Natural Language Engineering **19** (3): 385–407. © Cambridge University Press 2013
doi:10.1017/S1351324913000107

A lexical semantic approach to interpreting and bracketing English noun compounds

SU NAM KIM[1] and TIMOTHY BALDWIN[2]

[1]*Faculty of Information Technology, Monash University, Victoria, Australia*
e-mail: sunamkim@gmail.com
[2]*NICTA Victoria Research Laboratories, Department of Computing and Information Systems, The University of Melbourne, Victoria, Australia*
e-mail: tb@ldwin.net

(*Received 4 October 2011; revised 29 March 2013; accepted 30 March 2013*)

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 two- and three-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** 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 SRs has a long history in Natural Language Processing 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 Natural Language Processing (Butnariu and Veale 2008; Nakov 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 basic approaches: analogy-based interpretation (Rosario and Hearst

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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; Butnariu and Veale 2008; Nakov 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 two-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 three-term NC dataset developed for this research, in what we believe are the first published results over three-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 SRs 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 SRs, 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 (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 analysing 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 handwritten rules, making it labor-intensive and hard to repurpose to new domains or relation sets. 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

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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, Lapata and Lascarides 2004), that is NCs where the head noun is a deverbal noun such as *animation* (derived from *animate*). 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.* (2004) and 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 (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.* (2004) performed compound nominalization interpretation over biomedical data. Nakov and Hearst (2006) also focused on the biomedical domain, using verb semantics based on a web corpus.

Moldovan *et al.* (2004) and Girju *et al.* (2005) proposed methods for open-domain NC interpretation using the pairing of word senses of the component words. Later 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 (2005) used a method based on the nearest-neighbour classification over the union of senses of the modifier and head noun that forms the basis of this paper. In a separate work, Kim and Baldwin (2006) also proposed a method for interpreting NCs via verb semantics. 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 five superclasses. 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 (2009) later extended this work using kernel methods. Nulty (2007) investigated the effectiveness of three different learning algorithms for NC interpretation at the token level. More recently, 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 (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.

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

There has also been research on NC interpretation for languages other than English. Johnston and Busa (1996) used qualia structure from the Generative Lexicon (Pustejovsky 1995) to interpret NCs in Italian. Utsuro *et al.* (2007) learned the SRs of Japanese compound functional expressions by projecting them onto dependency relations. Zhao *et al.* (2007) used paraphrase patterns and web statistics to interpret Chinese nominalizations similar to Lapata and Keller (2004). 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 three 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 (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 where (N2 N3) is most frequent, it selects a right bracketing analysis. Note that adjacency model does not distinguish between the left and the right model since the probability is based on adjacent terms. 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, Pitler and Lin, 2010). The relevance of NC bracketing to this paper is that we explore the hypothesis that NC interpretation predictions can enhance NC bracketing.

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

3 Semantic relations

Noun compound 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, SRs are used to describe how the components of a given NC interact with each other. SRs 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 SRs have been proposed in linguistics for interpreting NCs. Levi (1978) defined nine coarse-grained SRs for NCs. 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 (1977) argued that the SR interpretation of an NC differs depending on pragmatic and contextual information. Vanderwende (1994) defined thirteen SRs based on a combination of WH-questions and syntactico–semantic roles such as SUBJECT, LOCATIVE and CAUSED-BY. Rosario and Hearst (2001) focused on the medical domain and claimed that pre-existing general-purpose sets of SRs did not fit the medical domain well. They went on to define a new set of thirty-eight SRs, of which they used eighteen 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 SRs. As such, we use the set of twenty SRs proposed by Barker and Szpakowicz (1998), as it is relatively well established in Natural Language Processing research and was found to adequately capture the dataset used in this paper. Table 1 details the twenty 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 with 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,

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

Table 1. *The semantic relation set of Barker and Szpakowicz (1998) (N1 = modifier, N2 = head noun)*

