Semantic Relatedness of Wikipedia Concepts – Benchmark Data and a Working Solution

Liat Ein Dor, Alon Halfon, Yoav Kantor, Ran Levy, Yosi Mass, Ruty Rinott, Eyal Shnarch, Noam Slonim
IBM Haifa Reseacrh Lab
Haifa University, Mount Carmel, Haifa, HA 31905, Israel
{liate,alonhal,yoavka,ranl,yosimass,rutyr,eyals,noams}@il.ibm.com

Abstract

Wikipedia is a very popular source of encyclopedic knowledge which provides highly reliable articles in a variety of domains. This richness and popularity created a strong motivation among NLP researchers to develop relatedness measures between Wikipedia concepts. In this paper, we introduce WORD (Wikipedia Oriented Relatedness Dataset), a new type of concept relatedness dataset, composed of 19,276 pairs of Wikipedia concepts. This is the first human annotated dataset of Wikipedia concepts, whose purpose is twofold. On the one hand, it can serve as a benchmark for evaluating concept-relatedness methods. On the other hand, it can be used as supervised data for developing new models for concept relatedness prediction. Among the advantages of this dataset compared to its term-relatedness counterparts, are its built-in disambiguation solution, and its richness with meaningful multi-word terms. Based on this benchmark we develop a new tool, named WORT (Wikipedia Oriented Relatedness Tool), for measuring the level of relatedness between pairs of concepts. We show that the relatedness predictions of WORT outperform state of the art methods.

Keywords: Concept Relatedness, Relatedness function, Dataset

1. Introduction

Wikipedia is the leading open encyclopedia with very good coverage on diverse topics. Many terms encountered in human written texts are in fact Wikipedia concepts. As a result, multiple NLP tasks, such as entity linking and document clustering, can benefit from quantitatively measuring the level of relatedness between such concepts. A variety of methods have been proposed for measuring relatedness between Wikipedia concepts, e.g. (Witten and Milne, 2008; Strube and Ponzetto, 2006; Gabrilovich and Markovitch, 2007; Sherkat and Milios, 2017). Evaluating these methods, and developing new ones naturally requires high quality large benchmark data. Here, we introduce a new dataset, named WORD, composed of $19,276$ Wikipedia concept pairs, manually annotated to determine their level of relatedness.

We exploit these data to assess several types of state of the art semantic-relatedness tools, including word and document similarity functions (Mikolov et al., 2013; Gabrilovich and Markovitch, 2007), and link based methods (Witten and Milne, 2008; Ceccarelli et al., 2013), for the task of concept relatedness. Moreover, we suggest several new utilities, that are explicitly designed for this task, such as PMI (Church and Hanks, 1990; Bullinaria and Levy, 2007) measured between *concepts* based on their statistical co-occurrence in the entire Wikipedia corpus. We further exploit the data for supervised learning of concept-relatedness function, and use our novel utilities along with known state-of-the-art semantic-relatedness methods as features in a Linear Regression (LR) model. The resultant concept-relatedness tool, termed henceforth WORT, clearly outperforms each individual feature.

Finally, to demonstrate the versatility of WORD, we suggest a mechanism for automatically generating a disambiguated term-relatedness dataset from a Wikipedia concept-relatedness dataset. This mechanism, which does not involve additional human annotation, yields a new term-relatedness dataset containing $37,309$ pairs. This dataset, WORD, and the annotation guidelines are all available in `http://www.research.ibm.com/haifa/dept/vst/debating_data.shtml`.

2. Related Work

Most existing semantic relatedness datasets are composed of pairs of *words* (e.g. (Finkelstein et al., 2002; Hill et al., 2014)). Nevertheless, dataets with other element types, such as multi-word terms (Levy et al., 2015b), knowledge base concepts (Ceccarelli et al., 2013) and documents (Lee and Welsh, 2005) are also available.

From all term-relatedness datasets, TR9856 (Levy et al., 2015b) is probably the most similar to the current work, mainly due to its large size, and its generation process. The major difference between the two is in the type of elements composing them, Wikipedia concepts vs. terms, and thus in the relatedness task underlying their scores. Although part of the terms in TR9856 can be linked to concepts, the concepts are not available in the data, and were not provided to the annotators.

Another related work (Ceccarelli et al., 2013) uses links within Wikipedia documents to create a ranking dataset, under the assumption that a concept which is actually mentioned in a document, is more related to other concepts in the same document, than the other, false-positive candidates. Their work differs from ours in several aspects. Their interest in the relatedness task is directly linked to the task of entity disambiguation. Consequently, their goal is learning to rank, which affects their definition of related terms. However, for a variety of NLP tasks, measuring the actual relatedness between concepts, rather than ranking their relatedness, is of interest. We thus rely on human annotators to obtain an actual relatedness score between pairs of concepts, instead of ranking pairs of concepts based on the aforementioned heuristics over Wikipedia.

