A Comparison of Word Similarity Performance Using Explanatory and Non-explanatory Texts

Lifeng Jin
Department of Linguistics
The Ohio State University
jin@ling.osu.edu

William Schuler
Department of Linguistics
The Ohio State University
schuler@ling.osu.edu

Abstract

Vectorial representations of words derived from large current events datasets have been shown to perform well on word similarity tasks. This paper shows vectorial representations derived from substantially smaller explanatory text datasets such as English Wikipedia and Simple English Wikipedia preserve enough lexical semantic information to make these kinds of category judgments with equal or better accuracy.

1 Introduction

Vectorial representations derived from large current events datasets such as Google News have been shown to perform well on word similarity tasks (Mikolov, 2013; Levy & Goldberg, 2014). This paper shows vectorial representations derived from substantially smaller explanatory text datasets such as English Wikipedia and Simple English Wikipedia preserve enough lexical semantic information to make these kinds of category judgments with equal or better accuracy. Analysis shows these results may be driven by a prevalence of commonsense facts in explanatory text. These positive results for relatively small datasets suggest vectors derived from slower but more accurate analyses of these resources may be practical for lexical semantic applications.

2 Background

2.1 Wikipedia

Wikipedia is a free Internet encyclopedia website and the largest general reference work over the Internet.[1] As of December 2014, Wikipedia contained over 4.6 million articles[2] and 1.6 billion words. Wikipedia as a corpus has been heavily used to train various NLP models. Features of Wikipedia are well exploited in research like semantic web (Lehmann et al, 2014) and topic modeling (Dumais, 1988; Gabrilovich, 2007), but more importantly Wikipedia has been a reliable source for word embedding training because of its sheer size and coverage (Qiu, 2014), as recent word embedding models (Mikolov et al, 2013; Pennington et al, 2014) all use Wikipedia as an important corpus to build and evaluate their algorithms for word embedding creation.

2.2 Simple English Wikipedia

Simple English Wikipedia[3] is a Wikipedia database where all articles are written using simple English words and grammar. It is created to help adults and children who are learning English to look for encyclopedic information. Compared with full English Wikipedia, Simple English Wikipedia is much smaller. It contains around 120,000 articles and 20 million words, which is almost one fortieth the number of articles and one eightieth the number of words compared to full English Wikipedia, so the average length of articles is also shorter. Simple English Wikipedia is often used in simplification research (Coster, 2011; Napoles, 2010) where sentences from full English Wikipedia are matched to sentences from Simple English Wikipedia to explore techniques to simplify sentences. It would be

[1] See https://en.wikipedia.org/wiki/Wikipedia
[2] See http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia
[3] See http://simple.wikipedia.org/wiki/Main_Page

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 990–994,
Denver, Colorado, May 31 – June 5, 2015.

reasonable to expect that the small vocabulary size of Simple English Wikipedia may be disadvantageous when trying to create word embeddings using co-occurrence information, but it may also be true that despite the much smaller vocabulary size and overall size, because of the explanatory nature of its text, Simple English Wikipedia would still preserve enough information to allow the performance of models trained with Simple English Wikipedia to be comparable to models trained on full Wikipedia, and perform equally well or better than non-explanatory texts like the Google News corpus.

2.3 Word2Vec

The distributed representation of words, or word embeddings, has gained significant attention in the research community, and one of the more discussed works is Mikolov's (2013) word representation estimation research. Mikolov proposed two neusral network based models for word representation: Continuous Bag-of-Words (CBOW) and Skip-gram. CBOW takes advantage of context words surrounding a given word to predict the word by summing all the context word vectors together to represent the word; whereas Skip-gram uses the word to predict the context word vectors for skip-gram positions, therefore making the model sensitive to positions of context words. Both of the models scale well to large quantities of training data, however it is noted by Mikolov that Skip-gram works well with small amounts of training data and provides good representations for rare words, and CBOW would perform better and have higher accuracy for frequent words if trained on larger corpora. The purpose of this paper is not to compare the models, but to use the models to compare training corpora to see how different arrangement of information may impact the quality of the word embeddings.

