Procesamiento del Lenguaje Natural, Revista nº 46 marzo de 2011, pp 59-66 recibido 16-01-11 revisado 21-02-11 aceptado 24-02-11

Interpreting noun compounds using paraphrases

Interpretación de los compuestos nominales mediante paráfrasis

András Dobó, Stephen G. Pulman
Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford, OX1 3QD, United Kingdom
andras.dobo@linacre.oxon.org, stephen.pulman@comlab.ox.ac.uk

Resumen: Los compuestos nominales abundan en inglés y su interpretación es crucial para muchas tareas de procesamiento del lenguaje natural. Proponemos un método para la interpretación automática de los compuestos formados por dos nombres que busca las paráfrasis adecuadas en corpus estáticos y, a continuación, realiza búsquedas con motores de Internet para validarlas. Se reclutaron hablantes nativos para evaluar las paráfrasis obtenidas para los compuestos nominales: las clasificadas en primer, segundo y tercer lugar fueron puntuadas con un promedio de 3,1842; 2,7687 y 2,5583 (en una escala de 1 a 5), respectivamente, lo que se considera un resultado prometedor dada la dificultad de la tarea.
Palabras clave: compuestos nominales, interpretación automática, paráfrasis, semántica léxica

Abstract: Noun compounds are abundant in English and their interpretation is crucial for many natural language processing tasks. We propose a method for automatic two-noun noun compound interpretation that searches for suitable paraphrases in static corpora and then issues Web search engine queries to validate them. Native speakers were recruited to evaluate the returned paraphrases for noun compounds: those ranked first, second and third received an average score of 3.1842, 2.7687 and 2.5583 (on a scale of 1 to 5), respectively, which is considered promising given the difficulty of the task.
Key words: noun compounds, automatic interpretation, paraphrases, lexical semantics

1 Introduction

Written and spoken English is full of noun compounds, which are, following the definition of Downing (1977), sequences of nouns functioning as a single noun. Their interpretation, especially given their abundance, is crucial for many natural language processing tasks, such as machine translation, question answering, information retrieval and information extraction. For example, an information retrieval system, when searching for information on *plastic bottles*, needs to know whether information found on *bottles that are made of plastic* is relevant or not.

At first, using dictionaries for interpreting noun compounds seems to be a feasible idea. However, even for relatively frequent noun compounds, static English dictionaries give low coverage (Butnariu et al., 2009), and according to Séaghdha (2008), the frequency spectrum of noun compounds shows a Zipfian distribution, meaning that most noun compounds display a very low frequency.

This paper investigates the automatic interpretation of two-noun noun compounds using large corpora. Following Wright (2003) and Nakov and Hearst (2006), we believe that interpreting noun compounds with paraphrases is better than using a limited number of abstract relational categories, since there exists an unlimited number of them and they can capture even subtle differences in meaning. We further assume that using a ranked list of several paraphrases is more suitable than using just one paraphrase, as one is often not enough to capture the full meaning of a noun compound. For example, a possible interpretation of the noun compound *malaria mosquito* is the following ranked list of paraphrases:

1. carry
2. spread
3. be infected with

, since a *malaria mosquito* is a mosquito that *carries* / *spreads* / *is infected with* malaria.

The general interpretation method described in this work aims to find those verbs and

ISSN 1135-5948

prepositions in the used corpus, which are suitable for paraphrasing the noun compounds. The basic idea is to search for those sentences that paraphrase the noun compound in focus, count how many times each paraphrase is found with that noun compound, and then create a ranked list of paraphrases based on these frequencies. The search for paraphrases is intended to be done with two static corpora, namely the British National Corpus and the Web 1T 5-gram Corpus. Web search engine queries are then used to validate the results.

2 Related work

2.1 Inventory-based approaches

There are some linguistic theories, such as Levi (1978), which suggest that noun compounds can be divided into a small number of categories based on the semantic relations between their nouns. Many previous noun compound interpretation approaches are based on these theories and aim to interpret noun compounds using a small number of abstract relational categories. For example, Rosario and Hearst (2001) propose 18 abstract classes and apply a standard machine learning algorithm with a domain-specific lexical hierarchy to classify noun compounds from biomedical texts. Nastase and Szpakowicz (2003) propose a method employing machine learning tools to place noun compounds into clusters. This is based on features extracted from WordNet and Roget's Thesaurus and uses 30 clusters, which are grouped into 5 super-categories.

