Word Sense and Semantic Relations in Noun Compounds

Su Nam Kim, Monash University
Timothy Baldwin, The University of Melbourne

In this paper, we investigate word sense distributions in noun compounds (NCs). Our primary goal is to disambiguate the word sense of component words in NCs, based on investigation of "semantic collocation" between them. We use sense collocation and lexical substitution to build supervised and unsupervised word sense disambiguation (WSD) classifiers, and show our unsupervised learner to be superior to a benchmark WSD system. Further, we develop a word sense-based approach to interpreting the semantic relations in NCs.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Text Analysis; Language Parsing and Understanding

General Terms: Artificial Intelligence

Additional Key Words and Phrases: Noun Compounds, Word Sense Disambiguation, Semantic Relations

ACM Reference Format:
Kim, S.N., Baldwin, T. ACM Trans. Embedd. Comput. Syst. 9, 4, Article 39 (March 2010), 17 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

This paper investigates the interaction between word sense and the semantic interpretation of English noun compounds ("NCs"), that is sequences of nouns which syntactically function as a noun, such as *paper submission* or *computer science department*. Throughout this paper, we will refer to the component words in an NC as "components".

The observation underlying this research is that, while the component words of NCs are generally polysemous in isolation, in the context of a given NC, the sense is almost always uniquely defined across all token occurrences of that NC. That is, while word sense is generally highly dependent on both local and global context [Ng and Lee 1996; Navigli 2009], in the case of NCs, sense determination occurs almost exclusively within the confines of the NC. To see this, consider the polysemous noun *plant* which has the following nominal senses in WORDNET 3.0: (1) "buildings for carrying on industrial labor"; (2) "(botany) a living organism lacking the power of locomotion"; and (3) "something planted secretly for discovery by another".[1] Looking across a range of NCs extracted from the Penn Treebank [Marcus et al. 1993], we find that in most cases there is a very strong sense preference, independent of context: *treatment plant* (sense 1); *coffee plant* (sense 2); *plant closing* (sense 1); and *plant sciences* (sense 2). In some of these cases, it is certainly possible to construct an interpretation using one of the other senses (e.g. *treatment plant* = "a vegetative plant that treats/purifies some substance, such as water"). However, as a broad generalisation, the higher the relative frequency of the NC, the stronger the preference for a particular interpretation of the NC

[1]There is also a fourth sense in WORDNET ("an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience") which we didn't find attested in the NCs in the Penn Treebank and we exclude from our discussion.

Author's addresses: S.N. Kim, School of Information Technology, Monash University and T. Baldwin, Computing and Information Systems Department, The University of Melbourne VIC, Australia.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

Kim, Su Nam and Timothy Baldwin (to appear) Word Sense Disambiguation and Noun Compounds, ACM Transactions on Speech and Language Processing.

(i.e. the stronger the lexicalisation) and for a particular sense of each of the component nouns. The corollary of this, of course, is that the sense restriction of components in lower-frequency NCs will tend to be weaker. For example, we were able to find web occurrences of *Moscow plant* covering all three senses of *plant*:

(1) Renault's Moscow plant was inaugurated in April 2005 with initial capacity of 60,000 vehicles per year and a 230 million euro investment.[2]
(2) However, S. myrsinifolia native to the Kola Peninsula differ from Moscow plants in the opposite way: their leaves are more pubescent and more green beneath.[3]
(3) Moscow tries to discredit Minstrel by feeding London a skillful line on how Minstrel is really a Moscow plant.[4]

Even here, however, the combination of words may lead to a strong preference for a particular interpretation/sense combination, as seen with *bomber plane production plant* (sense 1), *mixed-rainforest plant* (sense 2) and *quasi-Mossad plant* (sense 3).[5]

The focus of this paper is twofold: (1) evaluation of the strength of lexical preference within NCs, and the use of sense preferences to automatically sense tag component nouns in NCs; and (2) investigation of the interface between the word sense of component nouns and the semantic interpretation of the containing NC. First, we develop a novel method for disambiguating the word sense of component nouns, and we then apply the predicted senses in an semantic relation (SR) interpretation task.

