Methods of Evaluation of Word Embeddings

Amir Bakarov

amirbakarov@gmail.com

Novosibirsk State University

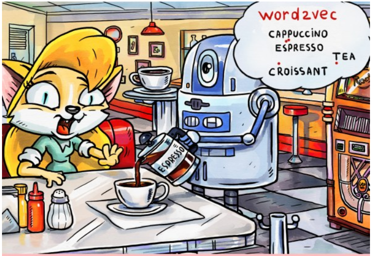
March 30, 2019

Outline

Introduction

Issues with Word Embeddings

- **Issue I: Black boxes.** People do not understand the linguistic motivation and the type of semantic relations that the embeddings capture.
- **Issue II: No proper evaluation.**



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

Evaluation in NLP

Any natural language processing system needs evaluation against human references.

Jones and Galliers (1995)
Evaluating natural language processing systems: An analysis and review
Springer Science & Business Media.

Approaches to evaluation:

1. **Extrinsic:** evaluation of model performance on downstream tasks.
2. **Intrinsic:** evaluation of the model itself (evaluation of properties of the model).

Issues with Evaluation

The problem of evaluation in meaning representations

1. Models that are trained on different corpora (Russian National Corpus, or Wikipedia, Or Aranea, etc);
2. Models that exploit different architecures (Word2Vec, GloVe, etc);
3. Models that have different hyperparameters;
4. and so on.

The problem of evaluation in meaning representations

- **Extrinsic evaluation.** If one wants to train and use the model for the specific task (NER, POS-tagging, etc), they are okay.
- **Intrinsic evaluation.** If the model is designed as a general-purpose one (as a formalism, not as a tool), one should able to evaluate its overall quality.

Some rhetorical questions here:

1. What could ‘quality of the model‘ mean in terms of general-purpose model?
2. How could such notion of quality be measured?
3. Can the model really be task-independent, or is it always task-specific?

The problem of meaning

Boundaries of notion of meaning are not clear.

Two types of information:

1. Lexical (embedded) information:
   - Graphematic form;
   - Phonology;
   - Word class;
   - Meaning.
2. Encyclopedic (world knowledge) information.

Do distributional semantic models capture only meaning itself, and not other type of information embedded in the linguistic units?

The problem of meaning

Dichotomy of views on the word meaning:

1. **Minimalist hypothesis:** nothing that we know about word is embedded into its meaning.
2. **Maximalist hypothesis:** all that we know about word is embedded in its meaning.

Diversity in views on semantics:

1. Referential;
2. Conceptual;
3. Structural;
4. Prototypical.

How is distributional semantics connected with it?

Methods of evaluation

Taxonomy

- 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

- Sentence Boundary Detection;
- Named Entity Recognition;
- Semantic Role Labeling;
- Paraphrase Detection;
- etc.

Problems of extrinsic evaluation

- Different tasks favor different embeddings;
- Different architectures and datasets could propose different results;
- Could not be used as a proxy for a general notion of embedding quality.

Intrinsic methods of evaluation

- **Word similarity**;
- **Word analogy**;
- Thematic fit;
- Word categorization;
- Synonymy detection;
- Outlier detection.

Baroni et al. (2014)
Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors
ACL.

Word similarity

Task: Given words *a* and *b*, the task is to find measure of similarity (**it could be not only similarity!**) between them.

Example:

- Words *mug* and *cup*. Human assessor score for similarity (synonymy) is 0.9. If the cosine distance between word vectors is close to 0.9, then the model works well.
- Other type of semantic relations: relatedness (co-hyponymy) for *cup* and *coffee*; meronymy, hyponymy, etc.

Approach: F1 on a list of pairs.

Motivation: check the relations between words in the lexicon.

Critique

- The notion of similarity is obscure; the annotation task is unclear;
- Human annotations tend to be subjective;
- It is unclear if the human judgements on similarity are absolutely correct;
- The model is considered "good" if it represents one type of semantic relations well; but what if such models are dedicated to represent another type of semantic relations?

Word analogy

Task: Given a set of three words $\{a, a*, b\}$, the task is to find such word $b*$ for which the relation b:$b*$ would be the same as the relation a:$a*$.

Example: a = Paris, a^* = France, b = Moscow. Correct answer: Russia

Primary approaches:

1. 3CosAdd: $argmax_{b^* \in V}(sim(b^*, b - a + a^*))$
2. PairDirection: $argmax_{b^* \in V}(cos(b^* - b, a^* - a))$
3. 3CosAvg: $argmax_{b^* \in V}(sim(b^*, a^* + avgoffset))$
4. 3CosMul: $argmax_{b^* \in V} \frac{cos(b^*, b) cos(b^*, a^*)}{cos(b^*, a) + \epsilon}$
5. etc.

