Wordnet-based metrics do not seem to help document clustering

Alexandre Passos[1] and Jacques Wainer[1]

Instituto de Computação (IC)
Universidade Estadual de Campinas (UNICAMP)
Campinas, SP, Brazil

Abstract. In most document clustering systems documents are represented as normalized bags of words and clustering is done maximizing cosine similarity between documents in the same cluster. While this representation was found to be very effective at many different types of clustering, it has some intuitive drawbacks. One such drawback is that documents containing words with similar meanings might be considered very different if they use different words to say the same thing. This happens because in a traditional bag of words, all words are assumed to be orthogonal. In this paper we examine many possible ways of using WordNet to mitigate this problem, and find that WordNet does not help clustering if used only as a means of finding word similarity.

1 Introduction

Document clustering is now an established technique, being used to improve the performance of information retrieval systems [11], as an aide to machine translation [12], and as a building block of narrative event chain learning [1]. Toolkits such as Weka [3] and NLTK [10] make it easy for anyone to experiment with document clustering and incorporate it into an application. Nonetheless, the most commonly accepted model—a bag of words representation [17], bisecting k-means algorithm [19] maximizing cosine similarity [20], preprocessing the corpus with a stemmer [14], *tf-idf* weighting [17], and a possible list of stop words [2]—is complex, unintuitive and has some obvious negative consequences.

One such drawback is the orthogonality of the words in the bag of words representation. Using the cosine similarity, the similarity between two documents $\mathbf{d}_i$ and $\mathbf{d}_j$ is defined as $\mathbf{d}_i.\mathbf{d}_j/(||\mathbf{d}_i||||\mathbf{d}_j||)$. Hence, in a standard bag of words representation, the similarity between a document that only contains the word "cat" and another that only contains the word "kitten" is the same as the similarity between one of them and a document that contains only the word "mobile", 0. In practical clustering systems this is not a deep problem because two documents about the same subject are likely to have at least a few of the subject-specific words in common. Nonetheless, it would be interesting to remove this limitation, possibly improving the quality of the obtained clusters. The simplest way of doing this is by changing the metric of the space of documents, defining the dot product $< \mathbf{a}, \mathbf{b} >$ as $\mathbf{b}^T \Sigma \mathbf{a}$, where Σ is a positive definite matrix.

To assess the similarity between two words one needs to know what they mean, either directly (as in via an ontology, or an electronic dictionary) or indirectly (by inferring meaning from usage). In this paper we concern ourselves with using Wordnet [13] as a source of semantic information with which to compute word similarity. Wordnet has an intuitively satisfying structure—words are represented as having many meanings (each such meaning forming a synset, which is the atomic structure of Wordnet), and relations between words (hyponymy, hyperonymy, antonymy, and other similar relations) are represented as a link in a graph. Many natural measures of similarity on such an ontology exist, such as Resnik's [16], Wu and Palmer's [21], Lin's [9], Leacock-Miller-Chodorow's similarity measure [7], and distance in the wordnet graph. Unfortunately, no such simple measure is any better than a plain bag of words clustering of the same data.

1.1 Related Work

Some work has been done on using Wordnet to improve the performance of clustering and classifying algorithms. The seminal paper is Hotho, Staab and Stumme [4], that shows that enhancing the bag of words with Wordnet synsets from the words in the text and their hypernyms (up to a certain distance) does make better clusters than a plain bag of words representation. As a follow up, Sedding and Kazakov [18] show that using a more precise word sense disambiguator one can obtain even better results than the results by Hotho. Reforgiato [15] uses Wordnet to perform dimensionality reduction prior to clustering. Hung et al. [5] uses a hybrid neural network model guided by Wordnet to cluster documents. Jing et al. [6] uses the same technique as Hotho et al. and enhances it by computing a word similarity measure based on what they call "mutual information" over their clustering corpus. However, their technique didn't produce any considerable improvement over Hotho et al.'s baseline.

While some of these papers prove that Wordnet can indeed be useful in improving document clustering systems, none of them explore removing the *a priori* assumption of word orthogonality using Wordnet. Jing et al. come close, but they do not use Wordnet to learn the Mahalanobis distance used in their paper.

1.2 Contributions

In this paper we evaluate the usage of similarity measures based on Wordnet to aid document clustering, and find that, in general, they do not help. Some measures, such as Resnik similarity [16] and Lin similarity [9] require more information than is avaliable at a blind document clustering task. Other measures, such as Wu-Palmer similarity and Leacock-Miller-Chodorow similarity add too much noise to the document vector that they end up producing close to random clusterings. Measures based only on the distance between the words in the Wordnet graph hurt less, but as they reduce their impact on performance they approach an identity measure. Similarities based on hypernymy reduce to the identity measure most of the time. This work suggests that to go around word orthogonality for document clustering one might have to look below Wordnet, and possibly into sume unsupervised method.

