Paul Nulty and Fintan J. Costello

A comparison of word similarity measures for noun compound disambiguation

Book section [Accepted]

Original citation:
Coyle, Lorcan and Freyne, Jill, (eds.) (2010) *Artificial Intelligence and Cognitive Science: 20th Irish Conference, AICS 2009, Dublin, Ireland, August 19-21, 2009, Revised Selected Papers*. Lecture Notes in Computer Science (6206). Springer, Berlin; Heidelberg, pp. 231-240. ISBN 9783642170799 DOI: 10.1007/978-3-642-17080-5_25

A Comparison of Word Similarity Measures for Noun Compound Disambiguation

Paul Nulty and Fintan Costello

School of Computer Science and Informatics,
University College Dublin,
Dublin 4, Ireland
paul.nulty@ucd.ie
fintan.costello@ucd.ie

Abstract. Noun compounds occur frequently in many languages, and the problem of semantic disambiguation of these phrases has many potential applications in natural language processing and other areas. One very common approach to this problem is to define a set of semantic relations which capture the interaction between the modifier and the head noun, and then attempt to assign one of these semantic relations to each compound. For example, the compound phrase *flu virus* could be assigned the semantic relation *causal* (the virus causes the flu); the relation for *desert wind* could be *location* (the wind is located in the desert). In this paper we investigate methods for learning the correct semantic relation for a given noun compound by comparing the new compound to a training set of hand-tagged instances, using the similarity of the words in each compound. The main contribution of this paper is to directly compare distributional and knowledge-based word similarity measures for this task, using various datasets and corpora. We find that the knowledge based system provides a much better performance when adequate training data is available.

Key words: noun compounds, word similarity, semantic classification, disambiguation

1 Introduction

A noun compound is a noun phrase in which the head noun is modified by another noun, for example *flu virus* or *desert wind*. Noun compounds occur frequently in many languages, and the problem of semantic disambiguation of these phrases has many potential applications in natural language processing and other areas. Search engines which can identify the relations between nouns may be able to return more accurate results. Hand-built ontologies such as WordNet at present only contain a few basic semantic relations between nouns, such as hypernymy and meronymy. If the process of discovering semantic relations from text were automated, more links could quickly be built up. Machine translation and question-answering are other potential applications. Noun compounds are

very common in English, especially in technical documentation and neologisms, while Latin languages tend to favour prepositional paraphrases instead of direct compound translation. To translate compound phrases effectively, knowing the semantic relation that holds between the two nouns is important [1]. Although noun compounds containing three or more nouns are common, in this paper we only consider compounds comprised of two nouns

One very common approach to this problem is to define a set of semantic relations which capture the interaction between the modifier and the head noun, and then attempt to assign one of these semantic relations to each noun compound. For example, the phrase flu virus could be assigned the semantic relation causal (the virus causes the flu); the relation for desert wind could be location (the wind is located in the desert). There is no consensus as to which set of semantic relations best captures the differences in meaning of various noun phrases. Work in theoretical linguistics has suggested that noun-noun compounds may be formed by the deletion of a predicate verb or preposition [2]. However, whether the set of possible predicates numbers 5 or 50, there are likely to be some examples of noun phrases that fit into none of the categories and some that fit in multiple categories.

In this paper we investigate methods for learning the correct semantic relation for a given noun compound by comparing the new compound to a training set of hand-tagged instances. As with all supervised learning approaches, the quality of the system depends on a method of measuring the similarity between a new instance and instances in the training set. The main contribution of this paper is to directly compare distributional and knowledge-based word similarity measures for this task.

2 Noun Compound Similarity

In this paper, we will focus on word similarity measures, i.e., methods for comparing the semantic similarity of two words, rather than two pairs of words. Semantic similarity between pairs of nouns is known as *relational similarity* [3]. Some previous approaches to noun compound disambiguation have used relational similarity measures.

2.1 Relational Similarity

Relational similarity is a measure of the similarity of the semantic relation that occurs in two pairs of words. For example, consider the noun pairs *street, traffic* and *riverbed, water*. *street* and *riverbed* are not highly similar words, and neither are *water* and *traffic*; however, there is a high similarity in the relationship between the pair of words in each case, i.e., water flows along a riverbed, and traffic flows along a street. Relational similarity may be measured using distributional methods by searching a large corpus for sentences in which both parts of a noun compound occur together [3].