The methods in this category, however, have been criticised for numerous reasons. Although they have the advantage of capturing the generalization of relations in noun compounds, they are constrained by the small number of categories they define (Butnariu et al., 2009). One of the most influential critiques is Downing (1977) who argues that there are so many possible noun compound relations that it is impossible to list all, and that there are many relations that do not fit into any of the standard relationship categories. She also claims that with a limited number of categories, the categories can be ambiguous and noun compounds with different relationships can be assigned to the same category. Furthermore, it is hard to determine which set of relational categories would be best for classifying the relations between noun compounds, since linguists specialized in noun compounds disagree even on the main categories (Lauer, 1995).

2.2 Paraphrasing approaches

A solution for the above mentioned problems is to employ paraphrases for the interpretation of noun compounds instead of predefined abstract semantic categories, with verbs and prepositions as possible paraphrases. By using paraphrases, the number of possible categories is only limited by the vocabulary of the language used, even subtle differences in meaning can be identified and there are no noun compounds that do not fit into any category (Butnariu et al., 2009). Therefore, paraphrasing methods have become popular in recent years.

One of the early automatic noun compound interpretation methods that involves paraphrases is proposed by Lauer (1995). Although using paraphrases, he only uses a small set of eight prepositional paraphrases, therefore this method is actually inventory based, and has the same problems as the other such methods. Nakov and Hearst (2006) and Nakov (2007) propose a method of noun compound interpretation by issuing exact Web search engine queries and extracting a list of paraphrases with their frequencies from the resulting snippets for each noun compound.

There have also been numerous methods proposed to solve the SemEval-2 Task #9 (Butnariu et al., 2009). Given a list of suitable paraphrases for each noun compound, this is a task to return a ranked list of paraphrases for each noun compound based on their aptness. Nulty and Costello (2010) proposed a method based on paraphrase co-occurrence statistics obtained from the training data favouring general paraphrases over less general ones. The best result was obtained by the system proposed by Wubben (2010), which employs a machine learning classifier based on features that were taken from WordNet, the training data and the Web 1T 5-gram Corpus.

3 Method

Our aim was to develop a general paraphrasing method for two-noun noun compound interpretation, such that given a list of noun compounds as its input, it returns a ranked list of paraphrases for each of them, with verbs and prepositions as possible paraphrases.

In almost all noun compounds, the second noun is the head and the first the dependent,

defining a property of the head. The compound of the two nouns behaves syntactically as the head would (Nakov and Hearst, 2006; Lauer, 1995). It will be assumed throughout this work that this holds for the noun compounds to be interpreted. Therefore only such paraphrases are searched for, whose subject is the second noun of the noun compound and whose object is the first noun of the noun compound.

3.1 The two main approaches taken

We took two approaches given the different types of paraphrase extraction and search.

3.1.1 The subject-paraphrase-object-triples version

The first (subject-paraphrase-object-triples) version searches for actual paraphrases for the input noun compounds in the used corpus. For this, it reads through the corpus and counts the frequency of all occurring (subject, paraphrase, object) triples, where:

- *paraphrase* is a verb, *subject* is its subject, and *object* is its direct object
- *paraphrase* is a verb with a preposition, *subject* is its subject, the preposition acts as a particle combining with the verb, and *object* is the direct object of the verb+preposition
- *paraphrase* is a single preposition, which is a non-clausal modifier of *subject*, and *object* is the direct object of the preposition

This is very similar to the extraction method used by Nakov (2007), when extracting features from parsed snippets of Web search query results for paraphrasing noun compounds.

After this paraphrase extraction, for each noun compound it searches for those extracted (subject, paraphrase, object) triples where *subject* is the second noun and *object* is the first noun of the noun compound. This results in a list of paraphrases for each noun compound including their frequency with that noun compound, which is counted as their score. For example, if there are 50 (story, be about, adventure) triples as (subject, paraphrase, object) triples extracted, then this version finds the paraphrase *be about* for the noun compound *adventure story*, with a score of 50.