2. RELATED WORK

2.1. Word Sense Disambiguation

Word sense disambiguation is the task of resolving the sense of word instances, usually relative to a predefined sense inventory [Agirre and Edmonds 2006; Navigli 2009]. It has been recognised as one of the hardest tasks in NLP (referred to as an *AI-complete* problem). WSD has been suggested as an intermediate task for NLP applications, although at current performance levels, most attempts to incorporate WSD in actual applications have been unsuccessful. For example, Sanderson [1996] identified the potential for WSD to enhance the performance of information retrieval (IR), but in practice found that it was impossible to achieve the required level of accuracy to achieve that potential gain. Vickrey et al. [2005] and Carpuat and Wu [2005], on the other hand, got mixed results for WSD in the context of machine translation (MT), while Agirre et al. [2008], Agirre et al. [2011] and MacKinlay et al. [2012] achieved modest gains in applying WSD to a parsing task.

WSD research is categorised into two primary categories: corpus-based and knowledge-based [Ide and Veronis 1998]. Corpus-based methods use features based on neighbouring (content) words in a fixed word window of the target word, while knowledge-based methods extract features from lexical resources such as dictionaries. Yarowsky [1995] famously developed a corpus-based WSD method based on bootstrapping based on collocations, while McCarthy et al. [2004] proposed a method for learning the first sense of a word based on grammatical context, content words in a word window, and ontological semantics. Leacock and Chodorow [1998] identified monosemous *hypernym*s and *hyponym*s of a given target word, and from these acquired sense-annotated examples automatically, in a knowledge-based technique. Banerjee and Pedersen [2003] showed the usefulness of *hypernym*s in WSD based on dictionary definition overlap. Mihalcea and Moldovan [1999] and Agirre and Martinez [2000] used lexical substitution to perform WSD.

[2] `http://www.renault.com/SiteCollectionDocuments/Communiqu%C3%A9%20de%20presse/en-EN/Pieces%20jointes/21865_20100301_CPAvtoframos-GB_45B25F99.pdf`
[3] `http://www.salicicola.com/announcements/skv/pages64-73.pdf`
[4] `http://www.scribd.com/doc/7163637/The-Deceiver`
[5] All of which had a single hit, or multiple hits for the same document, on the Google search engine on 15/6/2012.

While WSD has been the target of hundreds of research papers over the last decade, there has been almost no work done on WSD of MWEs, or the interaction between WSD and MWEs. Analysis of WSD specifically over MWEs is needed to develop a robust and accurate WSD method. A particular motivation of this claim is that word sense has been shown to be useful in interpreting NCs [Moldovan et al. 2004; Kim and Baldwin 2005].

2.2. Noun Compound Interpretation

Two main approaches have been taken to automatically interpret semantic relations in NCs: (1) semantic similarity, and (2) the use of verb semantics or paraphrases that are associated with the different semantic relations.

The semantic similarity approach relies on the assumption that when the components are similar between two NCs, they tend to share the same semantic relation. For example, given that *apple juice* has the semantic relation MATERIAL (i.e. "juice which is made from apples"), it is unsurprising that the highly similar unseen NC *orange juice* should also interpreted as MATERIAL. Various approaches have been proposed to compute the semantic similarity between two NCs. For example, Moldovan et al. [2004] used word senses directly (see Section 4.2 for details) while Kim and Baldwin [2005] and Nastase et al. [2006] employed them indirectly relative to a lexical resource. More recently, Tratz and Hovy [2010] employed various features including syntactic and semantic information such as synonyms and hypernyms to compute the semantic similarity.