To do this, it is necessary to find instances in the corpus where both constituents of the noun compound occur within a narrow window. Certain lexical patterns which occur between two nouns may give a good indication of the semantic relation that holds between the nouns[4]. For example, to compare the compounds *street traffic* and *river water*, a large corpus is used to find strings such as *traffic in the street* or *traffic along the street*. These contexts can then be compared to the contexts for *river water*; which may include similar or identical contexts, for example *water in the river*, *water flows in the river*. Then, without ever directly comparing the words in the compounds, the similarity of the compounds can be judged by the similarity of the strings occurring in their mutual contexts.

The contexts do not directly indicate a particular semantic relation, but they may be used to estimate the relational similarity between the pairs of words, which can then be used, with a training set, to assign a semantic relation to noun-compounds [5, 6]. One drawback of this approach is that it is not always possible to find many occurrences of both constituents within a short window, even using a very large corpus.

2.2 Lexical Similarity

Rather than using relational similarity, we are interested in how well the simpler method of comparing the similarity of the constituents of the compounds directly can work. To illustrate our method, we will consider an ideal example. Given a new noun compound *morning exercise*, which we wish to disambiguate, we may compare this compound to those in our training set of hand-tagged examples, and assign it the semantic relation of the compound which it is most similar to, based on the similarity of its constituent nouns. For this example, the most similar compound in our set might be *summer sport*, which has the relation *temporal* (i.e. the modifier indicates the time period in which the activity described by the head takes place). The similarity score is based on the sum of the similarities of *summer* and *morning*; and *sport* and *exercise*

2.3 Knowledge Based Measures

Knowledge-based word similarity measures work by measuring the distance between two words in a hand-crafted hierarchical knowledge base such as Cyc, Roget's thesaurus or Wordnet. Wordnet is a rich lexical database in which word senses are connected according to their hypernyms and hyponyms, with abstract concepts such as *physical object* and *living entity* near the top of the hierarchy, and more specific terms such as *dog* and *Labrador* below these entries. Each node in the hierarchy corresponds to a word sense or *synset*, rather than an actual token.

There are several similarity measures available which are designed to work using the Wordnet hierarchy. The simplest kind, PATH, counts the number of nodes in the path between the two words in the tree. The inverse of this count is the similarity between the two words. Another kind, LIN, uses this count and

also the information content of the nodes, which may be inferred from a separate corpus or from Wordnet itself [7]. We discuss these further in the experiments section.

2.4 Distributional Measures

Distributional measures of word similarity work by comparing the contexts in which each of the words occurs in a corpus. The simplest method of comparing contexts is known as the 'bag-of words' approach. Given two words, *w1* and *w2*, all words which occur within a certain window, n, of *w1* in a corpus are collected. These words are then compared with a similar 'bag-of-words' collected for *w2*, and the frequencies of words common to the contexts of both *w1* and *w2* are used to calculate the word similarity.

More recent approaches use parsed corpora to include some syntactic information about co-occurrence contexts of words. [14] describes a method for measuring the semantic similarity between two words based on the grammatical relationships which they are found to share in a corpus. The similarity between words is expressed as the sum of the frequencies of arguments to grammatical relations which are shared among both words. The similarity measure also specifies that the sum is weighted by the the probability of a particular argument occurring, so very common words are not given an unduly high weight. The implementation of this method will be discussed further in the following section.

For the experiments in this paper, we chose to use the UKWAC corpus. UKWAC is a very large corpus (over 2 billion tokens) of English text obtained by crawling the .uk web domain. It is annotated with part-of-speech tagging and is lemmatized. The corpus was searched through the Sketch Engine interface [11], which provides an API to many corpora, returning the grammatical relations with which a given word is most associated. Examples of our implementation of the similarity measure proposed in [14] in conjunction with the UKWAC corpus are discussed in the experiments section.

3 Experiments

The motivation for our experiments is investigate which word similarity measure works best for the task of disambiguating noun compounds. One advantage of the word similarity approach over relational similarity is that it does not require the system to have seen instances where both the head and modifier of the compound have occurred in the same sentence in a corpus. Instead, the distributional information about both the head and modifier, separately, is compared with that of the training instance and combined to measure the semantic distance between the compounds.

For all experiments, the classification was carried out using a nearest-neighbor technique, with leave-one-out cross-validation. This means that the semantic relation predicted for each noun-compound in the dataset is the relation of the compound which it is most similar to from the rest of the dataset.