3.1.2 The subject-paraphrase-and-paraphrase-object-pairs version

The logic behind the second (subject-paraphrase-and-paraphrase-object-pairs) version is that if there is a paraphrase that frequently has the second noun of the noun compound as subject, and it frequently has the first noun of the noun compound as object, then it is assumed that this paraphrase is a suitable one for the noun compound. Therefore, when reading through the used corpus, this version counts the frequency of all occurring (subject, paraphrase) pairs, where:

- *paraphrase* is a verb, and *subject* is its subject
- *paraphrase* is a verb with a preposition, *subject* is its subject and the preposition acts as a particle combining with the verb
- *paraphrase* is a preposition, which is the non-clausal modifier of *subject*

It also counts the frequency of all occurring (paraphrase, object) pairs, where:

- *paraphrase* is a verb, and *object* is its direct object
- *paraphrase* is a verb with a preposition, the preposition acts as a particle combining with the verb, and *object* is the direct object of the verb+preposition
- *paraphrase* is a preposition and *object* is its direct object

Then, for each noun compound, this version searches for such extracted (subject, paraphrase) and (paraphrase, object) pairs, where *subject* is the second noun and *object* is the first noun of the noun compound. This results in two lists of paraphrases for each noun compound, one for the second noun (subject), and one for the first noun (object). To compile the list of suitable paraphrases for the noun compound from these two lists, those paraphrases are searched for that appear in both of them; these are then included in the paraphrase list for the noun compound and their score is calculated from the (subject, paraphrase) and (paraphrase, object) frequencies. However, applying simply frequencies here has a serious problem; whether the noun is the subject or the object, the most frequent verbs combining with all nouns are very common ones, such as *be*, *do* or *make*. When the (subject, paraphrase) and (paraphrase, object) frequencies are combined, the highest scores are achieved by the paraphrases with those verbs not typical of the noun compounds and usually not suitable for paraphrasing them. To avoid this, both in (subject, paraphrase) and (paraphrase, object) relations, mutual information (Church and Hanks, 1989) is used

instead of frequencies. The mutual information of a (subject, paraphrase) pair and the mutual information of a (paraphrase, object) pair is then multiplied together to form a single score for the (noun compound, paraphrase) pair. For example, if there are 40 (bottle, be for) pairs as (subject, paraphrase) pair and 50 (be for, water) pairs as (paraphrase, object) pair extracted, *bottle* occurs 500 times in a (subject, paraphrase) pair, *be for* occurs 2000 times in a (subject, paraphrase) pair, *water* occurs 800 times in a (paraphrase, object) pair, *be for* occurs 1500 times in a (paraphrase, object) pair, there are 2000000 (subject, paraphrase) pairs and there are 1500000 (paraphrase, object) pairs found, then this version finds the paraphrase *be for* for the noun compound *water bottle*, with a score of 37.7153.

However, since a mutual information below 0 is equivalent to a genuine dissociation between the words, only those paraphrases, with a mutual information of the (subject, paraphrase) pair and the (paraphrase, object) pair both above 0 are considered for a noun compound. Furthermore, Church and Hanks (1989) note that mutual information is unstable for very small counts, therefore paraphrases with a (subject, paraphrase) or (paraphrase, object) frequency of at most 5 are also discarded.

In order to make these methods more efficient, all words are lemmatized when extracting the triples and the pairs from the corpus, and when searching for possible paraphrases for noun compounds, the search is conducted with the lemmatized nouns of the noun compound. The lemma for each word is obtained from WordNet (Fellbaum, 1998).

3.2 The corpora and their pre-processing

The British National Corpus and the Web 1T 5-gram Corpus are employed in the search for paraphrases, and the results are validated through Web search engine queries. In order to be able to extract (subject, paraphrase), (paraphrase, object) pairs and (subject, paraphrase, object) triples, the grammatical relations among the words in the corpora need to be identified; for this purpose automatic parsing methods can be used. The instance of the British National Corpus used had already been parsed with the C&C CCG parser (Clark and Curran, 2007) before, so further pre-processing was not needed.