3 Task Description

To evaluate the effectiveness of full English Wikipedia and Simple English Wikipedia as training corpora for word embeddings, the word similarity-relatedness task described by Levy & Goldberg (2014) is used. As pointed out by Agirre et al (2009) and Levy & Goldberg (2014), relatedness may actually be measuring topical similarity and be better predicted by a bag-of-words model, and similarity may be measuring functional or syntactic similarity and be better predicted by a context-window model. However, when the models are constant, the semantic information of the test words in the training corpora is crucial to allowing the model to build semantic representations for the words. It may be argued that when the corpus is explanatory, more semantic information about the target words is present; whereas when the corpus is non-explanatory, information around the words is merely related to the words. The WordSim353 (Agirre, 2009) dataset is used as the test dataset. This dataset contains pairs of words that are decided by human annotators to be either similar or related, and a similarity or relatedness gold standard score is also given to every pair of words. There are 100 similar word pairs, 149 related pairs and 104 pairs of words with very weak or no relation. In the evaluation task, the unrelated word pairs are discarded from the dataset.

The objective of the task is to rank the similar word pairs higher than related ones. The retrieval/ranking procedure is as follows. First, the cosine similarity scores are calculated using word embeddings from a certain model; then the scores are sorted from the highest to the lowest. The retrieval step is then carried out by locating the last pair of the first n% of the pairs of similar words in the sorted list of scores and determining the percentage of similar word pairs in the sub-list delimited by the last pair of similar words. In other words, the procedure treats similar word pairs as successful retrievals and determines the accuracy rate when the recall rate is n%. Because the accuracy rate would always fall to the percentage of similar word pairs in all word pairs, it is expected that the later and more suddenly it falls, the better the model is performing in this task.

4 Models

The word2vec python implementation provided by gensim (Rehurek et al, 2010) package is used to train all the word2vec models. For Skip-gram and CBOW, a 5-word window size is used to allow them to get the same amount of raw information, also words appearing 5 times or fewer are filtered out. The dimensions of the word embeddings from Skip-gram and CBOW are all 300. Both full English Wikipedia and Simple English Wikipedia are used as training corpora with minimal preprocessing procedures: XML tags are removed and

infoboxes are filtered out, thus yielding four models: Full English Wikipedia – CBOW(FW-CBOW), Full English Wikipedia – Skip-gram(FW-SG), Simple English Wikipedia – CBOW(SW-CBOW) and Simple English Wikipedia – Skip-gram(SW-SG). The pre-trained Google News skip-gram model with 300-dimensional vectors (GN-SG) is also downloaded from the Google word2vec website for comparison. This model is trained on the Google News dataset with 100 billion words, which is 30 times as large as the full English Wikipedia and 240 times as large as Simple English Wikipedia.

5 Results

Table 1 shows the accuracy rate at every recall rate point, with the sum of all the accuracy rates as the cumulative score. It is shown that GN-SG, although not far behind, is not giving the best performance despite being trained on the largest dataset. In fact, it is clear that it never excels at any given recall rate point. It outperforms various models at certain recall rate points by a small margin, but there is no obvious advantage gained from training using a much larger corpus even when compared with the models trained on Simple English Wikipedia, despite the greater risk of sparse data problems on this smaller data set.

For models trained on Simple English Wikipedia and full English Wikipedia, it is also interesting to see that the models almost perform equally well. The FW-CBOW trained on full English Wikipedia performs the best among the models overall, but for the first few recall rate points, it performs equally well or slightly worse than either SW-CBOW or SW-SG trained on Simple English Wikipedia. At the later points, it is also clear that although FW-CBOW is generally better than all the other models most of the time, the margin could be considered narrow and furthermore it is equally as good as SW-CBOW at the first two recall points.

Comparing FW-SG with SW-SG and SW-CBOW, there is almost no sign of performance gain from training using full Wikipedia instead of the much smaller Simple Wikipedia. FW-SG performs equally well or often slightly worse than both Simple Wikipedia models.

The main observation in this paper is that Google News is not out-performing other systems substantially and that full Wikipedia systems are not out-performing Simple Wikipedia substantially (that is, comparing the CBOW models to one another and the Skip-gram models to one another). The main result from the table is not that smaller training datasets yield better systems, but that systems trained using significantly smaller training datasets of explanatory text have very close performances in this task compared with systems trained on very large datasets, despite the big training data size difference.

