TR9856: A Multi-word Term Relatedness Benchmark

Ran Levy and **Liat Ein-Dor** and **Shay Hummel** and **Ruty Rinott** and **Noam Slonim**
IBM Haifa Research Lab, Mount Carmel, Haifa, 31905, Israel
{ranl,liate,shayh,rutyr,noams}@il.ibm.com

Abstract

Measuring word relatedness is an important ingredient of many NLP applications. Several datasets have been developed in order to evaluate such measures. The main drawback of existing datasets is the focus on single words, although natural language contains a large proportion of multiword terms. We propose the new TR9856 dataset which focuses on multi-word terms and is significantly larger than existing datasets. The new dataset includes many real world terms such as acronyms and named entities, and further handles term ambiguity by providing topical context for all term pairs. We report baseline results for common relatedness methods over the new data, and exploit its magnitude to demonstrate that a combination of these methods outperforms each individual method.

1 Introduction

Many NLP applications share the need to determine whether two terms are semantically related, or to quantify their degree of "relatedness". Developing methods to automatically quantify term relatedness naturally requires benchmark data of term pairs with corresponding human relatedness scores. Here, we propose a novel benchmark data for term relatedness, that addresses several challenges which have not been addressed by previously available data. The new benchmark data is the first to consider relatedness between multi–word terms, allowing to gain better insights regarding the performance of relatedness assessment methods when considering such terms. Second, in contrast to most previous data, the new data provides a context for each pair of terms, allowing to disambiguate terms as needed. Third, we use a simple systematic process to ensure that the constructed data is enriched with "related" pairs, beyond what one would expect to obtain by random sampling. In contrast to previous work, our enrichment process does not rely on a particular relatedness algorithm or resource such as Wordnet (Fellbaum, 1998), hence the constructed data is less biased in favor of a specific method. Finally, the new data triples the size of the largest previously available data, consisting of $9,856$ pairs of terms. Correspondingly, it is denoted henceforth as **TR9856**. Each term pair was annotated by 10 human annotators, answering a binary question – related/unrelated. The relatedness score is given as the mean answer of annotators where related $= 1$ and unrelated $= 0$.

We report various consistency measures that indicate the validity of TR9856. In addition, we report baseline results over TR9856 for several methods, commonly used to assess term–relatedness. Furthermore, we demonstrate how the new data can be exploited to train an ensemble–based method, that relies on these methods as underlying features. We believe that the new TR9856 benchmark, which is freely available for research purposes, [1] along with the reported results, will contribute to the development of novel term relatedness methods.

2 Related work

Assessing the relatedness between single words is a well known task which received substantial attention from the scientific community. Correspondingly, several benchmark datasets exist. Presumably the most popular among these is the **WordSimilarity-353** collection (Finkelstein et al., 2002), covering 353 word pairs, each labeled by $13-16$ human annotators, that selected a continuous relatedness score in the range 0-10. These hu-

[1] https://www.research.ibm.com/haifa/dept/vst/mlta_data.shtml

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 419–424,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

man results were averaged, to obtain a relatedness score for each pair. Other relatively small datasets include (Radinsky et al., 2011; Halawi et al., 2012; Hill et al., 2014).

A larger dataset is Stanford's Contextual Word Similarities dataset, denoted **SCWS** (Huang et al., 2012) with 2,003 word pairs, where each word appears in the context of a specific sentence. The authors rely on Wordnet (Fellbaum, 1998) for choosing a diverse set of words as well as to enrich the dataset with related pairs. A more recent dataset, denoted **MEN** (Bruni et al., 2014) consists of 3,000 word pairs, where a specific relatedness measure was used to enrich the data with related pairs. Thus, these two larger datasets are potentially biased in favor of the relatedness algorithm or lexical resource used in their development. TR9856 is much larger and potentially less biased than all these previously available data. Hence, it allows to draw more reliable conclusions regarding the quality and characteristics of examined methods. Moreover, it opens the door for developing term relatedness methods within the supervised machine learning paradigm as we demonstrate in Section 5.2.

It is also worth mentioning the existence of related datasets, constructed with more specific NLP tasks in mind. For examples, datasets constructed to assess lexical entailment (Mirkin et al., 2009) and lexical substitution (McCarthy and Navigli, 2009; Kremer et al., 2014; Biemann, 2013) methods. However, the focus of the current work is on the more general notion of term–relatedness, which seems to go beyond these more concrete relations. For example, the words *whale* and *ocean* are related, but are not similar, do not entail one another, and can not properly substitute one another in a given text.