Relation	Definition	Example
AGENT	N2 is performed by N1	*student protest, band concert, military assault*
BENEFICIARY	N1 benefits from N2	*student price, charitable compound*
CAUSE	N1 causes N2	*printer tray, flood water, film music, story idea*
CONTAINER	N1 contains N2	*exam anxiety, overdue fine*
CONTENT	N1 is contained in N2	*paper tray, eviction notice, oil pan*
DESTINATION	N1 is destination of N2	*game bus, exit route, entrance stairs*
EQUATIVE	N1 and N2	*composer arranger, player coach*
INSTRUMENT	N1 is used in N2	*electron microscope, diesel engine, laser printer*
LOCATED	N1 is located at N2	*building site, home town, solar system*
LOCATION	N1 is the location of N2	*lab printer, desert storm, internal combustion*
MATERIAL	N2 is made of N1	*carbon deposit, gingerbread man, water vapour*
OBJECT	N1 is acted on by N2	*engine repair, horse doctor*
POSSESSOR	N1 has N2	*student loan, company car, national debt*
PRODUCT	N1 is a product of N2	*automobile factory, light bulb, color printer*
PROPERTY	N2 is N1	*elephant seal, blue car, big house, fast computer*
PURPOSE	N2 is meant for N1	*concert hall, soup pot, grinding abrasive*
RESULT	N1 is a result of N2	*storm cloud, cold virus, death penalty*
SOURCE	N1 is the source of N2	*chest pain, north wind, foreign capital*
TIME	N1 is the time of N2	*winter semester, morning class, late supper*
TOPIC	N2 is concerned with N1	*computer expert, safety standard, horror novel*

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

	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	

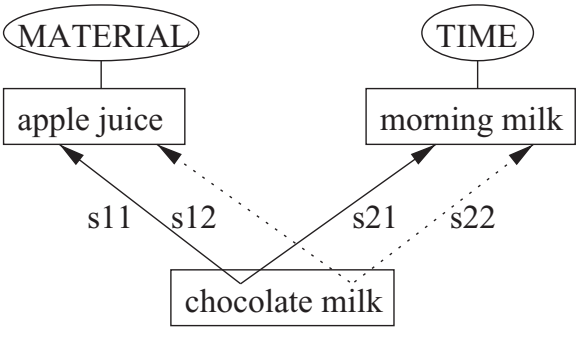


Fig. 1. Similarity between test NC *chocolate milk* and training NCs *apple juice* and *morning milk*.

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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

	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	

Banerjee and Pedersen 1994).[1] Based on this, the combined similarity of *chocolate milk* with *apple juice* is 0.77, and that for *morning 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.* 2004; 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 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 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

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

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

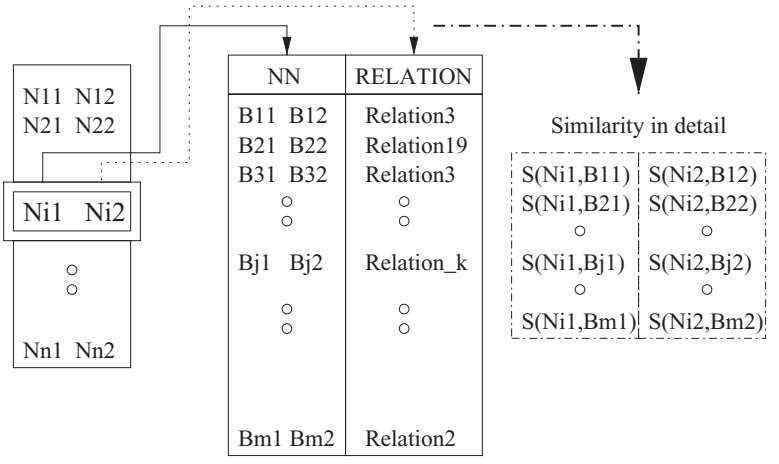


Fig. 2. Similarity between the ith NC in the test data and jth NC in the training data.

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 ((1)).

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

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

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 weights of component words with respect to the SRs in Section 6.2.