3. Benchmark Generation

The benchmark generation process is composed of several stages.

3.1. Concept Pairs Selection

The objective of the Concept Pairs Selection stage is to create a balanced population of related and unrelated concept pairs. Clearly, a set of randomly selected pairs is expected to contain a low fraction of related pairs. To overcome this we follow (Levy et al., 2015b) who suggested a similar procedure for n-grams, and hypothesize that concepts which are over represented within a given article, tend to be related to one another. Thus, we first apply the TagMe (Ferragina and Scaiella, 2010) wikification tool, to identify concepts mentioned in the article, and then use the Hypergeometric (HG) test to create a Concept Lexicon (CL) per article, composed of all concepts with a HG p-value ≤ 0.05 after Bonferroni correction.

We selected 123 Wikipedia articles covering various topics like ”affirmative action”, ”vegetarianism” and ”atheism”, and used the above mentioned procedure to generate 123 CLs. From each CL we created 160 concept pairs by randomly selecting 160 concepts from the CL, and pairing 40 of them with the article concept, and the rest 120 with another 120 concepts selected at random from the CL. For too small CLs, the maximal possible number of pairs was selected. In any case, we avoided selecting pairs of identical concepts.

3.2. Annotation

The above process yielded a total of $19,649$ concept pairs that were manually annotated via CrowdFlower, each by 10 annotators. The annotators were presented with a pair of URIs of Wikipedia articles, and were asked to mark them as ”related”, if they believe there is an immediate associative connection between them, and as ”unrelated” otherwise. The annotators were further instructed to consider antonyms as related.

3.3. Post Processing

In order to increase the reliability of our dataset, we applied the following post processing procedure. We analyzed the inter-annotator agreement using Cohen’s kappa coefficient, and filtered out annotators whose average kappa with the other annotators was smaller than 0.25. Annotators that did not have at least 50 common pairs with at least three other annotators were filtered out. The concept pairs that had less than 8 judgments after the annotators’ filtering process, were removed from the benchmark.

3.4. Dataset Statistics

After filtering out some annotators we were left with 247 distinct annotators with average pairwise Cohen’s kappa coefficient of 0.57. Removing concept pairs that were left with < 8 annotations, we ended up with a total of $19,276$ annotated concept pairs covering $10,871$ unique concepts. The relatedness score of each concept pair was computed by averaging over the binary answers of the annotators, yielding a relatedness score in the range $[0,1]$. We refer to these data as WORD, standing for Wikipedia Oriented Relatedness Dataset.

WORD is enriched with related concepts, where more than 50% of the pairs have a positive score, and more than 15% have a score > 0.5. WORD is also enriched with multiword terms (MWTs). 66% of the concepts are MWTs, among which 65% are bigrams. WORD is split into training and test set, where $2/3$ of the topic articles, are in the training set, and the rest are in the test set. The information provided in the data includes the concept URIs and titles; the title of the article from which they were selected; whether they belong to the train or test set; and their relatedness annotation score.

3.5. Advantages Compared to Existing Datasets

Most existing term-relatedness datsets suffer from several issues (Batchkarov et al., 2016; Chiu et al., 2016; Faruqui et al., 2016) for which WORD provides at least partial solutions. To the best of our knowledge, this is the largest dataset of its type, whose number of pairs is large enough to yield high quality learning and evaluation results. Unlike most existing datasets, it is split into a training and a test set, providing a well defined framework for developing supervised relatedness methods, and for comparing between different relatedness functions. Furthermore, we use a cross validation setting, named LOTO, whose advantages are described in Section 4.. Moreover, as a concept-relatedness dataset, it is enriched with meaningful MWTs, unlike most standard term-relatedness datasets that include mostly unigrams, or noisy n-grams. Furthermore, the entities in WORD are disambiguated, since their meaning is provided by their Wikipedia article. As a result, the agreement within the annotators is rather high compared to the typical agreement in term-relatedness tasks. Finally, models that learn from WORD can utilize the large amount of structure data and side information of Wikipedia for improving relatedness predictions.

3.6. Term Relatedness Dataset Generation

We suggest a method for automatically generating a term-relatedness dataset from Wikipedia concept relatedness dataset. Our method is based on redirects and commonness of Wikipedia links, which are used to find, for each concept, a disambiguated and reliable equivalence set of n-grams associated with it. Then, given a concept pair $\{c_1, c_2\}$, these sets are used to create all possible n-gram pairs, and are assigned the score of $\{c_1, c_2\}$. This procedure is repeated for all the pairs in WORD, and results in a term-relatedness dataset of 37,309 pairs.