The instance of the Web 1T 5-gram Corpus available was not previously parsed though. Automatically parsing this corpus encounters some problems. First, the n-grams are not complete sentences, so automatically parsing them would result in many errors. The second problem is that automatically parsing all of them (with the C&C CCG parser) would take more than half a year in CPU time. Given the lack of that much time, an alternative approach was chosen, namely tagging the corpus. As tagging also involves many errors on short n-grams, only the 4- and 5-grams were used. Although the grammatical relations cannot be directly obtained from a tagged text, the relations between the words can be inferred from part-of-speech patterns. For example, if a 4-gram has a part-of-speech pattern:

noun verb determiner noun

then it can be assumed that the first noun is the subject of the verb, and the second noun is the object of the verb. Patterns similar to this are used to deduce the grammatical relations inside the n-grams.

3.3 Prepositions

If a paraphrase with a preposition is encountered, then the subject-paraphrase-and-paraphrase-object-pairs version extracts a (subject, paraphrase) pair both including and excluding the preposition. The one without the preposition is extracted, since from the sentence “The professor teaches at a university”, for example, it seems reasonable to extract the (subject, paraphrase) pair (professor, teach); if a (paraphrase, object) pair (teach, anatomy) is also found, the two pairs can be combined to form the paraphrase *teach* for *anatomy professor*. It is necessary to save each (subject, paraphrases) pair with its preposition too, because otherwise this version would not find paraphrases including prepositions. The (paraphrase, object) pairs and (subject, paraphrase, object) triples with prepositions are not specially treated.

3.4 Passive paraphrases

Passive paraphrases are different from other paraphrases, because their surface subject is actually their underlying object. Therefore a (subject, paraphrase) pair with a passive *paraphrase* and without a preposition in fact has the same meaning (at least from our point of view) as the (paraphrase2, object) pair,

where *object* is the same as *subject* and *paraphrase2* is the active form of *paraphrase*. Thus, it makes sense to count their frequency together: whenever a (subject, paraphrase) pair is extracted with a passive *paraphrase* and no preposition, it is saved as a (paraphrase2, object) pair instead, with *object* as the original *subject* and *paraphrase2* as the active version of *paraphrase*. For example, from the sentence "The pizza was eaten", the subject-paraphrase-and-paraphrase-object-pairs version extracts the (paraphrase, object) pair (eat, pizza). Since passive verbs cannot have direct objects, there are no (paraphrase, object) pairs nor (subject, paraphrase, object) triples with a passive *paraphrase* and no preposition.

Furthermore, if a passive paraphrase includes the preposition *by* that refers to a direct object, then that direct object is actually the underlying subject of the paraphrase. Therefore a (subject, paraphrase, object) triple with a passive *paraphrase* and the preposition *by* in effect has the same meaning as the (subject2, paraphrase2, object2) triple, where *subject2* is equal to *object*, *object2* is equal to *subject*, and *paraphrase2* is the active version of *paraphrase* without preposition. Thus, it makes sense to count their frequencies together: if a (subject, paraphrase, object) triple is encountered where *paraphrase* is as described, it is instead extracted as a (subject2, paraphrase2, object2) triple, where *subject2* is the same as *object*, *obejct2* is the same as *subject* and *paraphrase2* is the active version of *paraphrase* without the preposition *by*. For example, from the sentence "This house was built by an architect", the subject-paraphrase-object-triples version extracts the (subject, paraphrase, object) triple (architect, build, house). Moreover, (subject, paraphrase) and (paraphrase, object) pairs with such *paraphrases* are treated very similarly. Passive paraphrases that include a preposition other than *by* do not need to be treated specially.

