A Method for Taxonomy-Aware Embeddings Evaluation (Student Abstract)

Navid Nobani,[1*] **Lorenzo Malandri,** [2] **Fabio Mercorio** [2]
Mario Mezzanzanica [2]

[1] Dept. of Informatics, Systems & Communication, Univ. of Milan-Bicocca, Milan, Italy
[2] Dept. of Statistics and Quantitative Methods, Univ. of Milan-Bicocca, Milan, Italy
{navid.nobani, lorenzo.malandri, fabio.mercorio, mario.mezzanzanica}@unimib.it

Abstract

While word embeddings have been showing their effectiveness in capturing semantic and lexical similarities in a large number of domains, in case the corpus used to generate embeddings is associated with a taxonomy (i.e., classification tasks over standard de-jure taxonomies) the common intrinsic and extrinsic evaluation tasks cannot guarantee that the generated embeddings are consistent with the taxonomy. This, as a consequence, sharply limits the use of distributional semantics in those domains. To address this issue, we design and implement MEET, which proposes a new measure -HSS- that allows evaluating embeddings from a text corpus preserving the *semantic similarity* relations of the taxonomy.

Introduction and Contribution

Despite their extensive usage in a wide variety of modern NLP tasks, evaluation of embeddings is still an open debate in the literature, as there is not a unique definition of what either an" effective" or a" performant" assessment measure is (Schnabel et al. 2015; Malandri et al. 2020b). We develop a new measure for estimating the semantic similarity between taxonomic elements that we use as supervision for the embedding selection. The contribution of our work goes towards two directions.(i) First, we develop a new measure, (i.e., HSS) to evaluate the semantic similarity between words in a taxonomy, considering concepts cardinality and multiple word senses. We compare HSS on classic benchmarks showing it outperforms current SOTA approaches;(ii) Second, we define and implement a methodology, namely MEET (A Method for Embeddings Evaluation for Taxonomic Data), that relies on the *semantic similarity* relationships automatically computed from a taxonomy for evaluating embeddings from a large text corpus. We show MEET outperforms SOTA methods, which rely on handcrafted resources for intrinsic embeddings evaluation.

Methodology and State of the Art

Hierarchical Semantic Similarity (HSS) measures semantic similarity in a taxonomy based on the similarity values encoded within the hierarchy itself (Malandri et al. 2020a). Compared with HSS, SOTA metrics (see (Jauhiainen et al. 2019) for a recent survey) suffer from two main drawbacks: 1. when a word has multiple senses, those methods compute a value of similarity for each word sense and then consider only the highest. As a consequence, more specific senses will have a higher value of similarity ; 2. though they consider the structure of the taxonomy (i.e., relationship between concepts) they do not take into account the number of child entities (i.e., words) belonging to those concepts. Since we want to extend a semantic hierarchy built from human experts, we adopt those values as a proxy of human judgements. Therefore, similarly to (Seco, Veale, and Hayes 2004) we compute $\hat{p}(c)$ using an intrinsic measure, exploiting the structure of the taxonomy instead of an external corpus. While Seco only uses the number of taxonomic concepts, we consider also the entities of the taxonomy: $\hat{p}(c) = \frac{N_c}{N}$ where N is the cardinality, i.e. the number of entities (words), of the taxonomy and N_c the sum of the cardinality of the concept c with the cardinality of all its hyponyms. Note that $\hat{p}(c)$ is monotonic and increases with granularity. Now, given two words w_1 and w_2, Resnik defines $c_1 \in s(w_1)$ and $c_2 \in s(w_2)$ all the concepts containing w_1 and w_2 respectively, i.e. the *senses* of w_1 and w_2. Therefore, there are $S_{w_1} \times S_{w_2}$ possible combinations of their word senses, where S_{w_1} and S_{w_2} are the cardinality of $s(w_1)$ and $s(w_2)$ respectively. We can now define $\mathcal{L}$ as the set of all the lowest common ancestor for all the combinations of $c_1 \in s(w_1), c_2 \in s(w_2)$.

The hierarchical semantic similarity between the words w_1 and w_2 can be defined as:

$$sim_{\mathrm{HSS}}(w_1, w_2) = \sum_{\ell \in \mathcal{L}} \hat{p}\left(\ell = LCA \mid w_1, w_2\right) \times I(LCA) \tag{1}$$

Where $\hat{p}\left(\ell = LCA \mid w_1, w_2\right)$ is the probability of LCA being the lowest common ancestor of w_1, w_2, and can be computed as follows applying the Bayes theorem:

$$\hat{p}\left(\ell = LCA \mid w_1, w_2\right) = \frac{\hat{p}\left(w_1, w_2 \mid \ell = LCA\right) \hat{p}\left(LCA\right)}{\hat{p}\left(w_1, w_2\right)} \tag{2}$$

We define N_ℓ as the cardinality of ℓ and all its descendants.

*corresponding author

	Task 1												Task2		
	HSS (ours)		WUP		LC		Shortest Path		Resnik		JC		categ.		classif.
	P	S	P	S	P	S	P	S	P	S	P	S	**battig**	**esslli**	(±0.01)
WSS	**0.68**	**0.65**	0.58	0.59	0.36	0.23	0.16	0.1	0.02	-0.03	0.04	0.06	-	-	-
MT287	**0.46**	**0.31**	0.4	0.28	0.26	0.12	0.11	0.11	0.03	0.04	0.18	0.16	-	-	-
MEN	**0.41**	**0.33**	0.36	**0.33**	0.14	0.05	0.07	0.03	0.05	0.03	-0.05	-0.04	0.58	0.77	0.82
HSS (ours)	-	-	-	-	-	-	-	-	-	-	-	-	**0.62**	**0.81**	**0.84**
SimLex999	-	-	-	-	-	-	-	-	-	-	-	-	0.49	0.78	0.83
WUP	-	-	-	-	-	-	-	-	-	-	-	-	0.43	0.73	0.72

Table 1: Experimental results on benchmark datasets and metrics (see (Jauhiainen et al. 2019)) for Task 1 and 2.

