Learning Semantic Relatedness from Human Feedback Using Relative Relatedness Learning

Thomas Niebler[1*], Martin Becker[1], Christian Pölitz[1], and Andreas Hotho[1,2]

[1] Data Mining and Information Retrieval Group, University of Würzburg (Germany)
{niebler, becker, poelitz, hotho}@informatik.uni-wuerzburg.de
[2] L3S Research Center Hanover (Germany)

Abstract An important topic in Semantic Web research is to learn ontologies from text. Here, assessing the degree of semantic relatedness between words is an important task. However, many existing relatedness measures only encode information contained in the underlying corpus and thus do not directly model human intuition. To solve this, we propose RRL (Relative Relatedness Learning) to improve existing semantic relatedness measures by learning from explicit human feedback. Human feedback about semantic relatedness is extracted from the publicly available MEN dataset. The core result is that we can generalize human intuition on datasets such as MEN using RRL. This way, we can significantly outperform semantic relatedness scores produced by current state-of-the-art methods.

1 Introduction

An important topic in Semantic Web research is to learn ontologies from text. In this scenario, assessing the semantic relatedness of words as perceived by humans is a crucial task. Often, the relatedness score of two words is approximated by calculating the cosine of their vector representations.

Problem Setting and Approach. While many methods using this approach come close to human intuition, they can only encode information from the underlying corpus and thus do not explicitly represent the actual notion of semantic relatedness as employed by humans. A natural way to solve this is to incorporate explicit human feedback, in order to account for the deviations of the respective semantic relatedness measure from human intuition. This can be achieved using metric learning. However, most metric learning algorithms use constraints such as "*w and w' are similar*" or "*w is more similar to w' than to w''*". With such constraint formulations, human relatedness scores, i.e., absolute information about the degree of relatedness (e.g., "*w and w' are 54% related*"), cannot be used. To address this issue, we propose *Relative Relatedness Learning (RRL)*, which exploits these scores to learn a semantic relatedness measure which fits human intuition by formulating relative constraints in the form "*w_1 is more related to w_1' than w_2 to w_2'*".

Contribution. The core result is that we can generalize human intuition from semantic relatedness datasets using RRL. We can significantly improve the measured semantic relatedness scores beyond the current state-of-the-art. Using two large, public word embedding datasets, we confirm this by learning from and evaluating on the MEN collection, which contains relatedness information generated from human feedback.

Data: $\mathbf{V} \subset \mathbb{S}^{n-1}$: word vectors; $\mathcal{H}$: semantic relatedness dataset (e.g. MEN); learning rate l
Result: a relatedness matrix M for Equation (1)
Let $M := I_n$
while M *not converged* **do**

$$\text{loss}(M) = \sum_{\mathcal{C}(\mathcal{H})} 0.5 \cdot \left(\max\left\{0, \sqrt{1-\cos_M(w_1, w_1')} - \sqrt{1-\cos_M(w_2, w_2')}\right\}\right)^2 + \frac{\text{tr}(M) - \log\det(M)}{n^2}$$

$$M \leftarrow M - l \cdot \nabla\text{loss}(M)$$

$$M \leftarrow \min_{M'} \{\|M - M'\|_{\mathcal{F}} | M' \in PSD\}$$

end
return M

Algorithm 1: The RRL algorithm to learn a relatedness measure from relative relatedness information. M is updated using projected gradient descent, regularization is performed via log det divergence.

2 Relative Relatedness Learning (RRL)

Given in Algorithm 1, we propose a supervised approach with a custom loss function to learn a symmetric, positive semidefinite (PSD) matrix M to parameterize the cosine measure so that it better measures semantic relatedness:

$$\cos_M(x, y) := \frac{x^T M y}{\sqrt{x^T M x}\sqrt{y^T M y}} \tag{1}$$

Our algorithm is inspired by a metric learning approach called LSML [2] which uses relative *distance* comparisons to learn a linear metric characterized by the matrix M. In contrast, the training constraints $\mathcal{C}(\mathcal{H})$ in our algorithm are relative *relatedness* comparisons:

$$\mathcal{C}(\mathcal{H}) := \{(w_1, w_1', w_2, w_2') : rel(w_1, w_1') > rel(w_2, w_2')\}$$

and are collected from semantic relatedness datasets $\mathcal{H} := \{(w_i, w_i', rel(w_i, w_i'))\}$ such as MEN, which contain word pairs (w_i, w_i') together with relatedness scores $rel(w_i, w_i')$ collected from human feedback. Each word is represented by a normalized vector, e.g., from a set of vector embeddings.