1.3 Structure of this paper

There are many problems involved in trying to extract word similarity data from Wordnet. Section 2 details how we chose which synsets to use for each word. Section 3 then explains in detail which are the similarity measures considered. Section 4 details our experiments, and some tradeoffs we performed when performing this evaluation. Section 5 presents our results and section 6 discusses their meaning and suggests new directions.

2 Choosing the synsets

The first non-trivial step in deriving a similarity measure between words based on Wordnet is choosing how to represent the words in question, since Wordnet, technically speaking, does not concern words per se, but word meanings. Sedding and Kazakov [18] have already found that doing a word sense disambiguation pass, while costly, clearly improves clustering. Since we are interested in studying word similarity (and not word sense similarity) we chose not to replicate that study.

Short of performing a full word sense diambiguation there are two strategies we contemplated in this paper. In the first we chose, between every word pair (w_i, w_j) the synsets from each word that had largest similarity to each other. This approach was really bad, and, for all measures analyzed, produced clustering almost indistinguishable from random clustering. We found out that Wordnet correctly represents many little-used meanings of words, therefore creating artificially higher similarities between many pairs of words, adding a huge amount of noise to the clustering process.

The second strategy we chose was to use, for each word, its most commonly used meaning. This information is easily obtained from Wordnet, and has a high chance of being useful. Nonetheless, we found that it still, in most cases, adds noise to the clustering process. A useful approach we did not evaluate is using the mean similarity for each meaning of the word weighted by its frequency. Unfortunately, Wordnet does not provide detailed usage information.

The synsets used in the evaluation presented in section 5 were the most frequent meanings of each word.

3 Similarity measures

In this paper we evaluate the following measures:

- **lch**: This is the Leacock-Miller-Chodorow similarity, defined as $-log(\frac{p}{2d})$, where p is the distance between the synsets and d the total depth of the taxonomy.

Similarity measure	Mean entropy	stddev	Minimal entropy
lch	266.93	0.74	266.05
wup	251.05	7.21	241.52
path	232.45	11.03	215.06
exp	188.55	28.29	135.69
inv	234.62	8.47	218.60
pow	250.78	6.50	238.50
hyper	256.47	8.93	241.68
hyper2	200.59	19.89	175.72
hyper3	204.82	12.58	180.99
hyper4	213.09	5.36	207.39
hyper5	203.45	11.89	189.06
syn	199.09	19.52	169.56
no-matrix	190.42	8.18	178.84

Table 1. Entropy for the different distance functions.

- **wup**: This is the Wu-Palmer similarity, which is based on the most specific ancestor node of each synset in Wordnet. It is defined as $\frac{2d}{p_1+p_2}$, where d is the depth of the taxonomy and p_1 and p_2 are the distances from the synsets to their most specific ancestor node. When such a node cannot be found, the similarity is 0.
- **inv**: Defined as $\frac{1}{p+1}$, where p is the shortest path between the two synsets.
- **exp**: Defined as e^{-p}, where p is the shortest path between the two synsets.
- **pow**: Defined as 1.1^{-p}, where p is the sortest path between the two synsets.
- **hyper**, **hyper2**, **hyper3**, **hyper4**, **hyper5**: These measures are an attempt to reproduce Hotho's result that adding hypernyms improves clustering performance. **hyper** is 1 whenever the two words share an hypernym; **hyper2** is 1.1^{-d}, where d is the depth of the shared hypernym; **hyper3** is an assymetrical measure that is 1 whenever the first word is a hypernym of the second; **hyper4** is 1 whenever one of the words is a hypernym of the other; **hyper5** is 1 whenever one word has a hypernym with the same name as a hypernym of another word. Measures **hyper2**, **hyper3**, **hyper4**, and **hyper5** often reduce to the identity measure **no-matrix**.
- **syn**: A measure that does not depend on the topology of Wordnet. It is the fraction of the synsets shared by each of the two words considered, or, $|S_a \bigcap S_b|/|S_a \bigcup S_b|$, where S_a is the set of synsets of the first word and S_b the same for the second word.
- **no-matrix**: This is the identity measure; a word is only similar to itself.

4 Experiment design

The evaluation in this paper was carried out using data from the Reuters-21578 [8] corpus. It contains over a million words of news stories, divided in categories like "potato", "instal-debt", "lumber", and others. The data reported in section 5 was obtained using all the stories in the categories "orange", "rubber", "soy-oil", “cocoa”, and “coconut”. They were ramdomly sampled from all categories. It is trivial to modify the source code to use any other set of categories. For distance function we run the clustering algorithm 5 times, each time computing 5 clusters.

To evaluate cluster quality we use total entropy. To compute total entropy, let s_j be the j-th cluster obtained, c_i the i-th category, and f_{ij} the observed frequency of category i in cluster j. Also, let $0 \log 0 = 0$. The formula is:

$$\sum_j |s_j| \sum_i -f_{ij} \log f_{ij}$$

Clusterizations with a smaller entropy separate the categories well, while clusters with a high entropy are close to random samples from the data set.

All the source code used to perform these experiments is available at `http://github.com/alextp/wncluster`. It is written in plain python and uses the NLTK and Numpy[1] libraries, easily obtained in the internet. The code is arranged so that a run of the program computes and displays the results found in this paper, and similar results were obtained in other runs.