Semantic Relation	Example compounds	Proportion of data
causal	flu virus, onion tear	.18
temporal	summer travel, morning class	.09
spatial	desert wind, home remedy	.12
participant	mail sorter, blood donor	.41
quality	rice paper, picture book	.20

Table 1. Examples of noun compounds and semantic relations from the Nastase and Szpackowicz dataset

Semantic Relation	Example compounds	Proportion of data
be	steel knife	.13
have	street name,	.14
in	forest hut,	.21
inst	rice cooker,	.19
actor	honey bee,	.16
about	fairy tale	.17

Table 2. Examples of noun compounds and semantic relations from the O'Seaghdha dataset

3.1 Datasets

We use two datasets in our experiments. The first dataset was created by Nastase and Szpakowicz [8] and used in experiments by [3]. The data consists of 600 noun-modifier compounds. Of the 600 examples, four contained hyphenated modifiers, for example test-tube baby. These were excluded from our dataset. The data is labeled with two different sets of semantic relations: one set of 30 relations with fairly specific meanings and another set of 5 relations with more abstract relations. Table 1 shows the five relations and some examples. For our research we are particularly interested in noun-noun combinations. Of the 596 examples in the dataset, 325 are clearly noun-noun combinations, e.g. picture book, rice paper, while in the remainder the modifier is an adjective, for example warm air, heavy storm. We used only the noun-noun combinations in our experiments, as this is the focus of our research. Because of the relatively small size of the noun-noun data, we did not experiment with the finer-grained semantic relations, as this subdivision leaves a sparse and unbalanced dataset

Table 1 lists the five semantic relations, example compounds for each relation, and the proportion of examples in the dataset tagged with each relation.

The second dataset we use is a set of 1443 noun compounds annotated with a set of six semantic relations created by O'Seaghdha [9]. This dataset consists of noun sequences extracted from the British National Corpus. Any compounds which were initially tagged by annotators as having an unknown, lexicalised or non-compositional meaning were discarded from an initial set of 2000 compounds. The remaining compounds were each classified with one of 6 semantic relations. The relations, examples, and the distribution of each relation are presented in Table 2.

The full collection and annotation process for this dataset is fully described in [9].

3.2 Wordnet Experiments

Word similarity can be judged in a number of ways, as discussed in section two. For the Wordnet experiments, we use the position of words in the Wordnet hypernym hierarchy as the measure of similarity.

A number of issues arise when using this measure. Firstly, Wordnet is a database of word senses, or *synsets*, rather than tokens or lemmas. Most words can have more than one sense, and the sense distinctions in Wordnet are quite fine-grained. Since we are attempting to disambiguate the compound out of context, the best available method of choosing the correct sense is to assign to each word its most frequent sense. The most-frequent-sense baseline is currently not out-performed by modern contextless word sense disambiguation systems [10].

Secondly, there are a number of possible measures to calculate the difference between two nodes on the hypernym tree. Six Wordnet-based measures are implemented in the python Natural Language Tool Kit (NLTK) [12]. Based on previous work [13] we chose to experiment with four of these. The PATH routine simply counts the number of edges between the two word senses in the Wordnet hierarchy. LCH counts the edges between the senses and also takes into account how deep each of the senses is in the hierarchy. WUP counts the edges between the senses, their depth in the hierarchy, and also the depth of their least common subsumer (deepest common ancestor). The LIN measure, also described in [14], calculates similarity based on the information content of the two Wordnet senses and their least common subsumer. The information content metric is computed using the Brown corpus.

Initially, to compute the similarity of two compounds we simply add the similarity of the heads and the similarity of the modifiers, i.e.:
$sim(A,B) = sim(ModifierA, ModifierB) + sim(HeadA, HeadB)$.

We also experimented with using the product of the word similarities as the compound similarity. To return to our ideal example, the compounds *morning exercise* and *summer sport* should be judged highly similar by the Wordnet measures since *summer* and *morning* share a common ancestor (*time-period*) within just two edges of the Wordnet tree, and *sport* and *exercise* also have a nearby common ancestor (*activity*).

We report accuracy and f-score as our evaluation metrics. Accuracy is simply the percentage of examples which were classified with the correct semantic relation, out of the total number of examples in each dataset. F-score is a more complex measure which balances for different relation class sizes. Table 3 shows results for the Nastase and Szpakowicz dataset, table 4 results for the O'Seaghdha and Copestake dataset. Although the second dataset has 6 rather than 5 possible semantic relations, the majority class backoff baseline is actually lower, since this dataset is more balanced.