Because of these conversions, the frequency counts for such (subject, paraphrase, object) triples, (subject, paraphrase) and (paraphrase, object) pairs with passive *paraphrases* and the preposition *by* are stored with their converted version. Therefore, in order to find paraphrases like this for noun compounds, both methods search for such paraphrases for the reverse noun compound (the noun compound where the order of the nouns is changed; it might not be an actual noun compound, but this is not problematic) that are active and have no preposition. If such a paraphrase is found for the reversed noun compound, its passive version with the preposition *by* is then saved for the (not reversed) noun compound, with its score. That is, in order to find paraphrases for the noun compound *band concert* that are passive and have the preposition *by*, the subject-paraphrase-object-triples version searches for such extracted (subject, paraphrase, object) triples where the subject is *band*, the object is *concert* and the paraphrase is active and has no preposition. For example, if there is a triple (band, give, concert), the paraphrase *be given by* is then saved for *band concert* with the score of the (band, give, concert) triple. This works very similarly with the subject-paraphrase-and-paraphrase-object-pairs version.

3.5 Ambitransitive verbs

English verbs can, among other options, be strictly intransitive, strictly transitive, or ambitransitive (Dixon and Aikhenvald, 2000), where the latter means that it functions both transitively and intransitively. The Unaccusative Hypothesis by Perlmutter (1978) proposes two subclasses of intransitive verbs; the unaccusative verbs being those with a surface subject acting as their underlying object (such as *arrive*), and the unergative verbs being those with a surface subject acting as their underlying subject (such as *run*). These two categories can also be applied to ambitransitive verbs; the patientive ambitransitive verbs are unaccusative in their intransitive use and the agentive ambitransitive verbs are unergative in their intransitive use (Mithun, 2000). A typical patientive ambitransitive is *break*; the sentence "The window broke" actually means that someone or something broke the window. A typical agentive ambitransitive is *read*; in the sentence "She reads" *she* is truly the subject of the action.

If a patientive ambitransitive verb is used in its intransitive form, its underlying object (which is its surface subject) is incorrectly extracted as its subject. This can result in paraphrasing errors. There is, however, a solution to this problem. Patientive ambitransitives in their intransitive use behave in the same way as passive verbs; their surface subject is their underlying object. Therefore, patientive ambitransitives in their intransitive form should be treated as if they were passive, which solves the problem. A comprehensive list

of these verbs is given by Levin (1993) in Section 1.1, which is used in this method to identify them.

3.6 Using synonyms, hypernyms, sister words and semantically similar words

Although the two static corpora used seem large enough, no paraphrases for several noun compounds are found in them. Following Kim and Baldwin (2007), it is hypothesised here too, that noun compounds comprising semantically similar words are interpreted in the same way. Thus, in order to improve the recall for noun compound interpretation, instead of just using the nouns in the noun compound when searching for paraphrases, the interpretation method is also tested by using their synonyms, hypernyms, sister words and words that are semantically similar. The synonyms, hypernyms and sister words for each noun are obtained from WordNet, and the semantically similar words for a noun are retrieved through the method proposed by Lin (1998) for measuring word similarity.

3.7 Validation of paraphrases

When searching for paraphrases, especially if using synonyms, hypernyms, sister words or semantically similar words, some of the extracted paraphrases are not correct. In order to improve the results, these paraphrases should be validated by some means. It was decided to use the Web through Web search engines to validate the paraphrases extracted from static corpora. Two search engines were chosen; Google and Yahoo!. It is assumed that if a paraphrase is suitable for a noun compound, at least some Web pages containing the noun compound paraphrased by that suitable paraphrase should show up. First, very simple queries, similar to the ones used by Nakov and Hearst (2006) and Nakov (2007), were tried; for a noun compound *n1 n2* and a paraphrase *p*, all the possible exact queries in the form

"n2Infl THAT p n1Infl"

were issued, where *n1Infl* and *n2Infl* are any of the inflections of *n1* and *n2*, respectively, and *THAT* can be one of the following relative pronouns: *that*, *which* or *who*. The returned page hit counts of all these queries for a (noun compound, paraphrase) pair were then added together to form the Web validation score for that pair. Queries without these relative pronouns were also tested for.

Since these simple queries sometimes do not return a single result even for suitable paraphrases, an extension of the simple method was undertaken by searching for other verb tenses of the paraphrase rather than simple present. Further, queries with wildcards were also tried. The wildcard characters were placed between the paraphrase (p) and the first noun of the noun compound (n1Infl). Queries with up to 9 wildcards were issued.