3 Dataset generation methodology

In constructing the TR9856 data we aimed to address the following issues: (i) include terms that involve more than a single word; (ii) disambiguate terms, as needed; (iii) have a relatively high fraction of "related" term pairs; (iv) focus on terms that are relatively common as opposed to esoteric terms; (v) generate a relatively large benchmark data. To achieve these goals we defined and followed a systematic and reproducible protocol, which is described next. The complete details are included in the data release notes.

3.1 Defining topics and articles of interest

We start by observing that framing the relatedness question within a pre-specified context may simplify the task for humans and machines alike, in particular since the correct sense of ambiguous terms can be identified. Correspondingly, we focus on 47 topics selected from Debatabase [2]. For each topic, 5 human annotators searched Wikipedia for relevant articles as done in (Aharoni et al., 2014). All articles returned by the annotators – an average of 21 articles per topic – were considered in the following steps. The expectation was that articles associated with a particular topic will be enriched with terms related to that topic, hence with terms related to one another.

3.2 Identifying dominant terms per topic

In order to create a set of terms related to a topic of interest, we used the Hyper-geometric (HG) test. Specifically, given the number of sentences in the union of articles identified for all topics; the number of sentences in the articles identified for a specific topic, i.e., in the *topic articles*; the total number of sentences that include a particular term, t; and the number of sentences *within the topic articles*, that include t, denoted x; we use the HG test to assess the probability p, to observe $\geq x$ occurrences of t within sentences selected at random out of the total population of sentences. The smaller p is, the higher our confidence that t is related to the examined topic. Using this approach, for each topic we identify all n–gram terms, with $n = 1, 2, 3$, with a p-value ≤ 0.05, after applying Bonfferroni correction. We refer to this collection of n–gram terms as the *topic lexicon* and refer to n–gram terms as n–terms.

3.3 Selecting pairs for annotation

For each topic, we define S_{def} as the set of manually identified terms mentioned in the topic definition. E.g., for the topic "The use of performance enhancing drugs in professional sports should be permitted", $S_{def} = \{$"performance enhancing drugs","professional sports"$\}$. Given the topic lexicon, we anticipate that terms with a small p–value will be highly related to terms in S_{def}. Hence, we define $S_{top,n}$ to include the top 10 n–terms in the topic lexicon, and add to the dataset all pairs in $S_{def} \times S_{top,n}$ for $n = 1, 2, 3$. Similarly, we define $S_{misc,n}$ to include an additional set of 10

[2] http://idebate.org/debatabase

n–terms, selected at random from the remaining terms in the topic lexicon, and add to the dataset all pairs in $S_{def} \times S_{misc,n}$. We expect that the average relatedness observed for these pairs will be somewhat lower. Finally, we add to the dataset $60 \cdot |S_{def}|$ pairs – i.e., the same number of pairs selected in the two previous steps – selected at random from $\cup_{n,m} S_{top,n} \times S_{misc,m}$. We expect that the average relatedness observed for this last set of pairs will be even lower.

3.4 Relatedness labeling guidelines

Each annotator was asked to mark a pair of terms as "related", if she/he believes there is an immediate associative connection between them, and as "unrelated" otherwise. Although "relatedness" is clearly a vague notion, in accord with previous work – e.g., (Finkelstein et al., 2002), we assumed that human judgments relying on simple intuition will nevertheless provide reliable and reproducible estimates. As discussed in section 4, our results confirm this assumption.

The annotators were further instructed to consider antonyms as related, and to use resources such as Wikipedia to confirm their understanding regarding terms they are less familiar with. Finally, the annotators were asked to disambiguate terms as needed, based on the pair's associated topic. The complete labeling guidelines are available as part of the data release.

We note that in previous work, given a pair of words, the annotators were typically asked to determine a relatedness score within the range of 0 to 10. Here, we took a simpler approach, asking the annotators to answer a binary related/unrelated question. To confirm that this approach yields similar results to previous work we asked 10 annotators to re-label the **WS353** data using our guidelines – except for the context part. Comparing the mean binary score obtained via this re-labeling to the original scores provided for these data we observe a Spearman correlation of 0.87, suggesting that both approaches yield fairly similar results.

4 The TR9856 data – details and validation

The procedure described above led to a collection of $9,856$ pairs of terms, each associated with one out of the 47 examined topics. Out of these pairs, $1,489$ were comprised of single word terms (SWT) and $8,367$ were comprised of at least one multi-word term (MWT). Each pair was labeled by 10 annotators that worked independently. The binary answers of the annotators were averaged, yielding a relatedness score between 0 to 1 – denoted henceforth as the *data score*.