The final SR is determined by:

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

where:

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

5 Data and resources

5.1 Data collection

In order to perform the experiments, we first collected two- and three-term NCs from *The Wall Street Journal*, based on simple part-of-speech (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

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

complicating the task for our method).[2] Based on this methodology, we collected a total of 2,169 unique two-term NC types and 1,571 unique three-term NC types.

Two trained human annotators then individually bracketed the three-term NCs, arrived at an agreed bracketing for each NC and subsequently tagged all the three-term NCs for the outermost SR. That is, the human annotators tagged each three-term NC via the outermost two-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 two-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 percent for the two-term NCs and 49.28 percent for the three-term NCs. Note that the inter-annotator agreement over three-term NCs is slightly lower than that over two-term NCs. We hypothesize that the reason for this was the syntactic ambiguity of three-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 percent of the instances as our test data and the remainder as our training data. The breakdown of the data across 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 two-term NCs, ten NCs are labelled with AGENT, among which one NC is also labelled with some other SR. The number of unique test and training instances was 1,081 and 1,088 for two-term NCs, and 785 and 786 for three-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 (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 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

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

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

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

Table 4. *Distribution of semantic relations in two- and three-term NCs 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)*

	Two-term NCs				Three-term NCs			
	Test		Training		Test		Training	
Relation	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	1,163	82	1,184	96	820	35	822	36

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 seven SRs, made up of 446 training and 254 test instances as detailed in Table 5.

We, in addition, reimplemented the methods of Moldovan *et al.* (2004) and Nastase *et al.* (2006). The Moldovan *et al.* (2004) method is called semantic scattering and is based on statistical affinity between particular word senses and relations. The Nastase *et al.* (2006) method performs NC interpretation based on word sense and hypernym information in a memory-based learner.[4]

[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.

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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*

	Binary classification		Seven-way SR labelling	
SR	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

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 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:

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

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

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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 (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).

Table 6 lists 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

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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

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

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

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

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 three-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 two-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.* (2004) and Nastase *et al.* (2006) both require sense data, which is not included in our dataset. We thus move on to experiment

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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

Class	Our method (WUP)	Moldovan *et al.* (2004)	Nastase *et al.* (2006)
Seven-way	52.8%	49.6%	55.7%
Two-way	65.0%	63.4%	66.7%

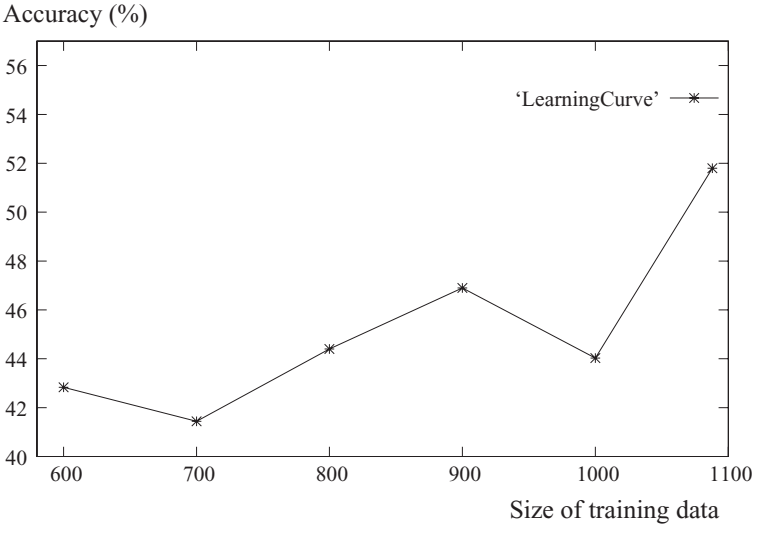


Fig. 3. Learning curve with respect to the size of the training data.

with the SemEval-2007 data, including comparison with these competitor methods, as detailed in Table 8 for seven- and two-way classification tasks.