Motivation: check the adequacy of word meaning.

Critique

- Result in Boolean type suffers an information loss;
- A single prediction is not enough to represent the quality of all 4 word vectors trained;
- Computational complexity;
- Meaning may be encoded in more complex ways than just summation in a vector space.

Thematic fit (also called selection preference)

Task: Classify pair $\{p, a\}$ where p is a *predicate* and a is an *argument* by the most adequate thematic role for a.

Example: for pair *{eat, people}* the most adequate role for *people* is *subject* (*people eat*, not *eat people*).

Approach: vectors in a given role of most frequent nouns are averaged to obtain a "prototype" vector for the relevant argument slot. Then pairwise cosines for the vector for a target noun and possible prototype vector are measured, and the closed prototype vector is picked.

Motivation: check the right usage of meaning captured by the model.

Word categorization

Task: Given a set of words, the task is to split it into subsets of words belonging to different categories.

Example: For a set of words *dog*, *cat*, *apple*, *banana* the first two make one cluster (animals) and the last two form another cluster (fruits). The model should cluster them correctly, and the measure (V-score, ARI, etc) should be high.

Approach: different unsupervised clustering techniques, different class-based clustering measures.

Motivation: check the adequacy of words semantic classes.

Synonymy detection

Task: Given a word a and a set $K = b_1, b_2, b_3$, the task is to find b_i which is the most synonymous to a.

Example: for the target *mug* the model must choose between *cup*, *glass*, *jar* and *plate*. The correct answer should be *cup*, because it is the closest my the meaning.

Approach: pairwise cosine distance.

Motivation: check the adequacy of semantic relation without grounding to artificial scalar values.

Outlier word detection

Task: Given a set of words, the task is to identify a semantically anomalous word in this set.

Example: For a set of words $\{orange, banana, lemon, book, orange\}$ the word *book* is the outlier, and the model should identify it.

Approach: Pairwise distances between words.

Motivation: check the adequacy of words semantic classes without biasing by clustering techniques.

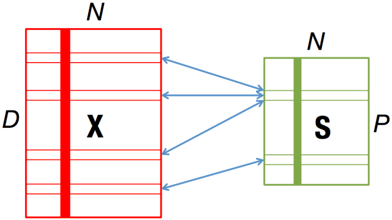
Other methods of intrinsic evaluation

- Knowledge-based:
  - Semantic networks (e.g. WordNet);
  - Explicit Semantic Analysis;
  - Dictionary graphs;
- Linguistic-driven:
  - Collocation frequency and lexical typology data;
  - Phonosemantic word representations.

Amir Bakarov (2018)
A Survey of Word Embeddings Evaluation Methods
arXiv:1801.09536.

QVEC

Alignment to a matrix of features extracted from manually crafted lexical resources: for example, lexicographer classes of WordNet which partitions nouns and verbs into coarse semantic categories.



Tsvetkov et al. (2015)
Evaluation of Word Vector Representations by Subspace Alignment
EMNLP.

Conclusion

Other questions

- How to evaluate subword-level representations? Contextualized representations (ELMO, BERT, GPT-2, whatever)?
- What is the most reliable way of obtaining distributional representations of **compositional linguistic units**?
- Should we avoid **fairness bias in representations**, and, if yes, how could we detect it?
  Example of fairness 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.*
- How should we deal with hubness problem and other limitations of distributional hypothesis?

A few more other questions

- Are there any reliable intrinsic measures for task-independent meaning representations, and are they able to predict downstream task performance?
- Does the stability factor in training representations exist, and what are the best practices for reproducible and reliable experiments?
- Could we explain the effect of architecture and parameter choices of deep learning models?
- What are the alternatives to cosine similarity and vector offsets in representation evaluation methods?
- How could real-valued representations be linguistically interpreted?

Resources

- `https://vecto.space`: framework that implements most of the described methods.
- `https://github.com/vecto-ai/word-benchmarks`: repository of benchmarks for several languages.
- `https://vectors.nlpl.eu/`: repository of pre-trained models.
- `https://ldtoolkit.space`: yet another evaluation framework.

RepEval at NAACL (June, 2019): A workshop on evaluation meaning representations for NLP

Thank you for your attention!