The paraphrase approach models the relationship between the nouns directly using verbs or paraphrases [Vanderwende 1994; Lapata 2002; Kim and Baldwin 2006b; Nakov and Hearst 2006; 2008; Butnariu and Veale 2008]. For example, *malaria mosquito* can be interpreted as "mosquito that <u>causes</u> malaria" using the verb, *cause*. Similarly, *olive oil* can be interpreted as "oil that <u>is extracted from</u> olive", indicating that the semantic relation is SOURCE. The first work in this direction was done by Vanderwende [1994], employing verb information automatically extracted from definitions in an online dictionary. In Nakov and Hearst [2008], the authors used distributional semantic information for verbs, prepositions and coordinating conjunctions acquired from a web search engine.

3. RESOURCES

In this section, we summarise the resources we use in this research. We also present the set of semantic relations (SRs) we use to interpret NCs, since our approach is centred around SRs.

3.1. Lexical Resources

WORDNET

WORDNET [Fellbaum 1998][6] is a large-scale lexical database of English developed at Princeton University under the direction of George A. Miller. It groups English words (nouns, verbs, adjectives and adverbs) into sets of synonyms called synsets. WORDNET provides short, general definitions for each synset, and records various conceptual-semantic and lexical relations between pairings of synsets. It contains both simplex words and multiword expressions. The total of all unique noun, verb, adjective, and adverb lexical items is 155,327, contained in 117,597 unique synsets (based on version 2.1). Many lexical items have a unique synset classification within a given syntactic category, but are described under more than one syntactic category.

WORDNET has been used in various natural language processing tasks such as word sense disambiguation [McCarthy et al. 2004; Moldovan et al. 2004; Nastase et al. 2006], PP-attachment [Kim and Baldwin 2006a; Atterer and Schütze 2007; Agirre et al. 2008] and question answering [Prager and Chu-Carroll 2001; Hermjakob et al. 2002], and has become a mainstream language resource in NLP. The current version is WORDNET 3.1, although most of our experiments were carried out using WORDNET 2.1 as it was the current version at the time we ran the experiments

[6] `http://wordnet.princeton.edu/`

CORELEX

CORELEX [Buitelaar 1989][7] is a noun classification based on a unified approach to systematic polysemy and the semantic underspecification of nouns, and is derived from WORDNET 1.5. While WORDNET 1.5 provides around 60,000 different noun synsets, CORELEX collapses these into a concise set of 126 coarse-grained classes, by taking into account systematic polysemy and underspecification.

CORELEX contains 45 basic CORELEX types, systematic polysemous classes and 39,937 tagged nouns. The semantic types are underspecified representations based on the Generative Lexicon [Pustejovsky 1995]. From a seed collection of hand-tagged nouns, a probabilistic tagger was built to classify unknown nouns (not in CORELEX) and to identify context-specific and new interpretations. The classification algorithm is centred around the computation of a similarity based on the Jaccard coefficient, that compares lexical items in terms of their shared attributes (linguistic patterns acquired from domain-specific corpora). In this paper, we used this lexical resource to interpret semantic relations in NCs.

As an illustration of the contents of CORELEX, the following is the class `fod` and its subclasses:

— **fod**
  — **atr fod(attribute)** *chocolate, vintage, wine*
  — **fod(food)** *ale, beefsteak, chili*
  — **fod frm(form)** *doughnut, suds*
  — **fod nat(natural_object)** *berry, java, milk*
  — **fod qui(quantify_indefinite)** *cocktail, syllabub, toast*
  — **fod sta(state)** *blackheart, pickle, stew*
  — **fod sub(substance)** *nectar, paste, wafer*

3.2. Tools

TIMBL

TIMBL [Daelemans et al. 2004][8] is a memory-based learner. Memory-based learning is based on the classical k-nearest neighbour (i.e. k-NN) approach to classification. For efficiency reasons, TIMBL makes extensive use of indexes rather than a typical flat file found in traditional k-NN systems. TIMBL includes a variety of similarity metrics such as overlap and dot product, feature weighting metrics such as information gain and chi square, and distance weighting metrics such as inverse distance and inverse linear distance. Also, it handles user-defined example weighting. For all our experiments, we use version 5.1 of TIMBL.