Using the notations above, pairs from $S_{def} \times S_{top,n}$ had an average data score of 0.66; pairs from $S_{def} \times S_{misc,n}$ had an average data score of 0.51; and pairs from $S_{top,n} \times S_{misc,m}$ had an average relatedness score of 0.41. These results suggest that the intuition behind the pair selection procedure described in Section 3.3 is correct. We further notice that 31% of the labeled pairs had a relatedness score ≥ 0.8, and 33% of the pairs had a relatedness score ≤ 0.2, suggesting the constructed data indeed includes a relatively high fraction of pairs with related terms, as planned.

To evaluate annotator agreement we followed (Halawi et al., 2012; Snow et al., 2008) and divided the annotators into two equally sized groups and measured the correlation between the results of each group. The largest subset of pairs for which the same 10 annotators labeled all pairs contained roughly 2,900 pairs. On this subset, we considered all possible splits of the annotators to groups of size 5, and for each split measured the correlation of the relatedness scores obtained by the two groups. The average Pearson correlation was 0.80. These results indicate that in spite of the admitted vagueness of the task, the average annotation score obtained by different sets of annotators is relatively stable and consistent.

Several examples of term pairs and their corresponding dataset scores are given in Table 1. Note that the first pair includes an acronym – *wipo* – which the annotators are expected to resolve to *World Intellectual Property Organization*.

4.1 Transitivity analysis

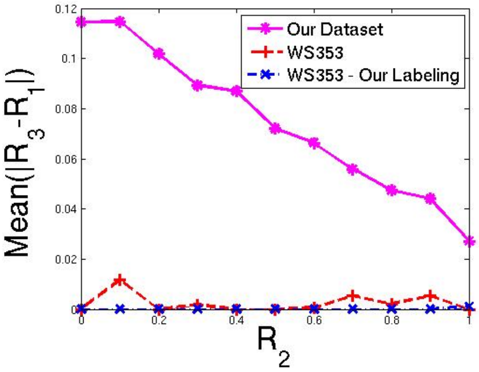
Another way to evaluate the quality and consistency of a term relatedness dataset is by measuring the transitivity of its relatedness scores. Given a triplet of term pairs (a, b) , (b, c) and (a, c), the transitivity rule implies that if a is related to b, and b is related to c then a is related to c. Using this rule, transitivity can be measured by computing the relative fraction of pair triplets fulfilling it. Note that this analysis can be applied only if all the three pairs exist in the data. Here, we used the following intuitive transitivity measure: let (a, b), (b, c), and (a, c), be a triplet of term pairs in the

Term 1	Term 2	Score
copyright	wipo	1.0
grand theft auto	violent video games	1.0
video games sales	violent video games	0.7
civil rights	affirmative action	0.6
rights	public property	0.5
nation of islam	affirmative action	0.1
racial	sex discrimination	0.1

Table 1: Examples of pairs of terms and their associated dataset scores.

dataset, and let R_1, R_2, and R_3 be their relatedness scores, respectively. Then, for high values of R_2, R_1 is expected to be close to R_3. More specifically, on average, $|R_3 - R_1|$ is expected to decrease with R_2. Figure 1 shows that this behavior indeed takes place in our dataset. The p-value of the correlation between $mean(|R_3 - R_1|)$ and R_2 is $\approx 1e - 10$. Nevertheless, the curves of the WS353 data (both with the original labeling and with our labeling) do not show this behavior, probably due to the very few triplet term pairs existing in these data, resulting with a very poor statistics. Besides validating the transitivity behavior, these results emphasize the advantage of the relatively dense TR9856 data, in providing sufficient statistics for performing this type of analysis.

Figure 1: $mean(|R_3 - R_1|)$ vs. R_2.



5 Results for existing techniques

To demonstrate the usability of the new TR9856 data, we present baseline results of commonly used methods that can be exploited to predict term relatedness, including ESA (Gabrilovich and Markovitch, 2007), Word2Vec (W2V) (Mikolov et al., 2013) and first–order positive PMI (PMI) (Church and Hanks, 1990). To handle MWTs, we used summation on the vector representations of W2V and ESA. For PMI, we tokenized each MWT and averaged the PMI of all possible single–word pairs. For all these methods we used the March 2015 Wikipedia dump and a relatively standard configuration of the relevant parameters. In addition, we report results for an ensemble of these methods using 10-fold cross validation.