6 Analysis

As mentioned previously, similarity may be better predicted by a context-window model because it measures functional or syntactic similarity. However, it is not clear in these models that the syntactic information is a major component in the word embeddings. Instead, it may be that the main factor for the performance level of the models is the general explanatory content of the Wikipedia articles, as opposed to the current events content of Google News.

For similar words such as synonyms or hyponyms, the crucial information making them similar is shared general semantic features of the words. For example, for the word pair *physics : chemistry*, the shared semantic features might be that they are both academic subjects, both studied in institutions and both composed of different subfields, as shown in Table 2. The '@' sign in table 2 connects a context word with its position relative to the word in the center of the window. These shared properties

Model	10% Recall Rate	20% Recall Rate	30% Recall Rate	40% Recall Rate	50% Recall Rate	60% Recall Rate	70% Recall Rate	80% Recall Rate	90% Recall Rate	100% Recall Rate	Cumulative Score
FW-CBOW	**0.91**	**0.95**	0.89	**0.83**	**0.72**	**0.74**	**0.61**	**0.51**	**0.46**	0.40	**7.03**
SW-CBOW	**0.91**	**0.95**	0.78	0.75	**0.72**	0.70	0.56	0.50	**0.46**	0.40	6.74
FW-SG	**0.91**	**0.95**	0.79	0.75	0.63	0.61	0.53	0.49	0.43	0.40	6.50
SW-SG	**0.91**	**0.95**	**0.91**	0.70	0.62	0.57	0.54	0.45	0.42	0.40	6.47
GN-SG	0.85	0.84	0.82	0.79	0.70	0.64	0.57	0.48	0.43	0.40	6.51

Table 1: Performance of Different Models at Different Recall Rate Points

of the core semantic identities for these words may contribute greatly to the similarity judgments for humans and machines alike, and these shared properties may be considered general knowledge about the words. For the related words, for example *computer : keyboard*, it may be difficult to pinpoint the semantic overlap between the components which build up the core semantic identities of these words, and none is observed in the data.

General knowledge of a certain word may be found in explanatory texts about the word like dictionaries or encyclopedias, but rarely found in texts other than that. It would be assumed by the writers of informative non-explanatory texts like news articles that the readers are well acquainted with all the basic semantic information about the words, therefore repetition of such information would be unnecessary. For a similarity/relatedness judgment task where basic and compositional semantic information may prove to be useful, using a corpus like Google News, where information or context for a particular word assumes one is already conversant with it, would not be as effective as using a corpus like Wikipedia where general knowledge about a word may be available and repeated. Also, the smaller vocabulary size of Wikipedia compared with Google News would suggest that general knowledge may be conveyed more efficiently with less data sparsity.

In the Simple Wikipedia vs. full Wikipedia case, both corpora are explanatory texts. Despite the much smaller size, the general semantic overlap between each pair of similar words seems as evident in Simple Wikipedia as in full Wikipedia. For measurements like cosine similarity where large values in the same dimensions are favored, the basic semantic components which contribute to the similarity judgments for the words are the same comparatively across two different corpora. This may not be surprising because although more information may be present in full Wikipedia, because of its explanatory nature, the core semantic components which make a concept distinct still dominate over new and sparser information added to it. In Simple Wikipedia, the size of the articles