SENSELEARNER

SENSELEARNER [Mihalcea and Faruque 2004] is a minimally-supervised WSD system that learns a semantic language model based on SemCor [Landes et al. 1998] and information from WORDNET for those words which do not appear in the training data. The system is based on three stages of processing. First, the system preprocesses the input by tokenising, POS-tagging and carrying out named entity detection and collocation extraction using a sliding window approach. Second, the system builds a semantic language model for each POS using SemCor. Finally, it uses syntactic dependencies and a conceptual network in order to generalise over words which do not occur in the training data.

3.3. Semantic Relations

In this section, we present the set of semantic relations (SRs) used in this work.

[7] `http://www.cs.brandeis.edu/~paulb/CoreLex/corelex.html`
[8] `http://ilk.uvt.nl/software.html`

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	*exam anxiety, overdue fine*
CONTAINER	n1 contains n2	*printer tray, flood water*
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*
PRODUCT	n1 is a product of n2	*automobile factory, light bulb, colour 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*

Fig. 1. The semantic relations (SRs) used in this research (n1 = modifier, n2 = head noun)

Prior research has sought to identify the SRs in NCs from either a linguistic perspective [Levi 1979; Finin 1980; Sparck Jones 1983; Downing 1977] or a computational perspective [Vanderwende 1994; Barker and Szpakowicz 1998; Rosario and Marti 2001; Moldovan et al. 2004].

In pioneering work, Levi [1979] proposed nine SRs for non-opaque (i.e. compositional) compounds. Finin [1980] countered Levi's earlier work in claiming that the number of discrete SRs needed to interpret NCs was infinite, citing the impact of various pragmatic factors on the semantics of NCs. Sparck Jones [1983] built on earlier work by Downing [1977] in claiming that the SRs in NCs can be described only in terms of tendencies or preferences, and not absolutes.

Different researchers have proposed varying sets of SRs for interpreting NCs, based on a variety of approaches to computational interpretation. Vanderwende [1994] defined SRs based on *WH*-questions. Barker and Szpakowicz [1998] developed 20 SRs in a bottom-up fashion over task-oriented data, and Moldovan et al. [2004] proposed 32 SRs for use in open-domain paraphrase-based interpretation. Rosario and Marti [2001] identified 36 domain-specific SRs for the biomedical domain. Nastase et al. [2006] defined 30 SRs, along with 5 superclasses, in an attempt to overcome the problems of fine-granularity and unbalanced distribution. Ó Séaghdha [2007] designed a set of SRs with careful annotation guidelines, with the intention of achieving greater class balance and higher inter-annotator agreement, based on the set of SRs proposed by Levi [1979].

The quest to define a commonly agreed-upon set of SRs remains unsolved due to a number of problems: (1) the granularity of SRs, (2) the coverage of SRs over data from different domains, and (3) the class distribution of SRs. Smaller sets of SRs tend to be hard to work with due to their coarse granularity [Levi 1979; Vanderwende 1994]. Larger sets of SRs tend to fit the data better but are associated with ambiguity and have skewed class distribution [Finin 1980; Rosario and Marti 2001; Moldovan et al. 2004]. Also, pragmatic effects lead to disagreements in SR labelling in- and out-of-context [Downing 1977; Sparck Jones 1983; Copestake and Lascarides 1997].

In this study, we used the set of SRs proposed by Barker and Szpakowicz [1998] instead of defining our own set. We found this set to be reasonably well-balanced and clearly defined, and to have been used broadly in previous work. Figure 1 details the 20 semantic relations defined by Barker and Szpakowicz [1998].

Sense	*art* Modifier	*art* Head noun	*art* Overall	*day* Modifier	*day* Head noun	*day* Overall
ws_1	**.85**	**.62**	**.67**	.13	.04	**.41**
ws_2	.11	.04	.22	.02	.04	.20
ws_3	.00	.03	.08	**.80**	.00	.12
ws_4	.04	.31	.03	.00	**.91**	.20
ws_5	—	—	—	.04	.01	.05
ws_6	—	—	—	.00	.00	.03
ws_7	—	—	—	.00	.00	.00
ws_8	—	—	—	.01	.00	.00
ws_9	—	—	—	.00	.00	.00
ws_{10}	—	—	—	.01	.00	.00

Fig. 2. Sense distribution for *art* and *day* as an NC modifier, head noun and overall in SEMCOR

4. MOTIVATION FOR DISAMBIGUATING NOUNS IN NCS

The primary observation underlying this research is that word sense is often highly restricted within NCs. A secondary observation, however, is that the sense distribution of component nouns in NCs doesn't necessarily correspond to that in simplex contexts, and that conventional WSD methods tend to perform poorly over NCs. Additionally, by focusing on the elements of the NC, we can bring the "one sense per collocation" heuristic [Yarowsky 1993] into play, in assuming that the elements of a given NC will always occur with the same sense irrespective of context, just as we explored the hypothesis above that the SR for a given combination of senses in an NC is fixed.

We compare this approach with a standard corpus-based approach, based on the off-the-shelf WSD system SENSELEARNER. We show that our disambiguation method based solely on word sense combinatorics is more successful at disambiguating word sense than existing methods. Note that we do not dispute the claim that context influences NC interpretation (e.g. Girju et al. [2007]). Rather, for our current purposes we focus exclusively on word sense at the type-level for NCs out of context and leave the harder task of token-level interpretation/WSD for future research.

4.1. Sense Distribution of Polysemous Elements in Noun Compounds

Our motivating observation for the task is that the word sense distribution of NC constituents is both different to that for simplex usages, and varies across head noun and modifier usages (which we will refer to as "grammatical roles"). For example, *art* as a modifier has a different sense distribution to that as a head noun. Figure 2 describes the sense distribution for the words *art* and *day* for each grammatical role within an NC, and also across all usages (NC or otherwise). The word senses and sense glosses are based on WORDNET 2.1, and the sense distributions are taken from SEMCOR. According to WORDNET 2.1, *art* has a total of 4 senses and *day* has 10 senses.

As a modifier in NCs, *day* occurs most frequently with ws_3 ("daytime, daylight"), while as a head noun, it is used mostly with ws_4 ("a day assigned to a particular purpose or observance"). On the other hand, *art* is mostly used as ws_1 ("the products of human creativity; works of art collectively") when used as both head noun and modifier. However, the distribution of senses of *art* as a head noun in NCs is more evenly distributed compared to its senses as a modifier in NCs. Our research would suggest that these sense distribution disparities are representative of what we observe across a range of nouns, and that the sense distribution of nouns is biased when they occur in NCs, with sense distributions varying according to syntactic role in the NC.

4.2. Sense Restrictions in Noun Compounds

As discussed above, Moldovan et al. [2004] used sense collocation to interpret NCs, based on the hypothesis that when the sense collocation of two NCs is the same, their SR is most likely also the same. Moldovan et al. [2004] encoded this hypothesis in the following formulation, based on conditional probability:

$$sr^* = \underset{sr_i}{\operatorname{argmax}} P(sr_i|ws(n_1), ws(n_2)) \quad (1)$$

where $ws(n_*)$ is the word sense of noun n_*, n_1 is the modifier, n_2 is the head noun and each sr_i is an SR.

Based on the above, we formulate our probabilistic model to disambiguate the word sense of polysemous element in NCs as:

$$ws^*(n_i) = \underset{ws(n_i)}{\operatorname{argmax}} P(ws(n_i)|ws(n_j), sr) \quad (2)$$

where n_i is the target noun to disambiguate, n_j is the other noun in the NC (which we are assuming has already been disambiguated), and sr is the SR between the modifier and head noun. Note that we assume we know the sense of the non-target noun n_j (either the head noun or the modifier) as well as the SR in this formulation, and use this to determine the word sense of the target noun n_i (either the modifier or the head noun, respectively). As it is unlikely that we will have reliable access to the sr for a given NC, we modify Equation (2) by replacing sr with the grammatical role (either modifier or head noun), to encode the observation from above that the sense distribution of a noun can vary greatly across the two roles. Hence, our final formulation is Equation (2) with sr replaced by the grammatical role (as in Equation (3)). We further experiment with the inclusion of sr to attest the contribution of sr to WSD.

4.3. One Sense per Collocation

The "one sense per collocation" heuristic was proposed by Yarowsky [1993] as a general bootstrapping method for WSD. It assumes that a word will be used with the same sense within a given word collocation — such as an NC or adjective-noun collocations — across all token occurrences. Yarowsky [1993] claims that this heuristic is effective at disambiguating words which occur in collocational contexts, and demonstrated that the accuracy of the heuristic over a range of binary disambiguation bootstrapping tasks was between 90% and 99%. He also showed that the heuristic was more successful in certain contexts than others. In the case of nouns, the best disambiguating context was directly adjacent adjectives or nouns, underlying the effectiveness of the heuristic for our work.

We draw on the one sense per collocation heuristic to disambiguate constituents in NCs. However, in our case, the heuristic is applied slightly differently to the original in Yarowsky [1995], in that we are seeking to disambiguate both nouns in NCs rather than one element based on what words it co-occurs with. We also apply it to the full WORDNET sense inventory rather than coarse-grained binary distinctions. Hence, we do not expect as high accuracy as reported by Yarowsky [1995]. We also do not make any linguistic claims about the potential for a given NC to have different senses based on context. Our basic claim is that the majority of token occurrences of a given NC will conform to a given sense combination.

5. OUR METHODS

In our experimentation on WSD of components in NCs, we use two classifiers, one supervised and one unsupervised.

5.1. Supervised Method

The first classifier is supervised and uses the sense collocation method of Moldovan et al. [2004], modified to model the grammatical roles of target nouns as described above, in the form:

$$ws^*(n_i) = \underset{ws(n_i)}{\operatorname{argmax}} P(ws(n_i)|ws(n_j), grammatical\_role) \quad (3)$$

Here, the semantics of n_j, which is assumed to be sense-determined, is extracted from two sources. We experiment with the use of two types of semantics: noun classes from CORELEX, and the first

sense	substituted NCs
1	**craft/artifact** museum
2	**artistic production/creative activity** museum
3	**artistry/superior skill** museum
4	**artwork/graphics/visual communication** museum

Fig. 3. Example of the substitution method for each word sense of the polysemous *art* in *art museum*

sense from WORDNET. We also test the method under various conditions, as presented below. First, we use only the predetermined semantics of the non-target noun as a feature but underspecify the grammatical role; second, we use both the semantics of the non-target noun and the SR; and third, we use all of the semantics of the non-target noun, the SR and the grammatical role. The first variant (proposed originally in Kim and Baldwin [2007]) is intended to test the contribution of the grammatical role relative to the original method; the second variant (novel to this work) is intended to test the importance of the SR in disambiguating the NC; and the third variant (also novel to this work) combines these two into a single model.

$$ws^*(n_i) = \underset{ws(n_i)}{\text{argmax}}\, P(ws(n_i)|ws(n_j), role) \tag{4}$$

$$ws^*(n_i) = \underset{ws(n_i)}{\text{argmax}}\, P(ws(n_i)|ws(n_j), sr) \tag{5}$$

$$ws^*(n_i) = \underset{ws(n_i)}{\text{argmax}}\, P(ws(n_i)|ws(n_j), role, sr) \tag{6}$$

5.2. Unsupervised method

We also built an unsupervised classifier based on lexical substitution, similarly to Mihalcea and Moldovan [1999] and Agirre and Martinez [2000]. We replace the target noun with its *synonym*s from WORDNET synsets, then compute the probability of each underlying word sense by calculating the frequency of the substituted NCs in a corpus. We used a web corpus (via the Google search engine) since it provides a large amount of data to compute the probabilities, notwithstanding noise in the results [Lapata and Keller 2004]. Note that as each word sense can have more than one *synonym*, we normalise the frequency across all *synonym*s to compute the final probability. Finally, we assign the word sense which has highest substitution-based frequency to the target noun.

Equation (7) shows how to compute the probability when the target noun (n_1) is the modifier and the non-target noun (n_2) is the head noun:

$$ws^*(n_1) = \underset{s_i \in ws(n_1)}{\text{argmax}} \frac{\sum_{n_j \in ss(s_i) \setminus \{s_i\}} freq(n_j, n_2)}{|ss(s_i) \setminus \{s_i\}|} \tag{7}$$

where each s_i is a word sense of n_1, and $ss(s_i)$ returns the synset containing sense s_i. The calculation in the case that the target noun is a head noun is analogous, with the only change being in the calculation of the corpus frequency.

Figure 3 shows how the substitution process works for the target noun *art* in the NC *art museum*.

In addition to the original method presented in Kim and Baldwin [2007], we also tested the impact of normalising in Equation (8) as well as weighting in Equation (9). Although the number of candidates per word sense varies, since we compute the score based on the web count which indicates the frequency of the word senses in use, we surmise that normalising the scores by the number of candidates per sense would not have much effect. With the weighting approach, the motivation is that the order of word senses indicates the frequency of the usage of the words. As

Noun	Number of senses
art	4
authority	7
bar	14
channel	8
child	4
circuit	6
day	10
nature	5
stress	4

Fig. 4. Target noun set, and the polysemy of each noun

such, the higher the word sense is listed in WORDNET, the more frequently it is used in the text. Based on this idea, we add the simple weights per word sense shown in Equation (9):

$$ws^*(n_1) = \operatorname*{argmax}_{s_i \in ws(n_1)} \sum_{n_j \in ss(s_i) \setminus \{s_i\}} \mathit{freq}(n_j, n_2) \tag{8}$$

$$ws^*(n_1) = \operatorname*{argmax}_{s_i \in ws(n_1)} \frac{\sum_{n_j \in ss(s_i) \setminus \{s_i\}} \mathit{freq}(n_j, n_2)}{|ss(s_i) \setminus \{s_i\}|} \cdot Weight_{sense_i} \tag{9}$$

Finally, we developed an indirect method to score the word senses based on the word overlap of neighbouring words in the context between the target word and word-substituted candidates. The motivation behind this is that although the web counts provide tentative information about the frequency of word use, they may not work well with infrequent words. Thus, instead of calculating the scores based on web counts, we count the words which occur in contexts (i.e. neighbouring words) between the target word and word-substituted candidates. Similarly, since the number of candidates per sense differs, we normalise the final score per sense. The method is shown in Equation (10). $Neighbour$ is the number of co-occurring neighbouring words in the contexts of the target word and word-substituted candidates. As context, we use snippets for the top-1000 results when querying Google with the original NCs and the substituted NCs. For example, we first query for *art museum*, then *artifact museum*, and so on. Second, we retrieve the first 1000 snippets for each target NC and word-substituted NC, and count the words occurring in both snippets to calculate the score for each sense.

$$ws^*(n_1) = \operatorname*{argmax}_{s_i \in ws(n_1)} \frac{\sum_{n_j \in ss(s_i) \setminus \{s_i\}} \mathit{freq}(\mathit{Neighbour}_{n_j} \cap \mathit{Neighbour}_{n_2})}{|ss(s_i) \setminus \{s_i\}|} \tag{10}$$

6. DATASET FOR DISAMBIGUATING NOUN COMPOUNDS

To evaluate our method, we initially collected the top-20 frequent polysemous nouns from SEMCOR and the English all-words task from Senseval-2 [Palmer et al. 2001]. Then we identified binary NCs (i.e. noun-noun sequences) in the British National Corpus [Burnard 1995] which contained each of 20 randomly-selected nouns in either the modifier or head noun position (but not both). From this, we extracted polysemous nouns which occurred as both modifier and head noun over at least 50 NC token instances. Finally, we selected the 9 nouns which occurred in the most NCs, as detailed in Figure 4.