Now we can rewrite the numerator of Eq. 2 as:

$$\hat{p}\,(w_1, w_2 \mid \ell = LCA)\,\hat{p}\,(LCA) = \frac{S_{<w_1,w_2>\in\ell}}{|descendants(\ell)|^2} \times \frac{N_\ell}{N}. \tag{3}$$

where the first leg of the *rhs* is the class conditional probability of the pair $< w_1, w_2 >$ and the second one is the marginal probability of class ℓ. The term $|descendants(\ell)|$ represents the number of subconcepts of ℓ. Since we could have at most one word sense w_i for each concept c, $|descendants(\ell)|^2$ represents the maximum number of combinations of word senses $< w_1, w_2 >$ which have ℓ as lowest common ancestor. $S_{<w_1,w_2>\in LCA}$ is the number of pairs of senses of word w_1 and w_2 which have LCA as lower common ancestor. The denominator can be written as:

$$\hat{p}\,(w_1, w_2) = \sum_{k\in\mathcal{L}} \frac{S_{<w_1,w_2>\in k}}{|descendants(k)|^2} \times \frac{N_k}{N} \tag{4}$$

Embeddings Evaluation. We generate vectors performing a FastText grid-search (the only method that allows setting the length of character n-grams), then the intrinsic evaluation selects the one which better represents the taxonomy. In essence, we want the similarity between word vectors to reflect as much as possible the semantic similarity between words in the taxonomy. We assess vectors' performance by the Pearson correlation between the cosine similarity of pair of word vectors and the HSS for the same pair of words.

Experimental Results on Benchmarks

Experimental Settings. Using an intel Core i7 machine equipped with 32GB RAM, our experiments rely on (1) a **Taxonomy**: The WordNet provides a structured hierarchy of meanings (senses) and synsets. (2) a **Corpus**: English Wikipedia dump (pre-processed) with no further cleaning.

We evaluate our approach over two benchmark tasks. **Task1: Semantic Similarity.** To demonstrate how $sim_{\texttt{HSS}}$ correlates with the similarity generated by humans, we compute the pairwise Pearson and Spearman correlation between $sim_{\texttt{HSS}}$ and benchmark's similarity. **Task2: Embeddings Evaluation.** To show the performance of HSS in downstream tasks, we compare the embeddings chosen by HSS, with the embeddings chosen using other methods: WUP (Pedersen et al. 2004) and similar to (Baroni, Dinu, and Kruszewski 2014), MEN (Bruni, Tran, and Baroni 2014), and SimLex999 (Hill, Reichart, and Korhonen 2015) benchmarks.

Results. **Task1.** Table 1 shows the results of computing Pearson(P) and Spearman(S) correlations among three human-annotated datasets, MEN, WSS, MT287 (Radinsky et al. 2011), and five SOTA similarity scores, WUP, LC, Shortest Path, Resnik and JC (see e.g. (Jauhiainen et al. 2019)). HSS outperforms the rest of measures (with the exception of SimLex999), in terms of correlations with the mentioned datasets; while focusing on execution time, HSS outperforms the others of one order of magnitude. These results confirm the performance superiority of HSS respect to benchmarks.**Task2.** Table 1 summarises the results of*categorisation* and *classification*. The embedding, which is chosen by HSS outperforms the other three embeddings chosen by MEN and SimLeX999 and WUP measures when applied on two categorisation datasets.

References

Baroni, M.; Dinu, G.; and Kruszewski, G. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, 238–247.

Bruni, E.; Tran, N.-K.; and Baroni, M. 2014. Multimodal distributional semantics. *JAIR* 1–47.

Hill, F.; Reichart, R.; and Korhonen, A. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 665–695.

Jauhiainen, T.; Lui, M.; Zampieri, M.; Baldwin, T.; and Lindén, K. 2019. Automatic language identification in texts: A survey. *JAIR* 675–782.

Malandri, L.; Mercorio, F.; Mezzanzanica, M.; and Nobani, N. 2020a. MEET: A Method for Embeddings Evaluation for Taxonomic Data. In *ICDMW*.

Malandri, L.; Mercorio, F.; Mezzanzanica, M.; and Nobani, N. 2020b. MEET-LM: A Method for Embeddings Evaluation for Taxonomic Data in the Labour Market. *Comp. Ind.* doi:10.1016/j.compind.2020.103341.

Pedersen, T.; Patwardhan, S.; Michelizzi, J.; et al. 2004. WordNet:: Similarity-Measuring the Relatedness of Concepts. In *AAAI*, 25–29.

Radinsky, K.; Agichtein, E.; Gabrilovich, E.; and Markovitch, S. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *WWW*, 337–346.

Schnabel, T.; Labutov, I.; Mimno, D.; and Joachims, T. 2015. Evaluation methods for unsupervised word embeddings. In *EMNLP*, 298–307.

Seco, N.; Veale, T.; and Hayes, J. 2004. An intrinsic information content metric for semantic similarity in WordNet. In *Ecai*, 1089.