Subsumption Preservation as a Comparative Measure for Evaluating Sense-Directed Embeddings

A. Patrice Seyed
3M HIS
Silver Spring, MD
apseyed@gmail.com

Abstract

While there has been a growing body of work on word embeddings, and recent directions better reflect sense-level representations, evaluation remains a challenge. We propose a method of query inventory generation for embedding evaluation that recasts the principle of subsumption preservation, a desirable property of semantic graph-based similarity measures, as a comparative similarity measure as applied to existing lexical resources. We aim that this method is immediately applied to populate query inventories and perform evaluation with the ordered triple-based approach set forth, and inspires future refinements to existing notions of evaluating sense-directed embeddings.

1 Introduction

Work in the area of word embeddings has exploded in the last several years. Approaches based on word prediction (Mikolov et al., 2013) show improvement over traditional and recent work on count based vectors (Baroni et al., 2014). There has been gradual movement toward sense-directed or sense-level embeddings (Huang et al., 2012; Faruqui et al., 2015; Trask et al., 2015) while existing evaluation strategies based on applications, human rankings, and solving word choice problems have limitations (Schnabel et al., 2015). A limitation of relying on downstream applications for evaluation is that results vary depending on the application (Schnabel et al., 2015). In recent work, Tsvetkov (2015) leverages alignment with existing manually crafted lexical resources as a standard for evaluation, which shows a strong correlation with downstream applications.

Along this vein, there is an increasing need for methodologies for word-sense level evaluation measures. The utility of word embeddings is to reflect notions of similarity and relatedness, and word embeddings intended to represent senses should in turn reflect structured relations like hypernymy and meronymy. Most existing resources on lexical similarity and relatedness rely on subjective scores assigned between word pairs. This style of evaluation suffers from limited size of the evaluation sets and subjectivity of annotators. To address the first issue, we propose a method for exploiting existing knowledge formalized in lexical resources and ontologies as a means to automating the process of populating a query inventory. To address the second issue, we propose an evaluation approach that, instead of human scoring of word pairs, relies on comparative similarity given a semantic ordering represented as 3-tuples (henceforth triples). The method applies the principle of subsumption preservation as a standard by which to generate a query inventory and evaluate word embedding by geometric similarity. For example, subsumption is preserved when the similarity score of embeddings representing *ferry* and *boat* is greater than that of *ferry* and *vessel*. In the following section we illuminate the method, evaluation approach, an exploratory experiment, its results, related work, and next steps.

2 Method

The foundation of the method is the principle of *subsumption preservation* (Lehmann and Turhan, 2012).[1] We define this principle with axiom schemata as follows:

[1] We reference the two principles of subsumption and reverse subsumption atomically via the disjunction. *Transitive* serves as syntactic shorthand for the corresponding axiom. We assume the relationship between A and C is not asserted but inferred by transitivity.

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 90–93,
Berlin, Germany, August 12, 2016.

$SP_{sim_{rel}}$(A,B,C) $=_{def}$
rel(A,B) $\wedge$ rel(B,C) $\wedge$ Transitive(rel) $\rightarrow$
sim(A,B) $\vee$ sim(B,C) $\geq$ sim(A,C)

$SP_{sim_{rel}}$(A,B,C) means that similarity measure *sim* conforms to the subsumption preservation principle with respect to relation *rel* for all triples $\langle$A,B,C$\rangle$, just in case for any tuple $\langle$A,B,C$\rangle$ of *rel* related via transitivity, the similarity score of $\langle$A,B$\rangle$ and that of $\langle$B,C$\rangle$ is greater than or equal to that of $\langle$A,C$\rangle$. The property of subsumption preservation provides a link between subsumption and similarity in that it expresses the constraint that A and B (B and C) are more similar than A and C since the former pair(s) are 'closer' in the corresponding graph. Note that *rel* serves as relational schema that is satisfied by transitive, generalization relations. This includes taxonomic or partonomic inclusion that are the foundation of lexical resources and ontologies (e.g., WordNet, Gene Ontology).