Our system surpassed the method of Moldovan *et al.* (2004) over both task settings, but was inferior to the method of Nastase *et al.* (2006). However, we note that both benchmark systems require manual word sense information, whereas we do not require this. Given this stringent requirement, the relative empirical superiority of the method of Nastase *et al.* (2006) 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 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, in addition, 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

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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

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*

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 do not shed any extra light on the task of NC interpretation.

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 that 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 NC components to NC interpretation by testing the relative impact of α weight in (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 increments of 0.1. For example, when the weight of the 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 noted 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.

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

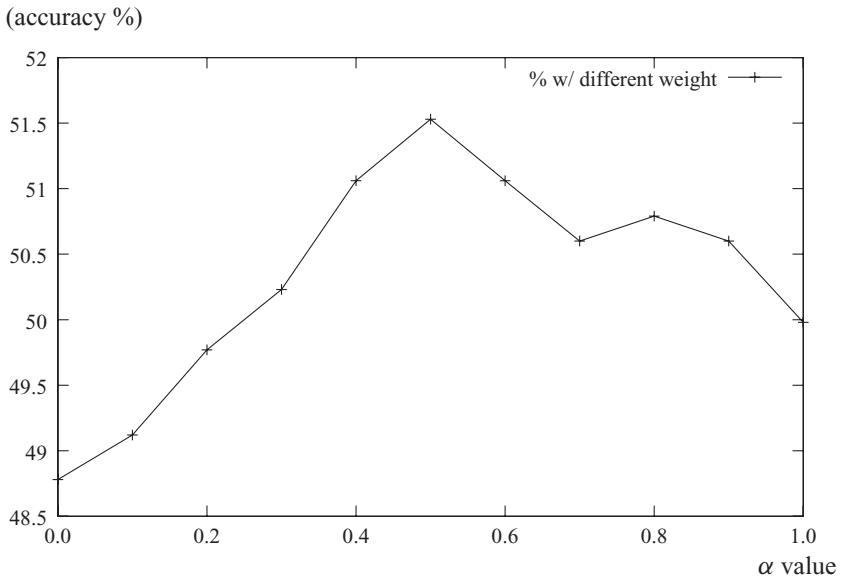


Fig. 4. Classifier accuracy at different α values.

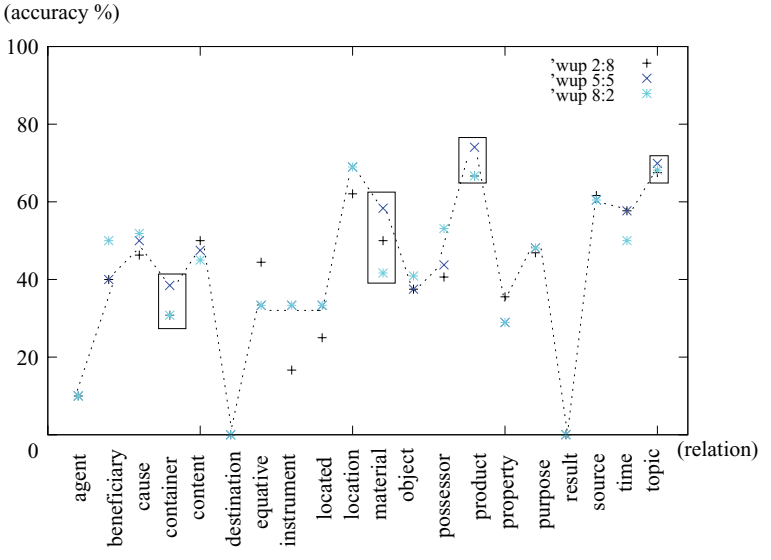


Fig. 5. (Colour online) Classification accuracy for each semantic relation at different α values ('*' = higher weight on the modifier; '×' = equal weighting on the modifier and head noun; '+' = higher weight on the head noun).

6.3 Experiment III: interpretation of three-term NCs

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

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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

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

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

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 two-term NCs that make up the three-term NC.