3 Datasets

We use two word embedding datasets and a semantic relatedness dataset with relatedness scores collected through human feedback to evaluate RRL.
WikiGloVe [3]. This dataset was trained on 6 billion tokens from Wikipedia articles from a 2014 dump and the Gigaword 5 corpus using the GloVe embedding algorithm and consists of 400,000 vectors with dimension 300.[3]

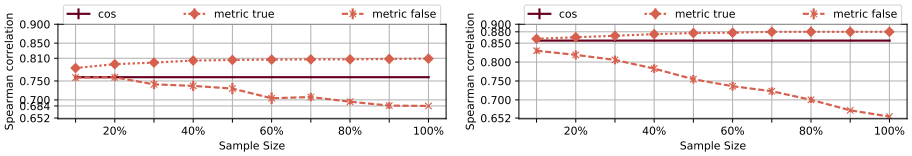
[3] `https://nlp.stanford.edu/projects/glove/`

ConceptNet Numberbatch [4]. Speer et al. combined Word2Vec and GloVe embeddings with relations from the semantic network ConceptNet to receive 426,572 300-dimensional word vectors currently posing the state-of-the-art on MEN.[4]
The MEN collection [1]. The MEN dataset contains 3,000 word pairs together with human-generated scores about their perceived semantic relatedness.[5] These scores reflect human feedback, which we use both to train our relatedness measure as well as for evaluation.

4 Experiments

In this section, we perform two experiments in order to demonstrate the usefulness of RRL for learning semantic relatedness. First we train several metrics on both vector datasets considering different amounts of user feedback and secondly assess the robustness of the learned measures by training on false information. We publish our code to enable reproducibility of our experiments.[6]
Experiment Setup. For both experiments, we randomly split MEN into a 80% training and a 20% test set. In the second experiment, we replace the relatedness scores in the training set by new random scores completely uncorrelated ($\rho < 0.0005$) to the original training scores, while the test scores stay the same. From the training data, we then sampled subsets of different sizes (10% - 100%) on which we train a metric each. The metric is evaluated on the previously sampled 20% test data by applying a standard approach of comparing artificial relatedness scores produced by the metric with human-collected ones using the Spearman correlation coefficient (cf. [3, 4]). We repeat sampling training sets and training a metric 25 times. Then, for each training sample size, we take the mean of the scores produced by the 25 trained metrics. In all training cases, the standard deviation was negligible so we do not report it here. As a baseline, we also report the Spearman correlation using the standard cosine measure on the 20% test data.
Integrating Different Levels of User Intentions. We first investigate how the amount of user feedback used for training influences the quality of the learned semantic relatedness measure. Figure 1 shows that we can inject user feedback information about semantic relatedness into our measure (dashed line, diamond markers) and in doing so, improve the fit of our measure to human intuition significantly. On the ConceptNet embeddings, it appears that we have reached a maximum boundary of achievable correlation on unseen data of 0.88. This is very close to the inter annotator agreement reported in [1]. Furthermore, it is important to note that although the correlation improvements seem very small, i) correlation scores are nonlinear, i.e., improving a high correlation score is much more difficult than improving a low correlation score, ii) with increasing amount of training data, the number of constraints grows roughly quadratically and iii) all differences are significant at $p < 0.05$ with at least 50% training data when comparing mean correlation scores with a Fisher transformation.

[4] https://github.com/commonsense/conceptnet-numberbatch/tree/16.09
[5] http://clic.cimec.unitn.it/~elia.bruni/MEN
[6] http://dmir.org/semmele



(a) WikiGloVe (b) ConceptNet

Figure 1: Results on different amounts of user feedback. Injecting true user feedback leads to a significant increase in correlation on the test data for both WikiGloVe and ConceptNet (diamond, dotted), while injecting false user feedback to RRL lets performance decrease dramatically (star, dashed). The continuous line serves as baseline, i.e., the standard cosine score on the test data. Results with >50% sample size are significant ($p < 0.05$).

Robustness of the Learned Semantic Relatedness Measure. Now we inject false user feedback into RRL to see if we can influence the score in not only a positive, but also a negative direction. Figure 1 shows that false user feedback ($\rho < 0.0005$ to the original scores) exhibits a large negative influence on the learned metric (dotted line, star markers), as expected. Nevertheless, we assume that the score decrease is mitigated by the inherent semantic content of the embeddings. Overall, this shows that although we can improve our measure's fit to human intuition, we need a high semantic quality of both the word embeddings and the latent collected relatedness scores through human feedback. Furthermore, we need a certain minimum amount of training data to produce significantly improved results.

5 Conclusion

In this work, we presented an approach to learn semantic relatedness from human intuition, using a relative constraint formulation. The core result is that we can inject this intuition into a relatedness measure with which we can produce significantly improved results compared to the standard cosine measure and more realistically assess human intuition of semantic relatedness. A noteworthy result is that we can even outperform the current state-of-the-art correlation with MEN on the ConceptNet embeddings, thus defining a new state-of-the-art result on MEN.

Acknowledgements. This work has been partially funded by the DFG grant "Posts II" and the BMBF funded junior research group "CLiGS" (grant identifier FKZ 01UG1408).

References

[1] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. "Multimodal Distributional Semantics." In: *JAIR* (2014).

[2] E. Y. Liu et al. "Metric Learning from Relative Comparisons by Minimizing Squared Residual." In: *ICDM*. Dec. 2012.

[3] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global Vectors for Word Representation." In: *EMNLP*. Vol. 14. 2014.

[4] Robert Speer, Joshua Chin, and Catherine Havasi. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." In: *AAAI*. 2017.