Using Distributional Similarity for Identifying Vocabulary Differences between Individuals

Kapila Ponnamperuma, Chris Mellish, Peter Edwards
Computing Science, University of Aberdeen, Aberdeen AB24 5UA, UK
{k.ponnamperuma,c.mellish,p.edwards}@abdn.ac.uk

Abstract

Distributional Similarity (DS) techniques have been widely used in corpus based thesaurus and taxonomy construction, language modelling, etc. However the application of DS across two corpora is a fairly novel concept, which leads to the identification of co-occurrences of words between corpora. In this paper we have used DS measures to study vocabulary differences (idiolect) of individuals in an interdisciplinary research environment. We have compared the performance of three DS measures including our extension to an established method. Our results show that distributional similarity can be successfully used in identifying vocabulary differences between individuals.

1. Introduction

Distributional Similarity (DS) makes the assumption that similar words share similar grammatical relationships. DS techniques have been widely used in corpus based thesaurus and taxonomy construction, language modelling, etc [11,13]. Motivated by J. Lin's work [8], which investigated terminology differences in cookery recipes by applying DS techniques across two corpora, we investigated the vocabulary differences of authors in an interdisciplinary research environment. Vocabulary differences have been identified as one of the major communication barriers among interdisciplinary researchers, which we intend to address within the *ourSpaces*[1] virtual research environment developed under the *Policy Grid II* project. The application of distributional similarity across two corpora is a fairly novel concept, which leads to the identification of co-occurrences of words between corpora. Our motivation was to establish a measure of similarity/differences of vocabulary between individuals, and in particular to look for situations where two authors from different disciplines might be using different words to express similar concepts.

Although there have been many studies using computational aspects of dialectic variations of languages [10], language variations at the individual level (idiolect) have received little attention. Barlow [1] has used corpus based bigram profiling to study the idiolect of 5 different US press secretaries. Louwerse [9] has studied the idiolect and sociolect of selected authors using Boolean and vector based similarity measures. A considerable body of work can be found in the area of authorship attribution [5] relevant to our work. Among many features used for authorship attribution, use of syntax and part of speech [3] and content words [4] have improved the results considerably. These methods draw some parallels with our approach as both use some syntactic analysis of documents, made possible by the availability of reliable statistical parsers.

In this paper we propose a novel approach based on Distributional Similarity (DS) and deep parsing of documents for studying individual variations in vocabulary. We hypothesise that a single author tends to use the same words in the same context and that one can predict different words being used for a similar concept by comparing the contexts of the words. The next section discusses the methodology including the experiment we carried out to test our hypothesis, followed by results and discussion.

[1] www.ourspaces.net

2. Method

Generally the distributional similarity of two words is computed by comparing corresponding feature vectors, which consist of object, subject, modifier, etc (or inverse relationships depending on whether the target is a verb or a noun) of each word with other words. In this section we discuss the grammatical parsing and feature vectors, feature weights, different DS measures followed by the explanation of our experiment and the evaluation.

Parsing and Constructing Feature vectors: Documents were parsed using the Rapid Accurate Statistical Parser (RASP) [2] and dependency triples involving nouns were extracted. Out of 16 grammatical relationship types used in RASP, direct and indirect objects, subjects and modifiers were selected for our experiment due to the facts that these relationships showed higher parsing precision and recall in RASP and usually higher presence in corpora. We also inferred a new object relationship type from transitive indirect objects (iobj) and direct object (dobj) relationship types as shown in Figure 1, which led to relationships with far richer information content (IC) than the two original relationships on their own by removing the preposition, which carries less information content.

(|iobj| |complexity:16_NN1| |of:17_IO|), (|dobj| |of:17_IO| |issue+s:19_NN2|)
=>(|obj| |complexity:16_NN1| |issue+s:19_NN2|)

Figure 1 – Transitive direct and indirect object relationships (in RASP output format)

The above RASP triples were simplified by removing tags and taking the singular form of each verb. These simplified triples can be represented as *(w1, r1, w2)* where *w1* is the head; *r* is the grammatical relationship (or role) and the *w2* is the dependent of any dependency triplet. Then the feature vectors were constructed for each noun for object, subject and modifier relationship types (roles). A sample feature vector for the word *analysis* is given in Figure 2.

Analysis	*obj_of[conclude, set, square],* *sub_of[affect, find, show, suggest]* *mod[causal, conceptual]*

Figure 2- A sample feature vector for word analysis

Feature Weights: Pairwise mutual information, first presented by [7] and used by much DS related work was chosen as the weighting function in our experiment. Mutual Information (MI) is defined as follows:

$$\mathrm{MI(w, r, w')} = \log_2 \frac{freq(w,r,w')freq(*,r,*)}{freq(w,r,*)freq(*,r,w)}$$

Where *freq(w,r,w')* is the number of triples *(w,r,w')* within the corpus, *freq(*,r,*)* is the number of triples with role r, *freq(*,r,w')* is the co-occurrence of w' with role r and *freq (w,r,*)* is co-occurrence of word w with role r. Positive mutual information indicates that there is a genuine association between w and w' and only positive values are considered for the similarity measure calculations.

Distributional Similarity Measures: Out of many distributional similarity measures [6], D. Lin's [7] Information Theoretic based technique has been used by many studies as a baseline over some of the distance and cosine based techniques, due to its higher accuracy and recall. Weeds et al [11] take document retrieval approaches to word similarity and use a relatively smaller corpus. Yarowsky [14] has claimed that the direct objects of a verb play a more dominant role than its subjects whereas modifiers play a more dominant role for nouns. In line with this finding we propose to enhance D. Lin's measure by assigning a higher (0.5) weight for modifiers when comparing nouns.

Experiment Setup: Our experiment was twofold. First, we investigated the performance of the 3 DS measures described above for examining the vocabulary differences of two researchers from different disciplines. In order to investigate the performance of DS measures, we created two corpora (corpora C and D) by selecting journal papers from two authors in the transport domain. Three distributional measures, namely Lin [7], Weeds [11] and Weighted Lin were calculated for each noun from corpora C and D. For the second part, we divided each corpus into two, allowing us to compare two corpora from the same author as well as corpora from different authors. The best performing measure from the first part of the experiment was selected for this exercise.

Corpus	C1+C2	D1+D2	C1	C2	D1	D2
Author	C	D	C	C	D	D
Triples	16778	16087	9594	7184	8126	7952
Unique Nouns	1360	1021	1397	1091	808	648

Table 1 –Corpus Statistics

Evaluation Methods: For each noun in the first corpus, we compute the similarity with all nouns in the second corpus and then, from all combinations, we select the 20 with the highest scores using each metric. In order to keep the workload of human evaluators at a manageable level, we used the top 20 pairs instead of comparing each word pair. To evaluate the performance of the three DS measures we used theWordNet based Wu and Palmer similarity measure (WuP) [12], (ranges from 0 to 1) and the average similarity (as judged by human evaluators in an online survey). The average human-judged similarity and the average WuP similarity of noun pairs in the top 20 (with identical nouns counted as completely similar with similarity of 1) are measures of overall success in capturing similarity, using DS. The WuP measure is rather conservative, however, and so the human judges were asked to rate any two words that could be used interchangeably by different authors in any context as to some extent similar.

We assume that in general the authors use words similarly and hence that the number of identical nouns occurring in the top 20 should be highest when two corpora from the same author are compared.Therefore for the second part of the experiment, we calculate the percentage of identical and similar noun pairs in the top 20. Similar nouns were selected based on an empirically determined threshold of 0.8 [15] for the WuP measure.

3. Results and Discussion

Table 2 shows the evaluation of 3 DS measures using the WuP measure and human judgment. WuP scored all three measures rather similarly though the weighted Lin measure has a slight edge over Weeds and Lin. However, weighted Lin has a considerable advantage when considering the scores given by human evaluators. Therefore we selected the weighted Lin measure for the second part of our experiment.

Table 3 shows the evaluation of the author comparison results. Trials 1 and 2 consist of corpus pairs from the same author while trials 3 to 6 consist of corpus pairs from different authors. We have calculated the percentage of identical nouns, percentage of similar (not identical) nouns as identified by the WordNet Wu and Palmer measure at threshold of 0.8.

As shown in Table 3, it is clear that using the weighted Lin measure, we found a high percentage (65%-100%) of identical pairs in the top 20 results and a low percentage (0%-5%) of semantically similar (not identical) pairs between corpora from the same author. Of course, if the percentage of identical nouns is large then there is little remaining space to see similar words in the top 20. These results support our hypothesis that the same author would use the same words in similar contexts.

Interestingly, a relatively high percentage (20%-30%) of semantically similar pairs and identical noun pairs were found in the top 20 results in cross author comparisons. This supports our initial intuition that different authors may use semantically similar but different words in the same context.

Method	Wu and Palmer	Human evaluators
Lin	0.819	0.668
Weeds	0.820	0.681
Weighted Lin	0.822	0.764

Table2 – Comparison of DS methods

Trial	Corpora	Identical Nouns (%)	Similar (not identical) nouns (%)
1	C1-C2	65	5
2	D1-D2	100	0
3	C1-D1	20	20
4	C1-D2	20	30
5	C2-D1	25	30
6	C2-D2	15	20

Table 3 – Comparison of corpora

We believe that the feature vectors and frequencies constructed from users' documents, blogs, etc can be successfully used as personal vocabulary models. These personal vocabulary models could be used as the basis for mapping between the words used by different users within the *ourSpaces* virtual research environment, which could in turn bridge bridge communication barriers between members of interdisciplinary projects.

Acknowledgments: The research described here is supported by the UK Economic & Social Research Council (ESRC) under the Digital Social Research programme; award: RES-149-25-1027 and RES-149-25-1075.

References

1. Barlow, M. 2010, "Individual Usage: A Corpus-based Study of Idiolects", 34 International LAUD Symposium.
2. Briscoe, T. 2006, "An introduction to tag sequence grammars and the RASP system parser", Cambridge University.
3. Hirst, G. &Feiguina, O. 2007, "Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts", *LLC,* 22(4), 405-417.
4. Koppel, M., Akiva, N. and Dagan, I. (2006a), "Feature Instability as a Criterion for Selecting Potential Style Markers", Journal of the American Society for Information Science and Technology 57(11), 1519-1525.
5. Koppel, M., Schler, J. &Argamon, S. 2009, "Computational methods in authorship attribution", *J.Am.Soc.Inf.Sci.Technol.,* 60(1), 9-26.
6. Lee, L. 1999, "Measures of Distributional Similarity", Proc of the 37th annual meeting of the ACL, Stroudsburg, PA, USA, 25
7. Lin, D. 1998, "An InformationTheoretic Definition of Similarity", In Proc. of the 15th Int. Conf. on Machine LearningMorgan Kaufmann, 296.
8. Lin, J. 2006. "Using Distributional Similarity to Identify Individual Verb Choice". Procs of the Fourth Int. Natural Language Generation Conference, 33-40, Sydney, ACL
9. Louwerse, M.M. 2004, "Semantic Variation in Idiolect and Sociolect: Corpus Linguistic Evidence from Literary Texts", Computers and the Humanities, 38(2), 207-221.
10. Nerbonne, J. &Kretzschmar, W. 2003, "Introducing Comp. Techniques in Dialectometry.", Computers and the Humanities, 37(3), 245-255.
11. Weeds, J., Dowdall, J., Schneider, G., Keller, B. & Weir, D.J. 2005, "Using distributional similarity to organise biomedical terminology", Terminology, 11(1),107-141.
12. Wu, Z. & Palmer, M.S. 1994, "Verb Semantics and Lexical Selection", *ACL*,133.
13. Yang, D. & Powers, D.M.W. 2008, "Automatic thesaurus construction", ACSC,147.
14. Yarowsky, D. 1993, "One Sense per Collection", ARPA Human Language technology Workshop, 266.
15. Zhang, X., Jing, L., Hu, X., Ng, M. & Zhou, X. 2007, "A comparative study of ontology based term similarity measures on PubMed document clustering", Procs of the 12th int. conference on Database systems for advanced applications, Springer-Verlag, 115.