After searching for a (noun compound, paraphrase) pair on the Web with one of the above described queries, the score of the (noun compound, paraphrase) pair is recalculated from its original score and its Web validation score. This is done as follows:

$$score_{new} = \ln(score_{original} + 1) * \ln(score_{webValidation} + 1)$$

where $score_{original}$ is the original score of the (noun compound, paraphrase) pair, and $score_{webValidation}$ is its Web validation score.

4 Results

The above method was tested on the noun compounds of the test data provided for the SemEval-2 Task #9. This method does not use the gold standard provided for the task, so it also returns paraphrases that are not in that standard, and does not return some that are in it. Therefore the scorer provided for the task is not suitable for the evaluation of this method. Thus, 5 English native speakers were recruited instead who were given the returned set of paraphrases for each noun compound and were asked to score each paraphrase between 1 and 5, 1 meaning that it is completely unsuitable and 5 meaning that it is perfectly suitable. Because of the limited amount of available human resources, the different versions of method were tested manually by us first. In the end, only the method considered to return the best results was evaluated by the recruited judges. Furthermore, for the evaluation only the first 50 nouns of the test data set were taken into account. As we believe that a ranked list of several paraphrases is perfectly suitable to interpret a noun compound, only the best 3 returned paraphrases were evaluated for each noun compound.

By manually testing, the subject-paraphrase-object-triples version used on the Web 1T 5-gram Corpus was found to be the best. It combines the usage of no substitute words with

the usage of sister words; the two returned list of paraphrases for each noun compound are merged after rescaling the scores in the list of the version with the sister words as:

$$score_{new} = \frac{score_{orig} * score_{lowest\_of\_nosubst}}{\text{score}_{highest\_of\_withsisterwords}}$$

where $score_{orig}$ is the original score of the (noun compound, paraphrase) pair, $score_{lowest\_of\_nosubst}$ is the score of the lowest scoring paraphrase for the given noun compound returned by the version not testing for any substitute words, and $\text{score}_{highest\_of\_withsisterwords}$ is the score of the highest scoring paraphrase for the given noun compound returned by the version applying sister words. With this rescaling, the best paraphrase for a noun compound returned by the method using sister words has the same score as the worst paraphrase for that noun compound returned by the method not using any substitute words. The ratio between the scores of the paraphrases returned by the same method remains the same. After the two lists of paraphrases are merged, all the paraphrases are validated using Web search engine queries (see Section 3.7). The different versions of web validation methods were tested on a part of the test data provided for the SemEval-2 Task #9, and evaluated by the scorer provided for the task. The best validation method out of the ones tried was found to be using the Google search engine, with only the simple present tense, without using relative pronouns, and using up to 1 wildcard, therefore this was employed.

Before the human judges' evaluation can be used, a certain agreement between the individual judges needs to be corroborated. In the case of significant disagreement, neither is data provided by them reliable nor can conclusions be deduced from it. The reliability of the data was checked using Krippendorff's alpha measure, which is a standard reliability measure proposed by Krippendorff (2004). The alpha returned was 0.435 for the evaluation provided by the judges, which means that there was significant disagreement between them. Therefore those 39 (noun compound, paraphrase) pairs (out of 150) with a standard deviation of at least 1.5 were discarded. Then the alpha measure became 0.696, which was considered acceptable for this task.

To evaluate the results, the average score given for those paraphrases ranked first, second and third by the method described here were calculated; they are 3.1842, 2.7687 and 2.5583, respectively. These results show that the returned paraphrases are considered moderately suitable on average. Given the difficulty of the task, we regard these as promising results, especially considering that significant disagreement exists about the suitability of paraphrases for noun compounds even among native speakers.

Those 5 noun compounds of the test data set, for which the judges' average score of all the returned (and not omitted) paraphrases are the best and the worst, can be found in Table 1 and Table 2, respectively.