In the second sub-experiment, we use two-term NCs extracted from three-term NCs based on gold-standard bracketing information, that is (N2 N3) is extracted from ((N1 N2) N3), 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 three-term NCs using all three components in the first case, and only a single two-term NC extracted from the three-term NC in the second case. Comparing the two NC representations we find that our method performs 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 two-term NC case, with three-term NCs, information content-based methods perform the best. While error analysis did not unearth a clear

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

reason for this reverse trend, we hypothesize that the differing level of information contained in two- and three-term NCs is captured differently by the two families of semantic similarity measure. Overall, the results for the three-term NCs were lower than those for the two-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: noun compound 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 three-term NCs based on the dataset described in Section 5. The basic intuition behind the method is that the SR prediction for the outermost two-term NC from the correct bracketing should agree with the SR prediction for the three-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 three-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 two-term NC (based on the dependency model) agrees with for the overall three-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 three-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 three-term NC and its two-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 three-term NC using SRs and the two-term NCs extracted from the three-term NC. In Figure 6 we extract the two outermost two-term NCs (*N1 N3*) and (*N2 N3*) from the three-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 two-term NCs agrees with that for the overall three-term NC, then this gives us our bracketing prediction. For example, *physics winter school*, with SR TOPIC, is associated with the two-candidate two-term NCs of *physics school* (i.e. (N1 N3)) and *winter school* (i.e. (N2 N3)). Since the SR of the first two-term NC (TOPIC) is the same as the three-term NC, while that for the second NC (TIME) is different, we can disambiguate the three-term NC as *(physics (winter school))* (right bracketing).

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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

Method	Model	Lexical only		Combined	
		Coverage	Accuracy	Coverage	Accuracy
Lauer (1995)	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 and Hearst (2005)	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%

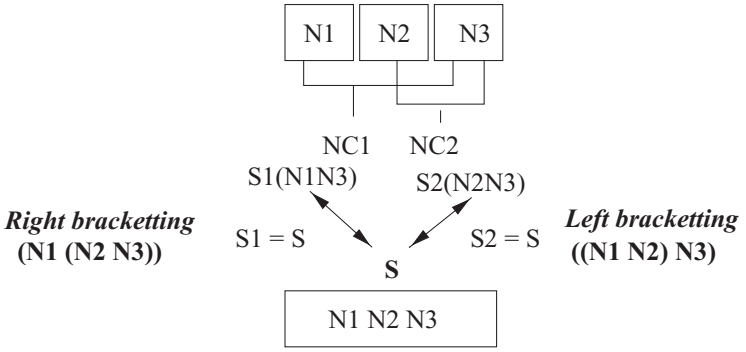


Fig. 6. Overview of the proposed NC bracketing method.

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 than 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 (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 lexical probabilities. Also, to compare the method with the current state-of-the-art, i.e. 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 percent coverage even with web data. When we combine

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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 was 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 (1995) and Nakov and Hearst (2005) 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 and Hearst (2005). 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, Baldwin and Martinez (2008) and others.

7 Conclusion

In this paper we have proposed an automatic method for interpreting the semantics of NCs with SRs. Our method uses an instance-based learning approach based on lexical similarity between each of the test instance and the corresponding components of training NCs using WordNet. We applied the method to both two- and three-term NCs using the SR set of Barker and Szpakowicz (1998). Overall, we found that our method did better over two-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.

Acknowledgments

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 *Proceedings of the 46th Annual Meeting of the ACL: HLT (ACL 2008)*, Columbus, OH, pp. 317–25.

Baldwin, T., and Kim, S. N. 2009. Multiword expressions. In N. Indurkhya, and F. J. Damerau (eds.), *Handbook of Natural Language Processing*, 2nd ed, pp. 267–292. Boca Raton, USA: CRC Press.

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

Barker, K., and Szpakowicz, S. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th International Conference on Computational Linguistics*, Montreal, Canada, pp. 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*, Uppsala, Sweden, pp. 865–74.

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)*, Manchester, UK, pp. 81–8.

Downing, P. 1977. On the creation and use of English compound nouns. *Language* **53**(4): 810–42.