The original intent of the subsumption preservation principle is that any quantitative semantic similarity measure *sim* is constrained by this desirable formal property. For instance, *Path* (Rada et al., 1989) abides by the subsumption preservation principle, and is defined as Path(A,B) $=_{def}$ 1/p, where *p* is the length of the path separating two concepts, *A* and *B*. A weakness of this and similar measures is that the length of path between two concepts is often a reflection of variability in the knowledge modeling technique or scope and not necessary a reflection of relatedness. To account for this shortcoming, Resnik (1995) applies the notion of information content: IC_{corpus} = -log(freq(A)), the inverse log of a concept *A*'s frequency in a given corpus, of a concept pair's least common subsumer as the similarity measure. There are other, varied approaches to semantic similarity that are based on a combination of corpus statistics and lexical taxonomy (Jiang and Conrath, 1997). Ultimately these approaches produce a score that is to some extent dependent on graph-based distances.

In the present work we take a different approach by proposing comparative similarity that hinges on semantic graph order preservation as the unit of evaluation. The intent is to apply only a basic geometric similarity measure (e.g., cosine) as *sim* within our definition of subsumption preservation, in order to provide a measure of how well embeddings abstract to the knowledge structure expected of a sense-directed embedding.

Thus given word embeddings, a knowledge resource and a similarity measure over the embedding space, an embedding does not conform to the subsumption preserving principle, if for example, the similarity score between terms *sparrow* and *bird* or *bird* and *vertebrate* is less than that of *sparrow* and *vertebrate*. A set of sense embeddings do not conform to the subsumption preserving principle to the proportion of cases that are violated. By adhering to the subsumption preserving principle a set of sense embeddings reflects notions of foundational semantic relationships and comparative similarity explicitly formalized in lexical and ontological resources. Thus, evaluation based on this method can serve as an indicator of how well approaches for learning embeddings can reflect relationships that are not present in knowledge resources.

3 Evaluation Approach

Traditionally word pairs of a query inventory are scored by similarity with a value between 0-1. We propose a different approach based on the unit of ordered triple instead of pairs, and that is relative rather than absolute and quantitative. Given a set of tuples of a relation *rel* that *sim* is potentially constrained by under subsumption preservation, we consider the candidate triples as instances of a query inventory for evaluation.

A similar approach has been applied in the evaluation of machine translation. Kahn (2009) describes a family of dependency pair match measures that are composed of precision and recall over various decompositions of a syntactic dependency tree. A dependency parser determines the relevant word triples where the relation is the second element. Reference and hypothesis sentences are converted to a labeled syntactic dependence tree, and the relations from each tree are extracted and compared. We draw inspiration from this approach, where the unit of evaluation is the ordered triple. Given the nature of our task we apply the measure of accuracy on the triples.

4 Exploratory Experiment Setup

For evaluation the BLESS dataset is selected as the basis for selecting a triple-based query inventory, (Baroni and Lenci, 2011), focusing on hypernymy and leaving meronymy as a future consider-

ation. For pairs that are related by hypernymy we identify intermediate words within the hypernym graph to generate candidate triples, including only nouns. For embeddings we used word2vec-based embeddings generated from google corpora.[2] For the similarity measure we selected cosine similarity, although the evaluation approach assumes embeddings and a similarity measure are two variables. So for example the score of sim(broccoli, vegetable) is greater than sim(broccoli,produce), therefore one part of the subsumption preservation principle is conformed to for the triple ⟨broccoli, vegetable, produce⟩. Also, sim(vegetable, produce) is greater than sim(broccoli, produce), therefore the triple is also in conformance with the other part of the subsumption preserved principle, namely reverse subsumption preservation.

We consider two approaches for calculating cosine similarity between words within the word2vec generated embeddings. The first is the simple approach and is performed by calculating the cosine between two word embeddings. The second is the aggregate approach, and requires, for each of the two words, exhaustively collecting all sister lemmas for the senses each word is a lemma of, calculating the centroid for all corresponding embeddings, and calculating cosine similarity between the resultant pair of centroid embeddings. The aggregate approach is in effort to simulate sense level embeddings for this exploration. We also consider the role of word generality in the evaluation.