4. Wikipedia Oriented Relatedness Tool (WORT)

WORT is a tool for measuring relatedness between pairs of concepts. It receives two concepts as input, and returns a value in the range $[0,1]$ indicating their predicted level of relatedness. The tool is based on a Linear Regression algorithm, which outperformed other models, including Nearest Neighbors. For model and feature selection we used a

Leave-One-Topic-Out (LOTO) cross validation setting on the training data, dividing the pairs into folds based on the article topic they originate from. This ensures that pairs included in the training set fold have little overlap with pairs associated with the left-out topic, hence avoiding a biased estimate of the model performance.

4.1. Model Features

WORT has 16 features which we classify into four categories, based on the type of information they rely on: *word-level* distribution, *concept-level* distribution, Wikipedia meta-data and article-text. The word-level distribution features treat a concept as the set of words composing its article title. Specifically, the relatedness between concepts is computed as follows. Let $r(x, y)$ be a word level relatedness measure between words x and y. We define the word-to-concept relatedness between a word x and a concept B to be $max_i r(x, y_i)$, where $\{y1, .., yn\}$ are the words of the article title of concept B. The relatedness between concept A and concept B $CC(A, B)$ is the average word-to-concept relatedness between the words in A and the concept B. The concept level relatedness between A and B is the average over the asymmetric functions $CC(A, B)$ and $CC(B, A)$. The word-level distribution features of the model are cosine similarity between Word2Vec(W2V) representation (Mikolov et al., 2013) and first order positive point mutual information (Church and Hanks, 1990), denoted by PMI1. Notice that word-level second order positive point mutual information (Bullinaria and Levy, 2007) was also computed, but was not included in the model since it did not contribute to its performance, probably due to its similarity to W2V (Levy et al., 2015a). The concept-level distribution features are novel relatedness methods, that adapt the word level measures to the level of concepts. These methods use concepts as basic units in the distribution computations. For examples, all appearances of the terms ”U.S.”, ”U.S.A”, ”United States” etc. are treated as appearances of the concept ”United States”. This unification results in richer statistics about the associated terms. WORT includes the concept-level versions of PMI1, PMI2, and their normalized versions (Bouma, 2009), obtained based on applying TagMe (Ferragina and Scaiella, 2010) to the entire Wikipedia May 2017 dump. The Wikipedia meta-data features, rely on information such as the category and link structure. A thorough list of features from this group is provided in (Ceccarelli et al., 2013), most of which are included in WORT. Finally, WORT contains two *article-text* features, ESA (Gabrilovich and Markovitch, 2007), and the cosine similarity between the tf-idf vectors of the two articles (TFIDF-CS). Other features, such as the Jensen Shannon divergence between the bag-of-words representation of two articles, were omitted from the final model due to their high computational time and their negligible contribution to the overall performance. The full list of the 16 WORT features and their description appears in Table 1. We use $in(e)$ and $out(e)$ to denote the number of incoming and outgoing links of concept e.

word-level distribution features	
PMI1	word level first order positive PMI (Church and Hanks, 1990)
W2V	cosine similarity between the Word2Vec representations of the two concepts (Mikolov et al., 2013)
concept-level distribution features	
NPMI1_C	normalized concept-level first order positive PMI1
PMI1_C	concept level normalized PMI1
PMI2_C	cosine similarity between the PMI1 values of each concept with the 10000 most frequent concepts
Wikipedia meta-data features	
INLINK-PMI	normalized PMI between the in links of the concepts (Ceccarelli et al., 2013)
MW	co-citation based similarity (Witten and Milne, 2008)
COND-PROB	$\frac{1}{2}(\frac{in(a)\cap in(b)}{in(b)} + \frac{in(a)\cap in(b)}{in(a)})$
J-IN	Jaccard similarity: $\frac{in(a)\cap in(b)}{in(a)\cup in(b)}$
J-OUT	Jaccard similarity considering the outgoing links
J-IN-OUT	Jaccard similarity considering the union of incoming and outgoing links
FRIEND	1 if a links to b and b links to a,
DIRECT-LINK	1 if $b \in in(a) \& a \in in(b)$, 0 otherwise
CATEGORY	1 if the two concepts share a Wikipedia category, 0 otherwise
article-text features	
TFIDF-CS	Cosine similarity between the tf-idf vectors of the two concept articles
ESA	Cosine similarity between ESA representation of the concept aricles