Similarity Metric	Sum or Product	Accuracy	F-Score	majority class baseline
LCH	*	.416	.411	.41
LCH	+	.416	.413	.41
Lin	*	**.453**	.448	.41
Lin	+	.447	**.470**	.41
PATH	+	.432	.440	.41
WUP	+	.436	.423	.41

Table 3. Results obtained using Wordnet similarity metrics on the Nastase and Szpackowicz dataset

Similarity Metric	Sum or Product	Accuracy	F-Score	majority class baseline acc
LCH	*	.498	.488	.21
LCH	+	**.501**	**.491**	.21
Lin	*	.474	.467	.21
Lin	+	.491	.484	.21
PATH	+	.497	.486	.21
WUP	+	.492	.483	.21

Table 4. Results obtained using Wordnet similarity metrics on the O'Seaghdha dataset

3.3 UKWAC Experiments

For comparison with the Wordnet knowledge base, we chose the UKWAC corpus as the source for our distributional similarity measures. The UKWAC (UK Web as Corpus) is a large corpus of English documents collected by crawling the .uk web domain [15]. The corpus was constructed by starting out with a seed set of URLs from a variety of domains, and crawling to collect more documents. HTML and other web-noise was stripped from the documents using systems developed for the CLEANEVAL 2007 task. Although there are some biases introduced by using a web-derived corpus, the UKWAC was chosen because its size (more than 2 billion tokens) should allow for detection of even rare grammatical relations among rare words.

In order to implement the similarity measure discussed in [14], we needed to extract grammatical relations from the sentences in the corpus. This was facilitated by the Sketch Engine resource, a web-based corpus query tool [11]. A part-of-speech tagged version of the UKWAC is indexed by this tool, which runs a shallow parser over a target sentence and returns a 'word sketch' containing grammatical relations and their arguments.

The Sketch Engine was queried using a python interface to their web-based javascript API. We retrieved and stored locally the word sketches for each noun involved in one of the compounds in the datasets, and compared the grammatical relation arguments of the constituent nouns for our experiments. To illustrate with an example, for the noun compounds *morning exercise* and *summer sport*, *morning* and *summer* both occur as the subjects of the following verbs: *follow (57.52), wake (53.6), rain (31.12), start (26.63), come (25.01), work (23.97),*

Grammatical Relation	product or sum	Accuracy	F-Score	Baseline
subjectof	*	.412	.326	.41
objectof	*	.452	.420	.41
andor	*	.470	**.468**	.41
combined	*	**.490**	.449	.41

Table 5. Results obtained using distributional similarity metrics on the Nastase and Szpackowicz dataset

Grammatical Relation	product or sum	Accuracy	F-Score	majority class baseline
subjectof	*	.343	.314	.21
objectof	*	**.430**	.418	.21
andor	*	.416	.404	.21
combined	*	.422	**.439**	.21

Table 6. Results obtained using distributional similarity metrics on the O'Seaghdha dataset

The mutual information scores returned by the Sketch Engine system, (displayed after each verb), are summed to give a score of the similarity between the two words.

Again, the system was tested using leave-one-out cross validation. We experimented using the grammatical relations *subject*, *object*, *conjunction* (and/or) and a combination of all these. For this method, we found that using the product, rather than the sum, of the similarities of the components gave better results.

4 Discussion

The best results obtained from both the knowledge-based and distributional word similarity measures are presented in Table 7. In some cases, the ranking of the systems evaluated by f-score is not the same as their ranking by accuracy. f-score is a per-class evaluation measure; the macro-averaged f-score (sum of f-score for each category divided by number of categories) compensates for bias which could be introduced if the number of examples in each class is unnaturally balanced, since it gives equal weight to all classes [3]. To calculate the f-score for each class, we compute precision and recall individiually for each class. F-score is the harmonic mean of precision and recall. However, if the true proportion of compounds per class is close to that of our sample datasets, accuracy is the most relevant measure for applications.