5 Results

The results of the exploratory evaluation are shown in Table 5. SS, RSS, AS, and RAS represent subsumption and reserve subsumption preservation by the simple and aggregate approaches. The triple inventory *w/o abstract* represents where triples including highly abstract terms *object* and *artifact* were removed, and the inventory *IC threshold* represents where triples only included terms with Information Content above 3.0. Therefore the number of triples in the three inventories are approximately 1900, 900, and 300, respectively. In all three cases 5k was used as the unigram frequency cutoff for all terms in the triples, and it was observed that increasing above this value did not improvement accuracy. The results of the latter two runs illustrate where the most general term in the triples is more likely a domain concept, which coincides which better overall accuracy.

[2] https://code.google.com/archive/p/word2vec/

triple inventory	SS	RSS	AS	RAS
baseline	.67	.68	.73	.68
w/o abstract	.78	.72	.78	.69
IC threshold	.88	.73	.78	.65

Table 1: Accuracy figures for the triple-based query inventory generated from the BLESS dataset and WordNet.

6 Related Work

Newman (2010) applies semantic similarity measures leveraging WordNet, among other resources, for measuring coherence of topic models. Word pairs of a topic's top N terms are scored by similarity measures, where all synset pairs for a word pair are exhaustively applied prior to calculating their mean. The goal is to determine, based on topics previously selected by Mechanical Turkers as coherent, how well similarity measures reflect the coherence. It was found that WordNet-based similarity measures varied greatly, while non-graph similarity measures using Wikipedia and more generally applying pointwise mutual information performed the best.

Schnabel (2015) performs a comparative intrinsic evaluation based on selected word embeddings and nearest neighbor terms by cosine similarity for different word embedding learning approaches. Mechanical Turk participants were asked to select the most similar term from nearest neighbors for a given target term. Embedding learning approaches are compared by average win ratio.

7 Discussion and Future Work

In this paper we put forth a method for generating a triple-based query inventory and evaluation to assist in determining how well word embedding abstract to the sense, conceptual level. This approach provides an evaluation of relative rather than absolute similarity, the latter of which can lead to drastic differences in similarity scoring. The results improved by applying filters to the BLESS-derived query inventory aimed at where the most general term in the triples are more "meaningful", or put simply, where we increased the proportion of domain knowledge being tested. Since this occurred at the cost of the size of the

triple set, it is worth considering other heuristics for augmenting the generated candidate triples to improve their utility. We hope that this approach be ultimately treated as a sort of unit test for embeddings aimed at the open or a particular domain.

In future work we will perform the evaluation on sense embeddings (Trask et al., 2015), and on embeddings that integrate with lexical resources (Faruqui et al., 2015; Rothe and Schütze, 2015). We will also investigate the use of other broader relations, such as meronymy, as well as consider other lexical and ontological resources that are more comprehensive for the domains we aim to evaluate. Another consideration is evaluating embeddings with other similarity measures that account for asymmetry. Further, we aim to test if the accuracy conforming to subsumption preservation correlates with an evaluation of a downstream task, to confirm whether it can serve as a valid proxy.

8 Acknowledgements

Many thanks to Amitai Axelrod, Omer Levy, and the reviewers for fruitful feedback.

References

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Processings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, page 110. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 238–247.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1606–1615.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 873–882.

Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Jeremy G. Kahn, Matthew G. Snover, and Mari Ostendorf. 2009. Expected dependency pair match: predicting translation quality with expected syntactic structure. *Machine Translation*, 23(2-3):169–179.

Karsten Lehmann and Anni-Yasmin Turhan. 2012. A framework for semantic-based similarity measures for $\{\text{ELH}\}$ concepts. In *Logics in Artificial Intelligence*, pages 307–319. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1):17–30.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *arXiv preprint arXiv:1507.01127*.

Tobias Schnabel, Igor Labutov, David M. Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 298–307.

Andrew Trask, Phil Michalak, and John Lui. 2015. "sense2vec-a fast and accurate method for word sense disambiguation. *Neural Word Embeddings*.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2049–2054.