Table 1: Full list of WORT features

5. Results

We used the test set of WORD to evaluate the performance of the different features in comparison to WORT. We computed both Spearman and Pearson correlations, to receive a wider picture of the relative strengths and weaknesses of the different features. The strongest features, according to Pearson correlation, are presented in Figure 1. As shown in the figure, each of the four feature categories mentioned in 4.1. has a representative within the five leading features. The strongest feature is the *text-based* feature, TFIDF-CS, which is followed by the *word-level distribution* features, PMI1 and W2V. Next, the novel concept-level normalized positive PMI1 (NPMI1_C) leads the *concept-level distri-*

bution features, while INLINK-PMI, which computes the PMI between the incoming links of the two concepts (Ceccarelli et al., 2013), leads the *Wikipedia meta-data* features. To examine the statistical significance of the obtained feature ranking, we used the MRDS method (Rastogi et al., 2015), which given a pair of features, computes the significance of the difference between their prediction capabilities, based on the difference between their observed correlations, the correlation between them, and the size of the data. Applying this test to the pairs within the top ranked features, the advantage of PMI1 over W2V and INLINK-PMI over PMI1_C was shown to be insignificant. The relative ranking of the other feature pairs was statistically significant, with p-value < 0.05. We further applied MRDS to check the advantage of the highest ranked feature from each category, over the other members of its category. The advantage of PMI1 over W2V was already shown to be insignificant. However, the advantage in the three other feature categories was significant, with p-value < 0.05.

Sorting the features by their *Spearman* correlation results in a slightly different feature hierarchy. The three leading features remain unchanged, but the concept-level features are replaced in the top list with link-structure features. This result implies that the link based features are stronger in concept ranking then in prediction of relatedness level, where the opposite is true for the *concept-level distribution* features.

WORT improvement of Spearman and Pearson correlation compared to the best feature (TFIDF-CS), is 16% and 13% respectively. This improvement is significant for both Spearman and Pearson correlation with MRDS p-value $< 10^{-308}$.

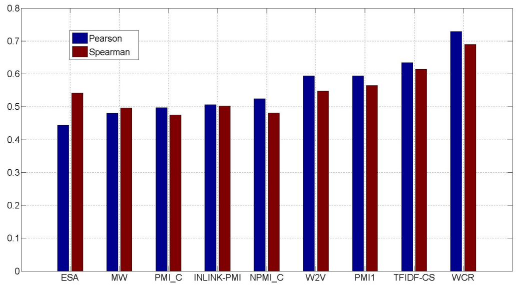

Figure 1: Pearson (blue) and Spearman (red) correlations of WORT (right most bars), and of the highest ranked WORT features according to Pearson correlation. The feature names appear below the bars.

6. Discussion and Future Work

We propose a novel type of semantic relatedness dataset, composed of pairs of Wikipedia concepts manually scored by human annotators. We display its usefulness for evaluating different families of relatedness functions, and study their relative performance. To demonstrate the versatility of WORD, we introduce a mechanism for automatically converting it into an even larger disambiguated term-relatedness dataset. Furthermore, by using the documents of the concept articles, WORD can serve also as a document relatedness dataset which is dramatically larger than the existing dataset (Lee and Welsh, 2005). From the supervised learning point of view, the performance of our new tool, WORT, which despite its simplicity, yields a relatedness function that significantly improves state of the art methods, demonstrates the high synergistic potential that lies in the diversity of the participating features. More complex models can be learned for further improving the relatedness function. For example, one can adopt the approach used by (Mueller and Thyagarajan, 2016) for metric learning of sentences, and use Siamese networks for learning concept representations and a metric between them. Finally, the concept pairs identified as related in WORD probably manifest various types of relatedness, including topical relations, functional relations, and so forth. More work is needed to characterize these relations types and their relative frequency in the data; different prediction methods could then be developed, specializing in predicting specific types of relations between concept pairs.

7. Acknowledgements

The authors would like to thank Vered Shwartz for her contribution to the preliminary stages of this work.

8. Bibliographical References

Batchkarov, M., Kober, T., Reffin, J., Weeds, J., and Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. *ACL 2016*, page 7.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, volume 156.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., and Trani, S. (2013). Learning relatedness measures for entity linking. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 139–148. ACM.

Chiu, B., Korhonen, A., and Pyysalo, S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. *ACL 2016*, page 1.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*.

Ferragina, P. and Scaiella, U. (2010). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *CIKM*.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611.

Hill, F., Reichart, R., and Korhonen, A. (2014). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

Lee, M. D. and Welsh, M. (2005). An empirical evaluation of models of text document similarity. In *In CogSci2005*, pages 1254–1259. Erlbaum.

Levy, O., Goldberg, Y., and Dagan, I. (2015a). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Levy, R., Ein-Dor, L., Hummel, S., Rinott, R., and Slonim, N. (2015b). Tr9856: A multi-word term relatedness benchmark. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 419–424, Beijing, China, July. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mueller, J. and Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In *AAAI*, pages 2786–2792.

Rastogi, P., Van Durme, B., and Arora, R. (2015). Multiview lsa: Representation learning via generalized cca. In *HLT-NAACL*, pages 556–566.

Sherkat, E. and Milios, E. (2017). Vector embedding of wikipedia concepts and entities. *arXiv preprint arXiv:1702.03470*.

Strube, M. and Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424.

Witten, I. and Milne, D. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30.