5 Results

Table 1 shows the results for clustering the first 10 words of each document. It is easy to see that the complex measures **lch** and **wup** perform considerably worse than the identity measure. Measures directly related to the distance between the synsets (**inv**, **exp**, and **pow**) perform better, although still no better than the baseline. From these, **exp** is the best measure, but it is also a measure in which most of the similarities are close to 0. Of the hypernym measures, no measure outperforms the **no-matrix** baseline. The **syn** measure appears as good as **exp**, but both appresent a very high variation in cluster quality. Still, it is interesting that one of the best measures ignores the topology of Wordnet.

To better elucidate the similarity functions’ behavior, table 2 shows their behavior on three sample documents, O1, O2, and R1. O1 and O2 are the first 20 words of the first and second full-text articles in the Reuters “orange” category, and R1 is the first article in the “rubber” category. Hence, it is clear that, for most functions not of the hypernym family, due to noise added by the similarity, the documents in different categories are as similar to each other as documents in the same category. In this small example, since no hypernyms were present, the hypernym similarities were equivalent to the baseline **no-matrix** similarity.

6 Conclusions and future work

As the evaluation shows, many similarity measures between words derived from Wordnet are either worse than a baseline (as is the case for **wup** and **lch**). This suggests that, for the purposes of text clustering, Wordnet does not provide good

[1] `http://numpy.scipy.org/`

Function	O1×O2	O1×R1	O2×R1
lch	0.66	0.80	0.99
wup	0.62	0.73	0.96
inv	0.34	0.38	0.51
exp	0.05	0.02	0.03
pow	0.54	0.67	0.97
hyper	0.66	0.75	1.04
hyper2	0.05	0.02	0.0
hyper3	0.05	0.02	0.0
hyper4	0.05	0.02	0.0
hyper5	0.05	0.02	0.0
syn	0.05	0.02	0.0
no-matrix	0.05	0.02	0.0

Table 2. An example of inter-class and intra-class similarities.

word similarity data. This might be due to a variety of reasons, one of which is that correctly representing a fuzzy concept such as the similarity between two words is not one of Wordnet's goals, and its structure does not fit well to the task (as seen in the selection of the synsets, in section 2). Also, for example, no measure based directly on Wordnet can relate a verb such as "to seat" to a noun such as chair. This suggests that a data driven approach to obtaining a non-trivial similarity measure between words might be more appropriate than an ontology-based similarity. This measure could be based on co-occurence (words that occur most often in the same documents are more similar), substitutability (words that can be oten used in the same context are similar), co-location (words that are often found very near each other are considered similar), or some more complex way.

References

1. Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL/HLT 2008*, 2008.
2. C. Fox. A stop list for general text. In *ACM SIGIR Forum*, volume 24, pages 19–21. ACM New York, NY, USA, 1989.
3. E. Frank and I.H. Witten. *Data Mining: Practical machine learning tools and techniques with Java implementations.* Morgan Kaufmann, 1999.
4. A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In *Proc. of the SIGIR 2003 Semantic Web Workshop*, pages 541–544, 2003.
5. C. Hung, S. Wermter, and P. Smith. Hybrid neural document clustering using guided self-organization and WordNet. *IEEE Intelligent Systems*, 19(2):68–77, 2004.
6. L. Jing, L. Zhou, M.K. Ng, and J.Z. Huang. Ontology-based distance measure for text clustering. In *Proc. of SIAM SDM workshop on text mining, Bethesda, Maryland, USA*, 2006.
7. C. Leacock, G.A. Miller, and M. Chodorow. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.
8. D. Lewis. Reuters-21578 text categorization test collection, Distribution 1.0, AT&T Labs-Research.

9. N. Lin, B. Wu, R. Jansen, M. Gerstein, and H. Zhao. Information assessment on predicting protein-protein interactions. *BMC bioinformatics*, 5(1):154, 2004.
10. E. Loper and S. Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 62–69, 2002.
11. Christopher Manning, P. Raghavan, and Hinrich Schtze. *Introduction to information retrieval*. Cambridge University Press New York, NY, USA, 2008.
12. D. Manning Christopher and H. Shutze. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press, 1999.
13. G.A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
14. MF Porter. An algorithm for suffix stripping. *Program*, 3(14):130–137, 1980.
15. D. Reforgiato Recupero. A new unsupervised method for document clustering by using WordNet lexical and conceptual relations. *Information Retrieval*, 10(6):563–579, 2007.
16. P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 1999.
17. G. Salton, A. Wong, and CS Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
18. J. Sedding and D. Kazakov. Wordnet-based text document clustering. *ROMAND*, page 104, 2004.
19. M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 34, page 35, 2000.
20. A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Proc. AAAI Workshop on AI for Web Search (AAAI 2000), Austin*, pages 58–64, 2000.
21. Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting of the Association for Computational Linguistics*, pages 133–138. Las Cruces, New Mexico, 1994.