Word Pair	COAST	SHORE	PHYSICS	CHEMISTRY
Simple Wikipedia	east@-1 164	Lake@2 39	particle@-1 34	organic@-1 86
	west@-1 137	**eastern@-1 25**	chemistry@2 29	physics@-2 29
	south@-1 75	**north@-1 17**	**quantum@-1 28**	physical@-1 21
	north@-1 64	**south@-1 17**	nuclear@-1 23	used@-3 20
	Africa@2 63	**Sea@4 14**	**theoretical@-1 21**	supramolecular@-1 18
	Sea@4 55	**western@-1 14**	**University@3 21**	chemistry@3 17
	Atlantic@-1 53	**northern@-1 12**	laws@-2 21	chemistry@-3 17
	western@-1 52	**southern@-1 11**	mathematical@-1 16	**theoretical@-1 16**
	northern@-1 52	lake@3 10	chemistry@-2 16	physics@2 16
	eastern@-1 50	River@4 8	**professor@-2 14**	placed@3 14
	North@2 46	close@-2 7	mathematics@2 13	biology@2 14
	Australia@2 43	Michigan@3 6	mathematics@-2 13	analytical@-1 12
	southern@-1 40	washed@-3 5	classical@-1 13	**University@3 12**
	Pacific@-1 37	**west@-1 5**	atomic@-1 13	**quantum@-1 12**
	America@3 33	**island@3 5**	modern@-1 11	Organic@-1 11
	city@-4 33	sea@-1 5	Nobel@-3 10	computational@-1 11
	island@3 30	Texas@-1 4	physics@3 10	**professor@-2 11**
Full Wikipedia	**west@-1 16279**	Lake@2 3700	particle@-1 2898	organic@-1 2733
	east@-1 13662	**north@-1 2718**	theoretical@-1 2366	physics@-2 1864
	south@-1 4574	**eastern@-1 2567**	**University@3 2053**	**University@3 1267**
	Atlantic@-1 3741	along@-3 2229	mathematics@-2 1929	physics@2 1192
	north@-1 3497	**western@-1 2163**	chemistry@2 1864	physical@-1 1080
	Pacific@-1 3383	**located@-4 1955**	nuclear@-1 1745	**professor@-2 977**
	western@-1 2802	**south@-1 1908**	laws@-2 1686	biology@-2 886
	southern@-1 2783	**southern@-1 1810**	**quantum@-1 1443**	biology@2 756
	eastern@-1 2771	Lake@3 1645	chemistry@-2 1192	**studied@-1 667**
	Sea@-1 2463	**northern@-1 1628**	**professor@-2 1192**	analytical@-1 633
	America@3 2446	batteries@1 1162	mathematical@-1 1136	inorganic@-1 575
	northern@-1 2383	lake@3 1121	mathematics@2 1032	**degree@-2 559**
	Island@3 2333	Bay@3 1050	matter@-1 786	**quantum@-1 554**
	North@2 2280	River@4 875	**degree@-2 741**	**University@4 517**
	Africa@2 2254	**east@-1 800**	state@-1 737	chemistry@3 418
	located@-4 2177	**west@-1 785**	**University@4 706**	chemistry@-3 418
	island@3 1966	bombardment@1 664	**studied@-1 679**	computational@-1 396

Table 2: Top 17 Context Words that Co-occur with the Sample Similar Word Pairs

and vocabulary may restrict it to be basic and precise to explain a certain concept with fewer presumptions of what the readers already know, and it is suggested by the analysis that such style is also reflected in full Wikipedia, leading to the domination of general knowledge over specific facts.

7 Conclusion

This paper has shown vectorial representations derived from substantially smaller explanatory text datasets such as Wikipedia and Simple Wikipedia preserve enough lexical semantic information to make these kinds of category judgments with equal or better accuracy than news corpora. Analysis shows these results may be driven by a prevalence of commonsense facts in explanatory text. These positive results for small datasets suggest vectors derived from slower but more accurate analysis of these resources may be practical for lexical semantic applications, and we hope by providing this result, future researchers may be more aware of the viability of smaller-scale resources like Simple English Wikipedia (or presumably Wikipedia in other languages which are substantially smaller in size than English Wikipedia), that can still produce high quality vectors despite a much smaller size.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches (pp. 19–27). Association for Computational Linguistics.

Coster, W., & Kauchak, D. (2011). Learning to simplify sentences using Wikipedia (pp. 1–9). Association for Computational Linguistics.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. the SIGCHI conference (pp. 281–285). New York, New York, USA: ACM. doi:10.1145/57167.57214

Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. IJCAI.

Lehmann, J., Isele, R., Jakob, M., & Jentzsch, A. (2014). DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web.

Levy, O., & Goldberg, Y. (2014). Dependency-Based Word Embeddings. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. ICLR Workshop Papers

Napoles, C., & Dredze, M. (2010). Learning simple Wikipedia: A cogitation in ascertaining abecedarian language. NAACL HLT

Pennington, J., & Socher, R. (2014). Glove: Global vectors for word representation. EMNLP.

Qiu, L., Cai, Y., Nie, Z., & Rui, Y. (2014). Learning Word Representation Considering Proximity and Ambiguity (pp. 1572–1578). AAAI Conference on Artificial Intelligence.

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. LREC Workshop on New Challenges for NLP Frameworks, 45–50.