Noun Compound (paraphrases)	**Average score**
broadway youngster (be in)	4.7500
cell membrane (surround)	4.6000
cattle population (be of)	4.4000
arts museum (be of, be devoted to, be for)	4.3333
business sector (be of)	4.2000

Table 1: The 5 best scoring noun compounds

Noun Compound (paraphrases)	**Average score**
anode loss (be at, be)	1.5000
bird droppings (be in, be for, be)	1.2667
bow scrape (be)	1.2500
activity spectrum (be in)	1.0000
altitude reconnaissance (-)	1.0000

Table 2: The 5 worst scoring noun compounds

5 Conclusion and future work

This paper presents a method that interprets two-noun noun compounds by searching for suitable paraphrases for them. It uses static corpora to search for paraphrases and issues Web search queries to validate them. Those paraphrases ranked first, second and third by this method were given an average score of 3.1842, 2.7687 and 2.5583 (on a scale of 1 to 5) by human judges, respectively, which is considered promising given the difficulty of the task.

As related in Section 3.2, due to a lack of time, the Web 1T 5-gram Corpus used in the search for the paraphrases was tagged and not parsed, and the grammatical relations inside the n-grams were deduced based on part-of-speech patterns. As this embodies a much higher error rate than when the relations are obtained with a parser, it is suggested that in the future the Web 1T 5-gram Corpus also be parsed and the general paraphrasing method be tested on that as well. This should improve the results

significantly. Moreover, the results might be further improved by extending the validation part of the method; for example the synonyms, hypernyms, sister words or semantically similar words of the nouns could be employed, or the different extensions could be combined.

References

Butnariu, C., SN. Kim, P. Nakov, DO. Seaghdha, S. Szpakowicz, and T. Veale. 2009. Semeval-2010 Task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden.

Church, KW. and P. Hanks. 1989. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (1):22-29.

Clark, S. and JR. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics* 33 (4):493-552.

Dixon, R. and A. Aikhenvald. 2000. Introduction. In *Changing valency: Case studies in transitivity*, edited by R. Dixon and A. Aikhenvald. Cambridge, UK: Cambridge University Press.

Downing, P. 1977. On the creation and use of English compound nouns. *Language* 53 (4):810-842.

Fellbaum, C. 1998. *WordNet: An electronic lexical database*. Cambridge, MA, USA: The MIT Press.

Kim, SN. and T. Baldwin. 2007. Interpreting noun compounds using bootstrapping and sense collocation. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*. Melbourne, Australia.

Krippendorff, K. 2004. *Content analysis: An introduction to its methodology*. Thousand Oaks, CA, USA: Sage Publications.

Lauer, M. 1995. Designing statistical language learners: Experiments on noun compounds, Macquarie University, Sydney, Australia.

Levi, JN. 1978. *The syntax and semantics of complex nominals*. New York, NY, USA: Academic Press.

Levin, B. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago, Il, USA: The University of Chicago Press.

Lin, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*. Madison, WI, USA.

Mithun, M. 2000. Valency-changing derivation in Central Alaskan Yup'ik. In *Changing valency: case studies in transitivity*, edited by R. Dixon and A. Aikhenvald. Cambridge, UK: Cambridge University Press.

Nakov, P. 2007. Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics, University of California at Berkeley, Berkeley, CA, USA.

Nakov, P. and M. Hearst. 2006. Using Verbs to Characterize Noun-Noun Relations. In *Artificial Intelligence: Methodology, Systems, and Applications*, edited by J. Euzenat and J. Domingue. Berlin / Heidelberg, Germany: Springer.

Nastase, V. and S. Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Proceedings of the 5th International Workshop on Computational Semantics*. Tilburg, The Netherlands.

Nulty, P. and F. Costello. 2010. UCD-PN: Selecting General Paraphrases Using Conditional Probability. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden.

Perlmutter, D. 1978. Impersonal passives and the unaccusative hypothesis. In *Proceedings of the 4th Annual Meeting of the Berkeley Linguistics Society*. Berkeley, CA, USA.

Rosario, B. and M. Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. Pittsburgh, PA, USA.

Séaghdha, DO. 2008. Learning compound noun semantics, University of Cambridge, Cambridge, UK.

Wright, DGS. 2003. Noun-verb associations for Noun-Noun Compound Interpretation. *Oxford University Working Papers in Linguistics, Philology & Phonetics* 8:175-190.

Wubben, S. 2010. UvT: Memory-based pairwise ranking of paraphrasing verbs. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden.