Natural Language Engineering **1** (1): 1–23. Printed in the United Kingdom

A Lexical Semantic Approach to Interpreting and Bracketing English Noun Compounds

Su Nam Kim♠ and Timothy Baldwin♡

♠ *Faculty of Information Technology*
Monash University

♡ *NICTA Victoria Research Laboratories*
Department of Computing and Information Systems
The University of Melbourne

(*Received day month year; revised day month year*)

Abstract

This paper presents a study on the interpretation and bracketing of noun compounds ("NCs"), based on lexical semantics. Our primary goal is to develop a method to automatically interpret NCs through the use of semantic relations. Our NC interpretation method is based on lexical similarity with tagged NCs, based on lexical similarity measures derived from WordNet. We apply the interpretation method to both 2-term and 3-term NC interpretation based on semantic roles. Finally, we demonstrate that our NC interpretation method can boost the coverage and accuracy of NC bracketing.

1 Introduction

This paper is concerned with determining the **semantic relations** ("SRs") in English noun compounds ("NCs"), as a means of representing the semantic connection between the head noun and modifier(s) of the NC. In this paper, we use the term **noun compound** or **NC** to refer to a sequence of two or more nouns, and refer to the individual nouns in the NC as "components". NC semantic relations are notoriously nuanced in nature; for example, *apple juice* is interpreted as "juice made from apple(s)" while the superficially highly-similar *morning juice* means "juice (drunk) in the morning" and *box juice* means "juice that is contained in a box".

Disambiguating NCs with semantic relations has a long history in NLP, due to the allure of a richer semantic representation to build off in applications such as question-answering and machine translation (Lauer, 1995; Venkatapathy and Joshi, 2006). However, due to difficulties such as high productivity, interaction between the interpretation and context of use, and semantic delineation of a discrete SR set (Downing, 1977; Spärck Jones, 1983; Lapata and Keller, 2004), interpreting SRs has been found to be a challenging task for NLP (Nakov, 2008; Butnariu and Veale, 2008; Baldwin and Kim, 2009; Ó Séaghdha, 2009; Tratz and Hovy, 2010).

As detailed in Section 2, recent attempts at automatic NC interpretation have taken two

Kim, Su Nam and Timothy Baldwin (2013) A Lexical Semantic Approach to Interpreting and Bracketing English Noun Compounds, Natural Language Engineering 19(3), pp. 385-407.

basic approaches: analogy-based interpretation (Rosario and Hearst, 2001; Girju et al., 2005; Kim and Baldwin, 2005; Turney and Littman, 2005; Girju, 2007; Ó Séaghdha and Copestake, 2007; Ó Séaghdha, 2009; Tratz and Hovy, 2010) and semantic disambiguation relative to an underlying predicate or paraphrase (Vanderwende, 1994; Lapata, 2002; Kim and Baldwin, 2006; Nakov and Hearst, 2006; Nakov, 2008; Butnariu and Veale, 2008). Our proposed approach uses analogy-based interpretation, taking WordNet as the basis to calculate lexical similarity. The intuition behind our method is that NCs made up of similar component words should have the same SR. That is, given the NC *apple juice* and the SR `made_of(N2,N1)` (i.e. N2 = the head noun [*juice*] is made of N1 = the modifier [*apple*]), we would expect the unseen NC *orange juice* to have the same SR of `made_of(N2,N1)`. To measure the semantic similarity between NCs, we calculate the lexical similarity between the corresponding components (head nouns and modifiers, respectively, of the two NCs). We first test our method over our own 2-term NC dataset, and then go on to apply it and two other methods to the dataset used for Task 4 at *SemEval-2007* (Girju et al., 2007). Second, we investigate the relative contribution of components on the similarity calculation. Third, we evaluate our method over a 3-term NC dataset developed for this research, in what we believe are the first published results over 3-term NC interpretation. Finally, we present an NC bracketing method which makes use of predicted interpretations between component words, and demonstrate that our method is able to boost the coverage and accuracy of a state-of-the-art NC bracketing method.

The remainder of the paper is structured as follows. In Section 2, we review the literature on NCs, and in Section 3, we detail the set of semantic relations used in this paper. In Section 4, we detail our motivation and basic approach to NC interpretation, and the datasets and resources used in our evaluation in Section 5. We then present the results for our method over the different datasets in Section 6, and finally conclude in Section 7.

2 Related Work

The primary focus of research on NCs has been on interpreting them with semantic relations, but there has also been work on syntactic analysis (i.e. bracketing) of NCs. We review the related work on these two areas in this section.

2.1 Interpreting Noun Compounds

Vanderwende (Vanderwende, 1994) presented one of the very first attempts to automatically interpret the semantics of NCs, based on a discrete set of SRs. She used semantic information automatically extracted by analyzing definitions in an online dictionary, and interpreted the NCs through the semantics of verbs corresponding to each relation. One drawback was that the system employed hand-written rules, making it labour-intensive and hard to repurpose to new domains or relation sets. Barker and Szpakowicz (Barker and Szpakowicz, 1998) developed a larger, more general-purpose set of SRs (detailed in Table 1), and evaluated a semi-automatic NC interpretation method over instances from the technical repair domain. More recent work on SR-based interpretation has focused on fully automatic methods, over a broad range of different SR sets (Fan et al., 2003; Girju, 2007; Ó Séaghdha, 2009; Tratz and Hovy, 2010).

Some research on NC interpretation has focused specifically on compound nominalizations (Isabelle, 1984; Hull and Gomez, 1996; Lapata, 2002; Grover et al., 2004), that is NCs where the head noun is a deverbal noun such as *animation* (derived from *animate*). Lapata (Lapata, 2002) proposed a fully automatic method for interpreting compound nominalizations, based on the assumption that the modifier is interpretable as either the subject (e.g. *child behaviour*) or object (e.g. *car lover*) of the base verb expressed by the head noun. Grover et al. (Grover et al., 2004) and Nicholson and Baldwin (Nicholson and Baldwin, 2005) extended this work in considering a wider selection of interpretation types, based around preposition selection.

In order to bound the set of NCs and SRs, some previous work has focused on NC interpretation in specific domains. Rosario and Hearst (Rosario and Hearst, 2001) used hierarchical tagged nouns from biomedical texts, and classified them according to a domain-specific set of SRs using neural networks. Grover et al. (Grover et al., 2004) performed compound nominalization interpretation over biomedical data. Nakov and Hearst (Nakov and Hearst, 2006) also focused on the biomedical domain, using verb semantics based on a web corpus.

Moldovan et al. (Moldovan et al., 2004) and Girju et al. (Girju et al., 2005) proposed methods for open-domain NC interpretation using the pairing of word senses of the component words. Later, Girju (Girju, 2007) integrated cross-lingual features from five Romance languages into the same method and showed that cross-lingual information enhances NC interpretation accuracy. Kim and Baldwin (Kim and Baldwin, 2005) used a method based on nearest-neighbour classification over the union of senses of the modifier and head noun, that forms the basis of this paper. In separate work, they also proposed a method for interpreting noun compounds via verb semantics (Kim and Baldwin, 2006). Nastase et al. (Nastase et al., 2006) used WordNet to retrieve the senses and hypernyms of components as well as the context-based word senses based on grammatical collocations over 20 SRs, grouped into 5 super-classes. Ó Séaghdha and Copestake (Ó Séaghdha and Copestake, 2007) used context and lexical/relation similarity to classify NCs, and developed an SR set intended to specifically address the unbalanced distribution of NCs across SRs in other SR sets. Ó Séaghdha later extended this work using kernel methods (Ó Séaghdha, 2009). Nulty (Nulty, 2007) investigated the effectiveness of three different learning algorithms for NC interpretation at the token level. More recently, Tratz and Hovy (Tratz and Hovy, 2010) proposed a method based on semantic similarity and patterns with a large set of Boolean features such as synonyms, hypernyms and suffix type.

In research closely related to this work, Turney and Littman (Turney and Littman, 2005) proposed a method for measuring the relational similarity between a pair of nominal phrases, for use in analogical reasoning. For example, the noun pair *cat:meow* is analogous to the pair *dog:bark*, because both represent an animal and the sound it makes. Their model is based on distributional similarity using a vector space model, and was developed based on experiments over NCs. They query the web for paraphrases of a given NC using prepositions and other connectives (e.g. *bark of dog* and *dog for bark* in the instance of *dog bark*), and classify instances based on the number of results for each of 128 paraphrase patterns, using a nearest-neighbour approach.

There has also been research on NC interpretation for languages other than English. Johnston and Busa (Johnston and Busa, 1996) used qualia structure from the Generative

Lexicon (Pustejovsky, 1995) to interpret NCs in Italian. Utsuro et al. (Utsuro et al., 2007) learned the SRs of Japanese compound functional expressions by projecting them onto dependency relations. Zhao et al. (Zhao et al., 2007) used paraphrase patterns and web statistics to interpret Chinese nominalizations, similarly to Lapata and Keller (Lapata and Keller, 2004). Sumita (Sumita and Iida, 1991) used a similar approach to translate Japanese phrases of form *N1 no N2* into English.

2.2 Disambiguating Syntactic Ambiguity: Bracketing

In addition to NC interpretation, a great deal of research has focused on the task of NC bracketing (Marcus, 1980; Lauer, 1995; Lapata and Keller, 2004; Nakov and Hearst, 2005; Vadas and Curran, 2008). NC bracketing is the task of disambiguating the internal structure of an NC made up of 3 or more component nouns. For example, the ternary (i.e. 3-ary) NC *computer science department* is most naturally interpreted as a department of computer science, corresponding to the "left bracketing" structure *((computer science) department)*. On the other hand, *linguistics graduate program* is most naturally interpreted as a graduate program of linguistics, corresponding to the "right bracketing" structure *(linguistics (graduate program))*, although the left bracketing interpretation of a program for linguistics graduates (i.e. *(linguistics graduate) program)*) is also available.

To disambiguate the syntactic structure of ternary or higher order NCs, Marcus (Marcus, 1980) proposed the "adjacency model". That is, with a given ternary NC (N1 N2 N3), the model compares the relative frequency of (N1 N2) and (N2 N3) in binary NC data. In the case that (N1 N2) is more frequent than (N2 N3), the method selects a left bracketing analysis, and in the case that (N2 N3) is the most frequent, it selects a right bracketing analysis. Note that adjacency model does not distinguish between the left or right model since the probability is based on adjacent terms. Lauer (Lauer, 1995) argued that the "dependency model" is a more accurate representation of the underlying syntax, and demonstrated that it is empirically superior to the adjacency model. For a given NC (N1 N2 N3), the dependency model compares the frequency of (N1 N2) with that of (N1 N3), based on the observation that the two dependency tuples in the left bracketing case are (N1 N2) and (N2 N3), while those in the right bracketing case are (N2 N3) and (N1 N3). Thus, in the instance that (N1 N2) is more frequent than (N1 N3), the model prefers a left bracketing analysis, and in the converse case, the model prefers a right bracketing analysis. State-of-the-art results for NC bracketing use web n-grams, combined adjacency and dependency features, and features such as whether the component words collapse into single words (Nakov and Hearst, 2005; Bergsma et al., 2010). The relevance of NC bracketing to this paper is that we explore the hypothesis that NC interpretation predictions can enhance NC bracketing.

3 Semantic Relations

Noun compound (NC) interpretation has a long history in both theoretical and computational linguistic research (Downing, 1977; Levi, 1978; Finin, 1980; Vanderwende, 1994; Barker and Szpakowicz, 1998; Rosario and Hearst, 2001; Lapata, 2002; Moldovan et al., 2004; Kim and Baldwin, 2005; Girju, 2007). Conventionally, semantic relations (SRs) are

Relation	***Definition***	***Example***
AGENT	N_2 is performed by N_1	*student protest, band concert, military assault*
BENEFICIARY	N_1 benefits from N_2	*student price, charitable compound*
CAUSE	N_1 causes N_2	*printer tray, flood water, film music, story idea*
CONTAINER	N_1 contains N_2	*exam anxiety, overdue fine*
CONTENT	N_1 is contained in N_2	*paper tray, eviction notice, oil pan*
DESTINATION	N_1 is destination of N_2	*game bus, exit route, entrance stairs*
EQUATIVE	N_1 and N_2	*composer arranger, player coach*
INSTRUMENT	N_1 is used in N_2	*electron microscope, diesel engine, laser printer*
LOCATED	N_1 is located at N_2	*building site, home town, solar system*
LOCATION	N_1 is the location of N_2	*lab printer, desert storm, internal combustion*
MATERIAL	N_2 is made of N_1	*carbon deposit, gingerbread man, water vapour*
OBJECT	N_1 is acted on by N_2	*engine repair, horse doctor*
POSSESSOR	N_1 has N_2	*student loan, company car, national debt*
PRODUCT	N_1 is a product of N_2	*automobile factory, light bulb, color printer*
PROPERTY	N_2 is N_1	*elephant seal, blue car, big house, fast computer*
PURPOSE	N_2 is meant for N_1	*concert hall, soup pot, grinding abrasive*
RESULT	N_1 is a result of N_2	*storm cloud, cold virus, death penalty*
SOURCE	N_1 is the source of N_2	*chest pain, north wind, foreign capital*
TIME	N_1 is the time of N_2	*winter semester, morning class, late supper*
TOPIC	N_2 is concerned with N_1	*computer expert, safety standard, horror novel*

Table 1. *The semantic relation set of Barker and Szpakowicz (Barker and Szpakowicz, 1998) (N_1 = modifier, N_2 = head noun)*

used to describe how the components of a given NC interact with each other. Semantic relations specify the underlying relation between a head noun and its modifier(s) in the form of a directed binary predicate. For example, *family car* involves the semantic relation POSSESSOR, which describes the ownership of *car* (head noun) by a *family* (modifier). On the other hand, *GM car* contains the relation MAKE, which represents the fact that *GM* (modifier) produces the *car* (head noun).

Various sets of semantic relations have been proposed in linguistics for interpreting NCs. Levi (Levi, 1978) defined 9 coarse-grained SRs for NCs. Finin (Finin, 1980) defined a large number of SRs, but went on to claim that the number of SRs required to interpret an arbitrary NC is unbounded. Downing (Downing, 1977) argued that the SR interpretation of an NC differs depending on pragmatic and contextual information. Vanderwende (Vanderwende, 1994) defined 13 SRs based on a combination of WH questions and syntactico-semantic roles such as SUBJECT, LOCATIVE and CAUSED-BY. Rosario and Hearst (Rosario and Hearst, 2001) focused on the medical domain and claimed that pre-

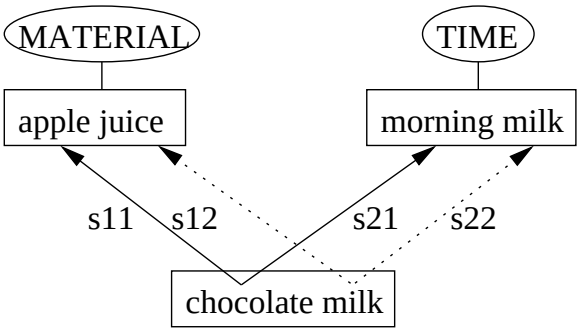

Fig. 1. Similarity between test NC *chocolate milk* and training NCs *apple juice* and *morning milk*

existing general-purpose sets of SRs did not fit the medical domain well. They went on to define a new set of 38 SRs, of which they used 18 in their experiments.

In this research, our focus is not on deriving a new set of SRs, but to compare different interpretation techniques relative to a fixed set of semantic relations. As such, we use the set of 20 SRs proposed by Barker and Szpakowicz (Barker and Szpakowicz, 1998), as it is relatively well-established in NLP research and was found to adequately capture the dataset used in this paper. Table 1 details the 20 SRs.

4 Motivation and Approach

The intuition behind this work on NC interpretation is the observation that NCs which contain similar words in corresponding positions tend to share the same semantics. To illustrate how the method works, let us consider *chocolate milk* as a test instance, which we will attempt to interpret based on the training instances *apple juice* and *morning milk*. The SR of *apple juice* is MATERIAL while that of *morning milk* is TIME. As shown in Figure 1, we compare the test instance to each of the training instances in a nearest-neighbour approach. That is, we calculate the word similarity between the nouns in corresponding positions in the test and each of the training instances, i.e. compare the modifier noun to each of the modifier nouns in the training data, and the head noun to each of the training head nouns.

Table 2 shows the component-wise similarity (modifier = N_1 and head noun = N_2, respectively) between the test instance and each of the training instances. Note that *Combined Similarity* is the weighted sum of the component word similarities. The word similarities S_{ij} are computed based on the `WUP` measure (Wu and Palmer, 1994) (detailed in Section 5.2), as implemented in WordNet::Similarity (Patwardhan et al., 2003).[1] Based on this, the combined similarity of *chocolate milk* with *apple juice* is 0.77, and that for *morn-*

[1] See Section 6.1 for empirical justification of the use of `WUP` as the basis of the word similarity calculation.

	Training noun	Test noun	S_{ij}	Combined Similarity
N_1	apple	chocolate	0.71	**0.77**
N_2	juice	milk	0.83	
N_1	morning	chocolate	0.27	0.64
N_2	milk	milk	1.00	

Table 2. *WordNet-based similarities for component nouns in the training and test data*

	Training noun	Test noun	S_{ij}	Combined Similarity
N_1	personal	loan	0.32	0.58
N_2	interest	rate	0.84	
N_1	bank	loan	0.75	0.80
N_2	interest	rate	0.84	

Table 3. *The effects of polysemy on the similarities between nouns in the training*

ing milk is 0.64 (see below for details). Since the similarity with *apple juice* is higher, the SR for *chocolate milk* is resolved to MATERIAL, the correct prediction in this case.

Unlike methods which use explicit sense information (e.g. Moldovan et al. (Moldovan et al., 2004) and Nastase et al. (Nastase et al., 2006)), and which are hence susceptible to the unavoidable noise associated with word sense disambiguation, word similarity makes no direct use of sense information. Instead, it models word similarity by the union of the senses of each word, and averages across the word pairings. As a result, our approach is not directly exposed to any sense ambiguity of the component words. That is not to say, of course, that we believe that context-sensitive word sense information (e.g. from a hypothetical high-accuracy word sense disambiguation method) would not boost the accuracy of our method, or that our simplistic assumption of a word being represented by the union of its senses is cognitively plausible. Rather, the focus of the paper is on empirically demonstrating that our simplistic method — which does not make use of context-sensitive word sense information — is remarkably adept at disambiguating the semantics of NCs.

The ability of our method to deal with the effects of word sense ambiguity can be seen in Table 3, where our training NCs are *personal interest* and *bank interest* (corresponding to SRs POSSESSOR and CAUSE/TOPIC), and our test NC is *loan rate*. Note that both training instances contain the head noun *interest*, but with different semantics: “a diversion that

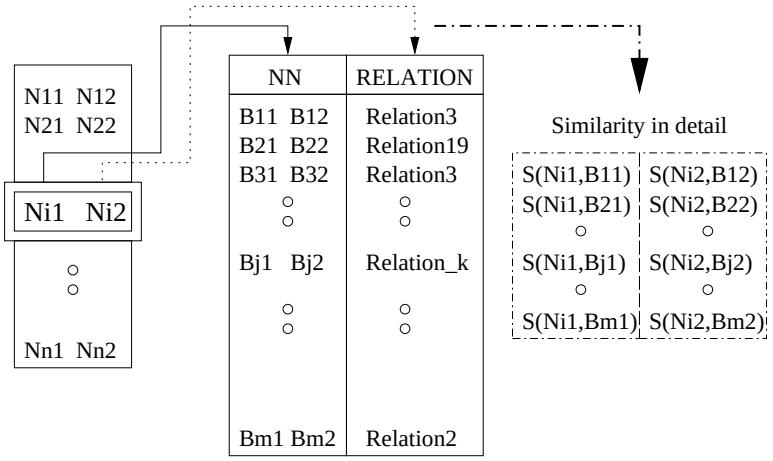

Fig. 2. Similarity between the i^{th} NC in the test data and j^{th} NC in the training data

occupies one's time and thoughts" in the case of *personal interest*, and "a fixed charge for borrowing money" in the case of *bank interest*. Despite this, our approach resolves the SR for the test instance *loan rate* correctly as CAUSE/TOPIC based on the semantic similarity between the different modifier pairings, i.e. the fact that *loan* is more similar to *bank* than *personal*.

Figure 2 shows the architecture of the constituent similarity method for interpreting NCs, where *(*N_{i1} N_{i2}*)* is a test instance and each *(*B_{j1} B_{j2}*)* is a training instance. $S(N_{i1}, B_{j1})$ and $S(N_{i2}, B_{j2})$ are the similarity between modifiers and head nouns, respectively, for *(*N_{i1} N_{i2}*)* and *(*B_{j1} B_{j2}*)*

The first step is to compute the similarities between the pair of modifiers and pair of head nouns for a given test and training instance ($S(N_{i1}, B_{j1})$ and $S(N_{i2}, B_{j2})$). The second step is to compute the combined similarity of modifiers and head nouns, as a weighted average (Equation 1).

Formally, S is the similarity between NCs $(N_{i,1}, N_{i,2})$ and $(B_{j,1}, B_{j,2})$:

$$S((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2})) = \alpha S1 + (1 - \alpha)S2 \quad (1)$$

where $S1$ is the modifier similarity ($S(N_{i,1}, B_{j1})$) and $S2$ is head noun similarity ($S(N_{i,2}, B_{j2})$); $\alpha \in [0, 1]$ is a weighting factor. Note that α will be used to test the different weight of component words with respect to the SRs in Section 6.2.

The final SR is determined by:

$$rel(N_{i,1}, N_{i,2}) \quad = \quad rel(B_{m,1}, B_{m,2}) \quad (2)$$

where:

$$m \quad = \quad \underset{j}{\operatorname{argmax}}\, S((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2})) \quad (3)$$

5 Data and Resources

5.1 Data Collection

In order to perform the experiments, we first collected 2- and 3-term NCs from the the Wall Street Journal, based on simple POS tag sequence data. We excluded NCs containing proper nouns, as on the one hand, proper nouns tend to occur with a reduced set of SRs (potentially simplifying the task), and on the other hand, the coverage of proper nouns in WordNet is poor (potentially complicating the task for our method).[2] Based on this methodology, we collected a total of 2,169 unique 2-term NC types and 1,571 unique 3-term NC types.

Two trained human annotators then individually bracketed the 3-term NCs, arrived at an agreed bracketing for each NC, and subsequently tagged all the 3-term NCs for the outermost SR. That is, the human annotators tagged each 3-term NC via the outermost 2-term NC, based on the bracketing information.[3] For example, ((N1 N2) N3) is considered as (N2 N3) while (N1 (N2 N3)) is considered as (N1 N3). The annotators were also asked to annotate all of the 2-term NCs. In both cases, annotators were allowed to annotate a given NC with multiple SRs in instances of genuine ambiguity. Any disagreements in the original annotations were resolved based on discussion between the annotators.

From the disagreement data, we observed certain SR pairings that commonly caused problems, including SOURCE and CAUSE (e.g. *debt cost*), PURPOSE and TOPIC (e.g. *customer service*), and OBJECT and TOPIC (e.g. *price reduction*).

The initial agreement over the SR annotation task was 52.31% for the 2-term NCs and 49.28% for the 3-term NCs. Note that the inter-annotator agreement over 3-term NCs is slightly lower than that over 2-term NCs. We hypothesize that the reason for this was the syntactic ambiguity of 3-term NCs: despite the NCs being provided with manual bracketing, the annotators sometimes indicated that their preferred interpretation was based on the alternate bracketing, an effect which was exacerbated by lack of context.

Finally, for evaluation, we randomly selected roughly 50% of the instances as our test data and the remainder as our training data. The breakdown of the data across the different SRs is presented in Table 4. Here, the numbers in each "N+" column represent the instances labelled with the given SR (and possibly other SRs), and the "M" columns represent the number of instances of the given SR and also other SRs. For example, in the set of 2-term NCs, 10 NCs are labeled with AGENT, among which, 1 NC is also labeled with some other SR. The number of unique test and training instances was 1,081 and 1,088 for 2-term NCs, and 785 and 786 for 3-term NCs, respectively. The final dataset is available for download from `http://www.csse.unimelb.edu.au/research/lt/resources/ncompound/ncompound.tgz`.

For direct comparison, we followed Kim and Baldwin (Kim and Baldwin, 2008) in repurposing the semantic relation dataset from SemEval-2007 (Girju et al., 2007), interpreting each noun pair as an NC and reusing the original annotation. Note that although Task 8

[2] We additionally have the problem of determining the extent of multiword named entities, e.g. *Glen Eira house*, which we would want to treat as a 2-term rather than 3-term NC.

[3] If the innermost 2-term NC were also tagged, it would be interesting to explore the interaction between the respective SRs, which we leave for future annotation and exploration.

Relation	*2-term NCs*				*3-term NCs*			
	Test		Training		Test		Training	
	N+	M	N+	M	N+	M	N+	M
AGENT	10	1	5	0	9	0	7	1
BENEFICIARY	10	1	7	1	2	0	3	0
CAUSE	54	5	74	3	21	0	18	0
CONTAINER	13	4	19	3	13	1	7	2
CONTENT	40	2	34	2	23	0	18	0
DESTINATION	1	0	2	0	0	0	1	0
EQUATIVE	9	0	17	1	1	0	2	1
INSTRUMENT	6	0	11	0	2	0	3	0
LOCATED	12	1	16	2	3	0	5	0
LOCATION	29	9	24	4	19	0	27	0
MATERIAL	12	0	14	1	10	0	11	0
OBJECT	88	6	88	5	22	6	26	3
POSSESSOR	33	1	22	1	25	4	21	6
PRODUCT	27	0	32	6	27	1	26	1
PROPERTY	76	3	85	3	33	0	43	0
PURPOSE	159	13	161	9	89	7	95	6
RESULT	7	0	8	0	3	0	4	0
SOURCE	75	11	99	15	61	0	44	1
TIME	25	1	19	0	19	0	24	0
TOPIC	465	24	447	39	438	16	437	15
TOTAL	1163	82	1184	96	820	35	822	36

Table 4. *The distribution of semantic relations in 2-term and 3-term noun compounds in our dataset (N+ = the number of instances labelled with the given SR [and possibly other SRs]; M = the number of instances of the given SR and also other SRs)*

at SemEval-2010 (Hendrickx et al., 2009) used a larger dataset for the same task, the dataset does not provide word senses for NCs (as required by the competitor methods detailed below). Thus, we chose to experiment only with the dataset from SemEval-2007. The original SemEval-2007 task was, for a given test instance, to judge whether a pre-determined SR applies or not. In addition to this style of binary classification, we generated a variant of the original dataset where all positive training instances were pooled together into a single classification task across all 7 SRs, made up of 446 training and 254 test instances, as detailed in Table 5.

We additionally reimplemented the methods of Moldovan et al. (Moldovan et al., 2004)

SR	Binary classification		7-way SR labelling	
	Test	Training	Test	Training
CAUSE-EFFECT	80	136	36	71
INSTRUMENT-AGENCY	78	135	36	68
PRODUCT-PRODUCER	93	126	55	78
ORIGIN-ENTITY	81	136	35	52
THEME-TOOL	71	129	27	50
PART-WHOLE	72	138	28	64
CONTENT-CONTAINER	74	137	37	63
Total	549	937	254	446

Table 5. *Breakdown of the SemEval-2007 data across SRs, in terms of the original binary classification and the repurposed SR label dataset used in this research*

and Nastase et al. (Nastase et al., 2006). The Moldovan et al. method is called semantic scattering, and is based on statistical affinity between particular word senses and relations. The Nastase et al. method performs NC interpretation based on word sense and hypernym information in a memory-based learner.[4]

5.2 WordNet::Similarity

To compute the semantic similarity between two nouns, we used the WordNet::Similarity open-source package to compute word similarities (Patwardhan et al., 2003).[5] WordNet::Similarity was developed at the University of Minnesota, and provides various methods to measure the similarity or relatedness between a pair of concepts or word senses. It contains implementations of a variety of lexical comparison methods, split across three basic types: similarity, relatedness and random. The similarity methods are categorized into two groups: path-based (LCH (Leacock and Chodorow, 1998) and WUP (Wu and Palmer, 1994)) and information-content based (RES (Resnik, 1995), JCN (Jiang and Conrath, 1998), and LIN (Lin, 1998)).

Path-based methods compute the shortest path length between senses of two nouns in WordNet. LCH calculates the shortest path between two target concepts (c_1 and c_2) in the

[4] They also found "grammatical collocations" to enhance the effectiveness of their method, which we ignore in this work, as we perform the interpretation task over NCs out of context, making no use of token-level data.

[5] `http://www.d.umn.edu/~tpederse/similarity.html`

WordNet a *is-a* hierarchy, and computes the similarity as follows:

$$Similarity_{lch}(c_1, c_2) = -\log\left(\frac{p}{2 \times depth}\right) \tag{4}$$

where $depth$ is the maximum depth of the hierarchy in WordNet and p is the number of nodes in the shortest path between the two concepts.

On the other hand, WUP uses the path length to the root node from the least common subsumer (LCS) of the two concepts, as follows:

$$Similarity_{wup}(c_1, c_2) = \frac{2 \times p3}{p1 + p2 + 2 \times p3} \tag{5}$$

where $p1$ and $p2$ are the number of nodes on the path from the LCS to c_1 to c_2 respectively, and $p3$ is the number of nodes on the path between LCS and the root.

RES, JCN and LIN augment the calculation of path length with the information content (IC) of the LCS, calculated as follows:

$$IC(c) = -\log\frac{freq(c)}{freq(root)} \tag{6}$$

where $freq(c)$ is the frequency of a given concept c, and $freq(root)$ is the frequency of the root of the hierarchy.

RES calculates the similarity of two concepts by the information of their LCS:

$$similarity_{res} = IC(lcs(c_1, c_2)) \tag{7}$$

JCN is an extension of RES, where the path length between the two concepts is included in the calculation, based on:

$$similarity_{jcn} = IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2)) \tag{8}$$

LIN is a further variant of RES, based on the Dice coefficient:

$$Similarity_{lin}(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{9}$$

WordNet::Similarity also includes implementations of a number of "relatedness" measures which make use of relations such as *has-part,is-made-of,is-an-attribute-of*, in addition to *is-a* relations. There are three relatedness measures: HSO, LESK and VECTOR. LESK is based on the weighted word overlap of different pairings of synset glosses, over a variety of relation types.

VECTOR is a corpus-based measure. Each word is represented as a multi-dimensional vector of co-occurring words. The similarity of a word pair is measured by the cosine similarity of the two vectors. In Equation 10, $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ are the vectors of the two target words:

$$Relation_{vector}(c_1, c_2) = \frac{\overrightarrow{v_1} \times \overrightarrow{v_2}}{||\overrightarrow{v_1}|| \times ||\overrightarrow{v_2}||} \tag{10}$$

Finally, as a baseline approach, we use the RANDOM method, which randomly assigns a score for each noun pair (with all occurrences of the same noun pair receiving the same score).

Basis	Method
Path-based similarity:	LCH, WUP
Information content-based similarity:	JCN, LIN
Relatedness:	LESK, VECTOR
Random:	RANDOM

Table 6. *WordNet::Similarity methods used to score noun pairs*

Table 6 lists the different similarity and relatedeness measures used in this work.

For the purposes of evaluation, we used the WUP and LCH path-based similarity measures, the JCN and LIN information content-based similarity measures, LESK and RANDOM.

6 Evaluation

We evaluate our proposed method under various experimental settings. The first experiment (Section 6.1) is based on the basic version of our method (1-NN, with uniform weighting of the modifier and head noun, and only the provided training instances), as applied to our own dataset and also the SemEval-2007 data. In the second experiment, we experiment with different weightings of the two components of the NC in our interpretation methodology (Section 6.2). In the third experiment (Section 6.3), we apply the method to 3-term NC interpretation. Finally, we use the NC interpretation method to perform NC bracketing in Section 6.4.

6.1 Experiment I: Automatic NC Interpretation

We first present results for our 2-term NC dataset of 1,088 training and 1,081 test instances, as detailed in Section 5. The baseline for this experiment is majority class classification, that is assigning the TOPIC SR to all test instances.

Table 7 shows that the WUP method achieves the highest NC interpretation accuracy, well above the baseline and marginally above the inter-annotator agreement (a nominal upper bound for the task). Among the four measures of similarity used in this first experiment, the path-based similarity measures have higher accuracy than the information content-based methods. We also include results for two lexical relatedness methods for completeness, both of which are found to be markedly less accurate than all of the similarity-based methods. Based on these results, we use the WUP method exclusively for the remainder of the paper, except where noted.

The methods of Moldovan et al. (Moldovan et al., 2004) and Nastase et al. (Nastase et al., 2006) both require sense data, which is not included in our dataset. We thus move on

Method		Accuracy
Human annotation	Inter-annotator agreement	52.3%
Majority class	Baseline	43.0%
Path-based	WUP	**53.3%**
	LCH	52.9%
Information content-based	JCN	46.7%
	LIN	47.4%
Relatedness	LESK	42.4%
Random	RANDOM	21.8%

Table 7. *Accuracy of NC interpretation for the different WordNet-based scoring methods over our 2-term NC dataset*

Class	Our method (WUP)	Moldovan et al.	Nastase et al.
7-way	52.8%	49.6%	55.7%
2-way	65.0%	63.4%	66.7%

Table 8. *NC interpretation accuracy over SemEval-2007*

to experiment with the SemEval-2007 data, including comparison with these competitor methods, as detailed in Table 8 for 7- and 2-way classification tasks.

Our system surpassed the method of Moldovan et al. over both task settings, but was inferior to the method of Nastase et al.. However, we note that both benchmark systems require manual word sense information where we do not. Given this stringent requirement, the relative empirical superiority of the method of Nastase et al. is modest.

Lastly, we tested the impact on test data accuracy of differing amounts of training data. As with any supervised method, the amount of training data is expected to impact on the accuracy of our method. Figure 3 shows the learning curve for our method as we increase the amount of training data. As the curve shows, our method clearly benefits from more training data, and is showing no sign of plateauing, suggesting that access to extra training data should lead to appreciable gains in accuracy.

We additionally experimented with a k nearest neighbour (k-NN) approach for varying values of k, but found the results to be slightly lower than those for the original 1-NN method, except for the less competitive relatedness-based measures of VECTOR and LESK, where the overall results improved very slightly. The results are omitted from the paper because ultimately they don't shed any extra light on the task of NC interpretation.

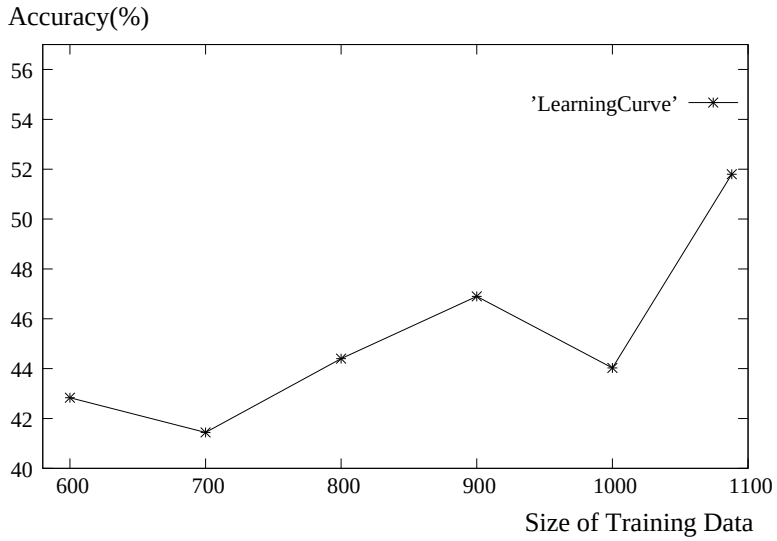

Fig. 3. Learning Curve with respect to the size of the training data

Relative contribution of modifier/head noun	Relation	Example
modifier $<$ head noun	PROPERTY	*elephant seal*
modifier $=$ head noun	EQUATIVE	*composer arranger*
modifier $>$ head noun	TIME	*morning class*

Table 9. *Predicted contribution of head noun and modifier for different semantic relations*

6.2 Experiment II: Relative Contribution of Head noun and Modifier

We next move on to investigate the relative contribution of the head noun and modifier in semantic interpretation. We expect that, for different SRs, the head noun and modifier will potentially have different impact in determining the overall similarity of the NC. For example, the head noun intuitively is the greater determinant of similarity with the PROPERTY SR (e.g. *fairy penguin*). Conversely, the modifier appears to be the primary semantic contributor for the TIME SR (e.g. *winter coat*). Table 9 shows a selection of SRs where we predict the head noun and modifier make different relative contributions to the overall interpretation.

Based on these intuitions, in our second experiment, we explore the relative contribution of the NC components to NC interpretation, by testing the relative impact of the α weight in Equation 1 to overall performance. Figure 4 shows the overall accuracy for SR interpretation as we modify the value of α. We tested values of α in the range $[0.0, 1.0]$ in

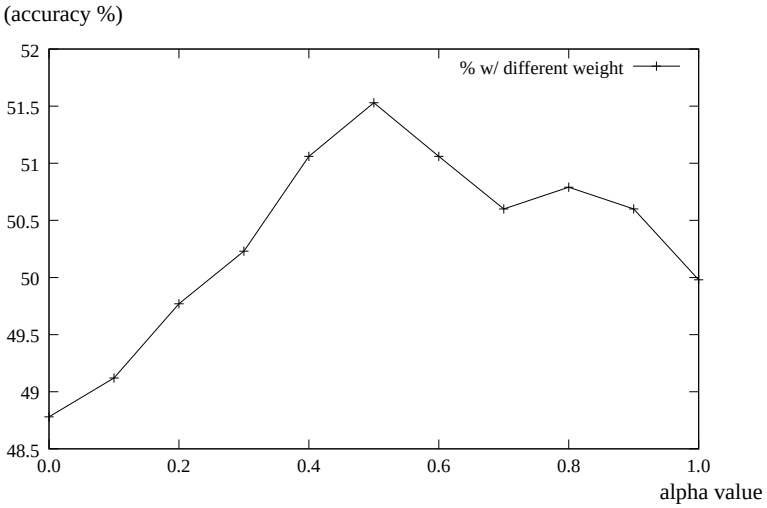

Fig. 4. Classifier accuracy at different α values

increments of 0.1. For example, when the weight of head noun is 0.3, that of modifier is 0.7. The best overall performance is achieved for $\alpha = 0.5$, i.e. when the head noun and modifier contribute equally to the determination of the SR.

Figure 5 shows the accuracy for different SRs at different α values ($\alpha \in 0.2, 0.5, 0.8$). First, we noticed that some SRs have a strong correlation between the target SR and either the modifier or the head. For example, some SRs such as CAUSE and POSSESSOR are more heavily influenced by the modifier, while SRs such as CONTENT and PROPERTY are affected more by the head noun, as predicted. Contrary to our expectations, the head noun seems to be a stronger determinant of the SR than the modifier for EQUATIVE NCs, and both nouns seem to play an equal role for TIME NCs. The results show that despite localized biases for individual SRs, the overall performance is the best when the components have equal say in the prediction of SR. Based on this finding, we use $\alpha = 0.5$ exclusively for the remainder of this paper.

6.3 Experiment III: Interpretation of 3-term NCs

In this experiment, we test our approach over 3-term NCs. To the best of our knowledge, these are the first reported results for NC interpretation over 3-term NCs.

The experiment is divided into two parts. In the first sub-experiment, we test our method using all three components of the NCs, and multiply all three component similarities together. Note that despite using all three components, our goal is to interpret only the top-level dependency tuple, not both 2-term NCs that make up the 3-term NC.

In the second sub-experiment, we use 2-term NCs extracted from the 3-term NCs based on gold-standard bracketing information. That is, (N2 N3) is extracted from ((N1 N2) N3),

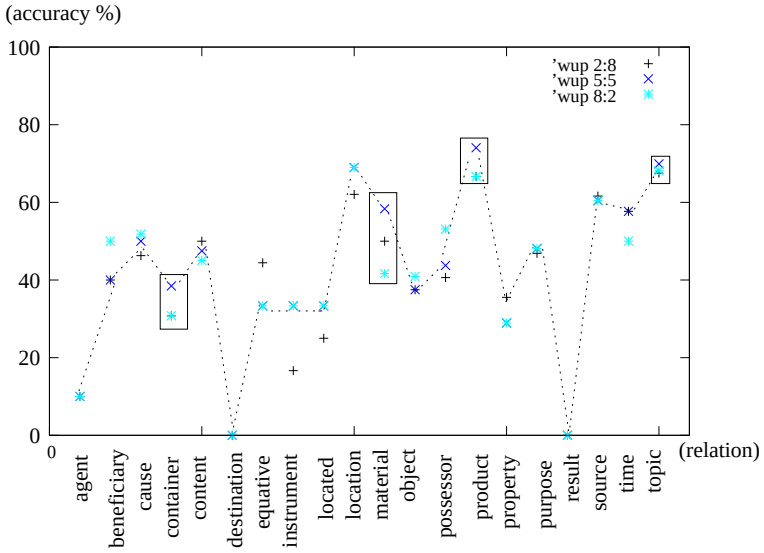

Fig. 5. Classification accuracy for each semantic relation at different α values ("$*$" = higher weight on the modifier; "$\times$" = equal weighting on the modifier and head noun; "$+$" = higher weight on the head noun)

Method		Accuracy
Human annotation	Inter-annotator agreement	49.3%
Majority class	Baseline	55.8%
Path-based	WUP	48.3%
	LCH	54.7%
Information content-based	JCN	55.8%
	LIN	**57.7**%
Relatedness	LESK	52.0%
Random	RANDOM	31.7%

Table 10. *Accuracy of NC interpretation for the different WordNet-based similarity measures over 3-term NCs (full)*

and (N1 N3) is extracted from (N1 (N2 N3)), and it is only these two components that are used by our method.

Tables 10 and 11 show the accuracy of our method over 3-term NCs, using all three components in the first case, and only a single 2-term NC extracted from the 3-term NC in the second case. Comparing the two NC representations, we find that our method performs

Method		Accuracy
Human annotation	Inter-annotator agreement	49.3%
Majority class	Baseline	55.8%
Path-based	WUP	43.18%
	LCH	**44.8%**
Information content-based	JCN	40.9%
	LIN	41.2%
Relatedness	LESK	42.7%
Random	RANDOM	26.6%

Table 11. *Accuracy of NC interpretation for the different WordNet-based similarity measures over 3-term NCs (binarised)*

better when it has access to all three components of the NC. This suggests that all the NC components contribute to determine the SR to some degree. Interestingly, whereas the path-based methods were the strongest performers in the 2-term NC case, with 3-term NCs, information content-based methods perform the best. While error analysis did not unearth a clear reason for this reverse trend, we hypothesise that the differing level of information contained in 2- and 3-term NCs is captured differently by the two families of semantic similarity measure. Overall, the results for the 3-term NCs were lower than those for the 2-term NCs, but equally, the inter-annotator agreement was slightly lower, and the results for all lexical similarity and relatedness methods are actually close to or above the level of agreement.

6.4 Experiment IV: NC Bracketing using Semantic Relations

Existing methods for bracketing (Nakov and Hearst, 2005; Bergsma et al., 2010) rely on a large amount of data to compute lexical probabilities. Unfortunately, however, a large amount of data is not always available, e.g. for specialist domains. Here, we test the utility of SRs in bracketing 3-term NCs, based on the dataset described in Section 5. The basic intuition behind the method is that the SR prediction for the outermost 2-term NC from the correct bracketing should agree with the SR prediction for the 3-term NC (without bracketing). For example, given the NC *automobile production target*, we would expect the SR for *production target* to be the same as for the overall 3-term NC, whereas we would not have the same expectation for *automobile target* (derived from the alternative bracketing hypothesis).

We use the WUP similarity measure to interpret NCs since it produced the highest accuracy, and test bracketing using various similarity thresholds. For example, if the SR for the outermost 2-term NC (based on the dependency model) agrees with for the overall 3-

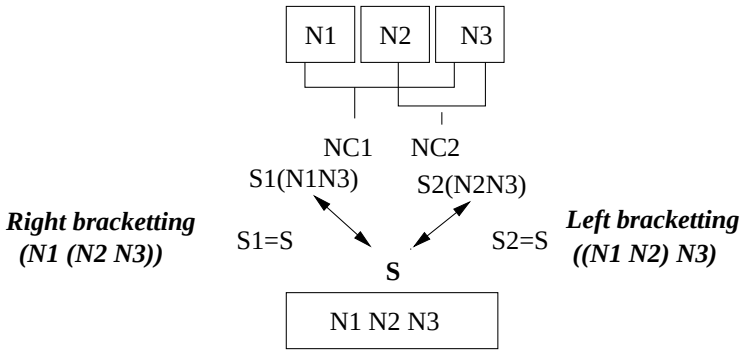

Fig. 6. Overview of the proposed NC bracketing method

term NC, both calculated at a similarity value greater than or equal to 0.5, then we use the SR predictions for bracketing. Otherwise, we ignore the SR predictions and fall back on a simple lexical model.

To disambiguate the syntactic structure in 3-term or higher term NCs, we use lexical probabilities based on a limited amount of data collected from three corpora: the British National Corpus, the Wall Street Journal and the Brown Corpus. SRs are calculated for the original 3-term NC and its 2-term NCs, based on a dependency model analysis of the components (Lauer, 1995).

Figure 6 depicts our method for identifying the syntactic structure of a 3-term NC using SRs and the 2-term NCs extracted from the 3-term NC. In Figure 6, we extract the two outermost 2-term NCs *(N1 N3)* and *(N2 N3)* from the 3-term NC *(N1 N2 N3)*, corresponding to a right bracketing and left bracketing analysis, respectively. We then classify the SR for each of *(N1 N3)*, *(N2 N3)* and *(N1 N2 N3)* (in the latter case, based on the three component model without bracketing information), and if the SR for only one of the two 2-term NCs agrees with that for the overall 3-term NC, that gives us our bracketing prediction. For example, *physics winter school*, with SR TOPIC, is associated with the two candidate 2-term NCs of *physics school* (i.e. (N1 N3)) and *winter school* (i.e. (N2 N3)). Since the SR of the first 2-term NC (TOPIC) is the same as the 3-term NC, while that for the second NC (TIME) is different, we can disambiguate the 3-term NC as *(physics (winter school))* (right bracketing).

Analysis of our method showed that it had limited coverage, due to its reliance on WordNet to calculate lexical similarity. We hence decided to combine our proposed bracketing method with a lexical probabilistic model. That is, we first attempt to determine the bracketing using the probabilistic model, and if the model does not provide an answer due to data sparseness (i.e. if we do not have any instances of both word pairs), then we back off to our SR-based model. Note that we also tested using the lexical model then applying the probabilistic model, but found the results to be uncompetitive and omit them from this paper. As our probabilistic model, we used adjacency and dependency from Lauer

Method	Model	Lexical only		Combined	
		Coverage	Accuracy	Coverage	Accuracy
Lauer	Adjacency	87.1%	60.2%	93.1%	64.2%
	Left Dependency	64.9%	56.0%	81.3%	**65.8%**
	Right Dependency	63.1%	23.5%	80.1%	34.2%
Nakov & Hearst	Adjacency	99.7%	71.2%	100.0%	71.2%
	Left Dependency	99.8%	74.0%	99.9%	**74.1%**
	Right Dependency	99.8%	40.9%	99.9%	40.9%

Table 12. *Complementing existing bracketing methods with semantic relation-based disambiguation*

(Lauer, 1995) (described in Section 2.2), using the combination of our three corpora — the Brown Corpus, the Wall Street Journal and the British National Corpus — to compute the lexical probabilities. Also, to compare the method with the current state-of-the-art, i.e. Nakov and Hearst (Nakov and Hearst, 2005), we used Google web data to compute the probabilities. Table 12 shows the coverage (i.e. proportion of instances where we are able to make a prediction) and accuracy of the probabilistic model in isolation, and also the combined model.

Bracketing using only the probabilistic models also suffers from data sparseness, and does not achieve 100% coverage even with web data. When we combine the probabilistic model with our semantic relation method, both the coverage and the accuracy increase in all cases. The best performance for each of the corpus- and web-based methods were achieved with the left-dependency model. Over web data, the increment in coverage and accuracy is tiny due to the majority of instances being disambiguated by the probabilistic model. However, even here, by combining it with the semantic relation method, we are able to increase accuracy marginally.

From Table 12, we can clearly see that when SR-based bracketing resolves the syntactic ambiguity over instances which are untagged by the probabilistic method, the accuracy is higher than for the probabilistic method in isolation, and also higher than for the SR-based bracketing method over all NCs. This suggests that the SR-based bracketing method complements the lexical probability method over our dataset, for both the Lauer and Nakov & Hearst methods. This result is highly significant, in showcasing an application where the underlying SR interpretation method boosts coverage and accuracy, if only modestly for the web-based method of Nakov & Hearst. Given the relative impact of NC bracketing on overall parser accuracy (Vadas and Curran, 2008), this finding potentially suggests a novel research direction in improving parsing with lexical semantic features, in the vein of Agirre et al. (Agirre et al., 2008) and others.

7 Conclusion

In this paper, we have proposed an automatic method for interpreting the semantics of NCs with semantic relations. Our method uses an instance-based learning approach based on lexical similarity between each a test instance and the corresponding components of training NCs, using WordNet. We applied the method to both 2-term and 3-term NCs, using the semantic relation set of Barker and Szpakowicz (Barker and Szpakowicz, 1998). Overall, we found that our method did better over 2-term NCs, at a level nearing inter-annotator agreement over the task. Finally, we proposed an NC bracketing method based on our NC interpretations, and demonstrated that we were able to boost the accuracy and coverage of standard NC bracketing methods based on lexical probabilities.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Agirre, E., Baldwin, T., and Martinez, D. (2008). Improving parsing and PP attachment performance with sense information. In *Proc. of the 46th Annual Meeting of the ACL: HLT (ACL 2008)*, pages 317–325, Columbus, USA.

Baldwin, T. and Kim, S. N. (2009). Multiword expressions. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing*. CRC Press, Boca Raton, USA, 2nd edition.

Barker, K. and Szpakowicz, S. (1998). Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th international conference on Computational linguistics*, pages 96–102.

Bergsma, S., Pitler, E., and Lin, D. (2010). Creating robust supervised classifiers via web-scale n-gram data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 865–874, Uppsala, Sweden.

Butnariu, C. and Veale, T. (2008). A concept-centered approach to noun-compound interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 81–88, Manchester, UK.

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

Fan, J., Barker, K., and Porter, B. W. (2003). The knowledge required to interpret noun compounds. In *Seventh International Joint Conference on Artificial Intelligence*, pages 1483–1485.

Finin, T. W. (1980). *The semantic interpretation of compound nominals*. PhD thesis, University of Illinois, Urbana, Illinois, USA.

Girju, R. (2007). Improving the interpretation of noun phrases with cross-linguistic information. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 568–575, Prague, Czech Republic.

Girju, R., Moldovan, D., Tatu, M., and Antohe, D. (2005). On the semantics of noun compounds. *Computer Speech and Language*, 19(4):479–496.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2007). Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proc. of the 4th International Workshop on Semantic Evaluations*, pages 13–18, Prague, Czech Republic.

Grover, C., Lapata, M., and Lascarides, A. (2004). A comparison of parsing technologies for the biomedical domain. *Journal of Natural Language Engineering*, 1(1):1–38.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2009). Semeval-2010 task 8: Multi-way classification of semantic

relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 94–99, Boulder, USA.

Hull, R. D. and Gomez, F. (1996). Semantic interpretation of nominalizations. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-1996)*, pages 1062–1068, Portland, Oregon.

Isabelle, P. (1984). Another look at nominal compounds. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING-1984)*, pages 509–516, San Francisco, USA.

Jiang, J. and Conrath, D. (1998). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, pages 19–33.

Johnston, M. and Busa, F. (1996). Qualia structure and the compositional interpretation of compounds. In *Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, pages 77–88, Santa Cruz, USA.

Kim, S. and Baldwin, T. (2006). Interpreting semantic relations in noun compounds via verb semantics. In *Proc. of COLING/ACL 2006*, pages 491–498, Sydney, Australia.

Kim, S. and Baldwin, T. (2008). Benchmarking noun compound interpretation. In *Proc. of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, pages 569–576, Hyderabad, India.

Kim, S. N. and Baldwin, T. (2005). Automatic interpretation of noun compounds using WordNet similarity. In *Second International Joint Conference On Natural Language Processing*, pages 945–956, JeJu, Korea.

Lapata, M. (2002). The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

Lapata, M. and Keller, F. (2004). The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the Human Langauge Techinology Conference and Conference on Empirical Methods in National Language Processing (HLT/NAACL-2004)*, pages 121–128, Boston, USA.

Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Noun Compounds*. PhD thesis, Macquarie University.

Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

Levi, J. N. (1978). *The Syntax and Semantics of Complex Nominals*. Academic Press, New York, USA.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*.

Marcus, M. (1980). *A Theory of Syntactic Recognition for Natural Language*. MIT Press, Cambridge, USA.

Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., and Girju, R. (2004). Models for the semantic classification of noun phrases. In *Proceedings of HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 60–67, Boston, USA.

Nakov, P. (2008). Noun compound interpretation using paraphrasing verbs: Feasibility study. In *Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA'08)*, pages 103–117, Varna, Bulgaria.

Nakov, P. and Hearst, M. (2005). Search engine statistics beyond the n-gram: Application to noun compound bracketting. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 17–24, Ann Arbor, USA.

Nakov, P. and Hearst, M. (2006). Using verbs to characterize noun-noun relations. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA'06)*, pages 233–244, Varna, Bulgaria.

Nastase, V., Sayyad-Shirabad, J., Sokolova, M., and Szpakowicz, S. (2006). Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 781–787, Boston, USA.

Nicholson, J. and Baldwin, T. (2005). Statistical interpretation of compound nominalisations. In *Proceedings of the Australian Language Technology Workshop*, pages 152–159, Sydney, Australia.

Nulty, P. (2007). Semantic classification of noun phrases using web counts and learning algorithms. In *Proceedings of the Association of Computational Linguistics 2007 Student Research Workshop*, pages 79–84, Prague, Czech Republic.

Ó Séaghdha, D. (2009). Semantic classification with WordNet kernels. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 237–240, Boulder, USA.

Ó Séaghdha, D. and Copestake, A. (2007). Co-occurrence contexts for noun compound interpretation. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, pages 57–64, Prague, Czech Republic.

Patwardhan, S., Banerjee, S., and Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*, pages 17–21, Mexico City, Mexico.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT press, Cambridge, USA.

Resnik, P. (1995). Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the 3rd Workship on Very Large Corpus*, pages 77–98.

Rosario, B. and Hearst, M. (2001). Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, USA.

Spärck Jones, K. (1983). *Compound noun interpretation problems*. Prentice-Hall, Englewood Cliffes, USA.

Sumita, E. and Iida, H. (1991). Experiments and prospects of example-based machine translation. In *Proc. of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 185–192, Berkeley, USA.

Tratz, S. and Hovy, E. (2010). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Uppsala, Sweden.

Turney, P. D. and Littman, M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278.

Utsuro, T., Shime, T., Tsuchiya, M., Matsuyoshi, S., and Sato, S. (2007). Learning dependency relations of Japanese compound functional expressions. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, pages 65–72, Prague, Czech Republic.

Vadas, D. and Curran, J. R. (2008). Parsing noun phrase structure with CCG. In *Proceedings of ACL-08: HLT*, pages 335–343, Columbus, USA.

Vanderwende, L. (1994). Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th conference on Computational linguistics*, pages 782–788.

Venkatapathy, S. and Joshi, A. (2006). Using information about multi-word expressions for the word-alignment task. In *Proc. of the COLING/ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 53–60, Sydney, Australia.

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, USA.

Zhao, J., Liu, H., and Lu, R. (2007). Semantic labeling of compound nominalization in Chinese. In *Proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions*, pages 73–80, Prague, Czech Republic.