Fan, J., Barker, K., and Porter, B. W. 2003. The knowledge required to interpret noun compounds. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, pp. 1483–5.

Finin, T. W. 1980. *The Semantic Interpretation of Compound Nominals*. PhD thesis, University of Illinois, Urbana, IL.

Girju, R. 2007. Improving the interpretation of noun phrases with cross-linguistic information. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, pp. 568–75.

Girju, R., Moldovan, D., Tatu, M., and Antohe, D. 2005. On the semantics of noun compounds. *Computer Speech and Language* **19**(4): 479–96.

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 *Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic, pp. 13–8.

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)*, Boulder, CO, pp. 94–9.

Hull, R. D., and Gomez, F. 1996. Semantic interpretation of nominalizations. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-1996)*, Portland, OR, pp. 1062–8.

Isabelle, P. 1984. Another look at nominal compounds. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING-1984)*, San Francisco, CA, pp. 509–16.

Jiang, J., and Conrath, D. 1998. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, pp. 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*, Santa Cruz, CA, pp. 77–88.

Kim, S. N., and Baldwin, T. 2005. Automatic interpretation of noun compounds using WordNet similarity. In *Proceedings of the Second International Joint Conference On Natural Language Processing*, JeJu, Korea, pp. 945–56.

Kim, S., and Baldwin, T. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of COLING/ACL 2006*, Sydney, Australia, pp. 491–8.

Kim, S., and Baldwin, T. 2008. Benchmarking noun compound interpretation. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, Hyderabad, India, pp. 569–76.

Lapata, M. 2002. The disambiguation of nominalizations. *Computational Linguistics* **28**(3): 357–88.

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

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 Technology Conference and Conference on Empirical Methods in National Language Processing (HLT/NAACL-2004)*, Boston, MA, pp. 121–8.

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

Leacock, C., and Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*, pp. 265–284. Cambridge, MA: MIT Press.

Levi, J. N. 1978. *The Syntax and Semantics of Complex Nominals*. New York, NY: Academic Press.

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

Marcus, M. 1980. *A Theory of Syntactic Recognition for Natural Language*. Cambridge, MA: MIT Press.

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*, Boston, MA, pp. 60–67.

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)*, Varna, Bulgaria, pp. 103–17.

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)*, Ann Arbor, MI, pp. 17–24.

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)*, Varna, Bulgaria, pp. 233–44.

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)*, Boston, MA, pp. 781–7.

Nicholson, J., and Baldwin, T. 2005. Statistical interpretation of compound nominalisations. In *Proceedings of the Australian Language Technology Workshop*, Sydney, Australia, pp. 152–9.

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*, Prague, Czech Republic, pp. 79–84.

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*, Boulder, USA, pp. 237–40.

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*, Prague, Czech Republic, pp. 57–64.

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)*, Mexico City, Mexico, pp. 17–21.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

Resnik, P. 1995. Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the 3rd Workshop on Very Large Corpus*, MIT, Cambridge, MA, pp. 77–98.

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

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press

Spärck Jones, K. 1983. *Compound Noun Interpretation Problems*. Englewood Cliffs, NJ: Prentice-Hall.

Sumita, E., and Iida, H. 1991. Experiments and prospects of example-based machine translation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, pp. 185–92.

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*, Uppsala, Sweden, pp. 678–87.

Turney, P. D., and Littman, M. L. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning* **60**(1–3): 251–78.

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*, Prague, Czech Republic, pp. 65–72.

Vadas, D., and Curran, J. R. 2008. Parsing noun phrase structure with CCG. In *Proceedings of ACL-08: HLT*, Columbus, OH, pp. 335–43.

Vanderwende, L. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th Conference on Computational Linguistics*, Kyoto, Japan, August 5–9, pp. 782–88.

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

Wu, Z., and Palmer, M. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, pp. 133–8.

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*, Prague, Czech Republic, pp. 73–80.

https://doi.org/10.1017/S1351324913000107 Published online by Cambridge University Press