Both measures perform above the majority class baseline for both datasets. The Wordnet-based system clearly achieves the best results on the O'Seaghdha dataset. For the Nastase dataset, the results are less clear. The Wordnet system achieves a better accuracy, while the corpus system achieves a better f-score. It may be that the results are less clear-cut on the second dataset because it is smaller and more unbalanced than the O'Seaghdha set. To test this, we repeated

Condition	Dataset	Accuracy	F-Score	majority class baseline
Wordnet LCH	O'Seaghdha	**.501**	**.491**	.21
Corpus Object.	O'Seaghdha	.433	.418	.21
Corpus Comb.	O'Seaghdha	.422	.439	.21
Wordnet LIN	Nastase	.447	**.470**	.41
Corpus andor	Nastase	.470	.468	.41
Corpus Comb.	Nastase	**.490**	.449	.41

Table 7. Best results on each dataset using the Wordnet-based and distributional similarity measures

Condition	Dataset	Accuracy1443	Accuracy325	Baseline
Wordnet LCH	O'Seaghdha	.501	.429	.21
Corpus Object.	O'Seaghdha	.433	.402	.21
Corpus Comb.	O'Seaghdha	.422	.411	.21

Table 8. Best results on the full O'Seagdha dataset (Accuracy 1443) and a reduced subset of that dataset (Accuracy325)

the experiments which achieved best results on the second dataset, while limiting the available data to 325 randomly chosen instances.

The results (Table 8) show that the Wordnet based method is clearly able to take advantage of the larger dataset better than corpus-based system. However, our implementation of the distributional similarity method could possibly be improved by experimenting with different corpora and different methods of comparing word contexts.

5 Conclusion

We directly compared knowledge-based and distributional word similarity measures for the task of semantically disambiguating noun compounds. We experimented with different measures of Wordnet similarity and different parameters for the corpus similarity technique described in [14] using a very large, web-derived corpus.

Both measures achieved performance well above baseline on both datasets. The results suggest that, given enough data, the Wordnet measure produces better results, even without any word-sense-disambiguation beyond the most-frequent sense heuristic.

Experiments on a random subset of the larger dataset indicates that the Wordnet measure can take advantage of more training data better than the distributional method. Given the availability of large, hand-tagged training sets, the ease of querying resources such as Wordnet quickly, and the expensive nature of indexing and searching gigaword corpora to obtain distributional features, our results suggest that the knowledge based approach is more efficient when lexical similarity is used for disambiguation.

References

1. Johnston, M. and Busa, F.: Qualia structure and the compositional interpretation of compounds In Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons, (1996).
2. Levi, Judith: The Syntax and Semantics of Complex Nominals Academic Press, New York (1978)
3. Turney, P.D., Waterman, M.S.: Similarity of Semantic Relations Computational Linguistics 32(3), 379-416 (2006)
4. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora In Proceedings of Conf. Computational Linguistics (COLING-92), (1992)
5. O Seaghdha, D. and Copestake, A. Using Lexical and Relational Similarity to Classify Semantic Relations. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09), Athens,Greece, (2009)
6. Nakov, P. and Heast, M.: Solving Relational Similarity Problems using the Web as a Corpus In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08), Columbus, OH, (2008)
7. Seco, N., Veale, T. and Hayes, J. : An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In the proceedings of ECAI 2004, the 16th European Conference on Artificial Intelligence. Valencia, Spain. John Wiley. (2004)
8. Nastase, V and Szpakowicz, S. : Exploring noun-modifier semantic relations. In Proceedings of the 5th International Workshop on Computational Semantics, 2003
9. O Seaghdha, Diarmuid, M.S.: Annotating and Learning Compound Noun Semantics. In Proceedings of the ACL-07 Student Research Workshop, Prague, Czech Republic, (2007)
10. Kolhatkar, V and Pedersen, T: WordNet::SenseRelate::AllWords - A Broad Coverage Word Sense Tagger that Maximimizes Semantic Relatedness in the Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies 2009 Conference, June 1-3, 2009, Boulder, CO. (2009)
11. Kilgarriff, A.; Rychly, P.; Smrz, P.; and Tugwell, D. 2004. The Sketch Engine. In Proc. of EURALEX 2004, 105116. (2004)
12. Bird, S., Loper, E.: NLTK: The Natural Language Toolkit In Proceedings of the 42nd meeting o the ACL (Demonstration session),(2004)
13. Kim, S. N. , T. Baldwin. : Automatic interpretation of noun compounds using WordNet similarity. In: Proc. of IJCNLP-05, pp. 945956. (2005)
14. Lin, D.: An Information-Theoretic Definition of Similarity In Proceedings of the 15th International Conference on Machine Learning, Madson, WI, (1998)
15. Ferraresi, A., E. Zanchetta, S. Bernardini and M. Baroni 2008: Introducing and evaluating ukWaC, a very large web-derived corpus of English . In Proceedings of 4th WAC workshop, LREC, Marrakech, Morocco.