5.1 Evaluation measures

Previous experiments on **WS353** and other datasets reported Spearman Correlation (ρ) between the algorithm predicted scores and the ground–truth relatedness scores. Here, we also report Pearson Correlation (r) results and demonstrate that the top performing algorithm becomes the worst performing algorithm when switching between these two correlation measures. In addition, we note that a correlation measure gives equal weight to all pairs in the dataset. However, in some NLP applications it is more important to properly distinguish related pairs from unrelated ones. Correspondingly, we also report results when considering the problem as a binary classification problem, aiming to distinguish pairs with a relatedness score ≥ 0.8 from pairs with a relatedness score ≤ 0.2.

5.2 Correlation results

The results of the examined methods are summarized in Table 2. Note that these methods are not designed for multi-word terms, and further do not exploit the topic associated with each pair for disambiguation. The results show that all methods are comparable except for ESA in terms of Pearson correlation, which is much lower. This suggest that ESA scores are not well scaled, a property that might affect applications using ESA as a feature.

Next, we exploit the relatively large size of TR9856 to demonstrate the potential for using supervised machine learning methods. Specifically, we trained a simple linear regression using the baseline methods as features, along with a *token*

Method	r	ρ
ESA	0.43	**0.59**
W2V	**0.57**	0.56
PMI	0.55	0.58

Table 2: Baseline results for common methods.

length feature, that counts the combined number of tokens per pair, in a 10-fold cross validation setup. The resulting model outperforms all individual methods, as depicted in Table 3.

Method	r	ρ
ESA	0.43	0.59
W2V	0.57	0.56
PMI	0.55	0.58
Lin. Reg.	**0.62**	**0.63**

Table 3: Mean results over 10-fold cross validation.

5.3 Single words vs. multi-words

To better understand the impact of MWTs, we divided the data into two subsets. If both terms are SWTs the pair was assigned to the SWP subset; otherwise it was assigned to the MWP subset. The SWP subset included $1,489$ pairs and the MWP subset comprised of $8,367$ pairs. The experiment in subsection 5.2 was repeated for each subset. The results are summarized in Table 4. Except for the Pearson correlation results of ESA, for all methods we observe lower performance over the MWP subset, suggesting that assessing term–relatedness is indeed more difficult when MWTs are involved.

Method	r		ρ	
	SWP	MWP	SWP	MWP
ESA	0.41	0.43	0.63	0.58
W2V	0.62	0.55	0.58	0.55
PMI	0.63	0.55	0.63	0.59

Table 4: Baseline results for SWP vs. MWP.

5.4 Binary classification results

We turn the task into binary classification task by considering the $3,090$ pairs with a data score ≥ 0.8 as positive examples, and the $3,245$ pairs with a data score ≤ 0.2 as negative examples. We use a 10-fold cross validation to choose an optimal threshold for the baseline methods as well as to learn a Logistic Regression (LR) classifier, that further used the token length feature. Again, the resulting model outperforms all individual methods, as indicated in Table 5.

Method	Mean Error
ESA	0.19
W2V	0.22
PMI	0.21
Log. Reg.	**0.18**

Table 5: Binary classification results.

6 Discussion

The new TR9856 dataset has several important advantages compared to previous datasets. Most importantly – it is the first dataset to consider the relatedness between multi–word terms; ambiguous terms can be resolved using a pre–specified context; and the data itself is much larger than previously available data, enabling to draw more reliable conclusions, and to develop supervised machine learning methods that exploit parts of the data for training and tuning.

The baseline results reported here for commonly used techniques provide initial intriguing insights. Table 4 suggests that the performance of specific methods may change substantially when considering pairs composed of unigrams vs. pairs in which at least one term is a MWT. Finally, our results demonstrate the potential of supervised–learning techniques to outperform individual methods, by using these methods as underlying features.

In future work we intend to further investigate the notion of term relatedness by manually labeling the type of the relation identified for highly related pairs. In addition, we intend to develop techniques that aim to exploit the context provided for each pair, and to consider the potential of more advanced – and in particular non–linear – supervised learning methods.

Acknowledgments

The authors thank Ido Dagan and Mitesh Khapra for many helpful discussions.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfre-

und, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. *ACL 2014*, page 64.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49:1–47.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

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

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414. ACM.

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

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden, April. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language resources and evaluation*, 43(2):139–159.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Shachar Mirkin, Ido Dagan, and Eyal Shnarch. 2009. Evaluating the inferential utility of lexical-semantic resources. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 558–566. Association for Computational Linguistics.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.