398 word pairs (289 for Word2Vec-f and 376 for other models were used), scaled labels;
   - **RT: Synonyms and Hypernyms from the Thesaurus RuThes** (test chunk), 9550 word pairs (2481/5640 were used), binary labels;
   - **AE: Cognitive Associations from the Sociation.org Experiment** (test chunk), 3004 word pairs (1861/2721 were used), binary labels.
2. Semantic Relatedness
   - **Our dataset**, contains 2663 Russian pairs of short (up to 216 symbols) texts with binary labels (reporting existence of relatedness); the distribution of classes is 48% to 52%. The sentences were annotated with the help of 3 native speaking volunteers.

Results

Table:
Performance of the vectors of the compared models across different tasks. The word similarity task reports Spearman's p and average precision (AP) with human judgements; the semantic relatedness task reports $F1$. In all cases, larger numbers indicate better performance.

Model	Word Similarity			
	HJ, p	RT, AP	AE, AP	Semantic Relatedness, F_1
Word2Vec	0.51	0.72	0.78	0.85
GloVe	0.4	0.74	0.77	0.85
Word2Vec-f	0.04	0.73	0.74	0.78
Wang2Vec	0.41	0.72	0.78	0.85
AdaGram	0.11	0.57	0.66	0.81
FastText	0.44	**0.76**	**0.79**	0.85
Swivel	**0.52**	0.74	0.76	0.85

Our contributions

- Our work is the first towards **an extensive survey** of the word embedding evaluation methods;
- We propose an **evaluation of embedding models** applied to the textual data of Russian language on two tasks.

Code, datasets, trained models and used corpus could be found in our repository: `https://github.com/bakarov/2ch2vec`

Conclusions and future work

Conclusions

- Our hypothesis that different word representations propose different results on different evaluation tasks was confirmed;
- We have surveyed different methods of evaluation of word embeddings.

Open questions

- How to evaluate **cross-language word embeddings** and **multi-sense word embeddings**?
- What is the most adequate way of obtaining distributional representations of **compositional linguistic units** (compositional distributional semantics)?
- Should we avoid **bias in word embeddings**, and, if yes, how could we detect it?
  Example of bias: the word "man" is closer to the word "programming" than the word "woman", but there is no reason why men should be connected to programming more than women.
- ...and many more of still unspoiled questions.

Thank you for your attention!

Amir Bakarov, Andrey Kutuzov

Feel free to ask about preprint! Write to aabakarov@edu.hse.ru