As the final dataset, we randomly selected 50 NCs for each of the modifier and head noun positions of the 9 polysemous nouns. Hence, we have 100 NCs for each polysemous noun, totalling 900 instances.

art **lesson**

Sense 1 lesson
=> teaching, instruction,...
=> education
=> profession
=> ...
Sense 2 example, deterrent example..

Fig. 5. Word sense of *art* in different sense collocations

To annotate the word senses of the target nouns in the 900 NCs, we hired two linguistically-trained human annotators. We extracted 50 sentences for each modifier or head noun-positioned target noun from the British National Corpus and provided them to the human annotators. These sentences were intended to help the annotators determine the word senses of target nouns in the case that a given NC was used over a range of sense assignments. Also, the set of sentences was used to take the majority class assignment for the NC type. The initial type-level inter-annotator agreement was 69.2%, and the human annotators met to discuss all instances of disagreement. A single expert annotator also annotated the 900 NCs for SR, once again with reference to their token occurrences in the British National Corpus. In post-analysis of the annotation, we observed that NC monosemy (i.e. a given noun occurring with only one sense across all NCs) helps significantly when sense annotating NCs. We also observed that although one sense tends to be dominant in NCs, determining the majority sense was harder than we expected in some cases where the senses were relatively evenly distributed.

We specified the semantics of each non-target noun by both: (a) CORELEX, and (b) the first-sense and three direct *hypernym*s (all from WORDNET 2.1). In the set of 900 NCs, 61.6% of the collocating nouns were contained in CORELEX. For the remainder, we manually assigned a CORELEX class following the CORELEX clustering methodology. For example, when *lemon* was not found in CORELEX, we assigned the class FOOD, which also contains *orange*. The determination of first sense in WORDNET was based on the first sense learning method of McCarthy et al. [2004]. Figure 5 shows the specification of the semantics of the non-target noun using three direct *hypernym*s. In this case, *art* is the target noun and *lesson* is the non-target noun *art lesson*.

To compute the probability for the unsupervised classifier, we calculated the web count of each synonymy-substituted NC using Google. Lapata and Keller [2004] showed that the web provides reliable probability estimates for tasks including unsupervised noun compound interpretation and bracketing. For our purposes, we generated both the singular and plural forms of each NC using MORPH [Minnen et al. 2001] to calculate the frequency of a given NC. Note that we do not include the target noun itself. If no *synonym*(s) of the target noun are found, then we use *hypernym*s (excluding substitution candidates which have lexical overlap with the target noun). For example, the synset membership of sense_1 of *art* is *art* and *fine art*. As *fine art* includes the word *art*, we look to the *hypernym*s, and end up with the candidates *artifact* and *craft*.

7. EVALUATING THE WSD METHOD

In our first experiment, we attempt to disambiguate the word sense of the target noun in a given NC based on each of the proposed methods. We evaluate the supervised WSD method via 10-fold cross-validation over the 900 NC instances using TIMBL, and we evaluate the unsupervised method over the same 900 instances. We use two unsupervised baselines: (1) random sense assignment, and (2) the first sense prediction for the target noun by the method of McCarthy et al. [2004], based on the full British National Corpus (comprising both NC and non-NC instances of a given target noun). We also have one supervised baseline in the form of a majority class classifier, based on 10-fold cross-validation over the 900 instances. In order to benchmark our results, we ran SENSELEARNER over the dataset using the pre-trained word class models, randomly selecting one of the original

Target noun	Role in NC	Baselines Random	First	Majority	Benchmark SenseLearner
art	modifier	25.00	68.00	68.00	54.00
	head noun	25.00	54.00	54.00	50.00
	both	25.00	61.00	61.00	52.00
authority	modifier	14.29	6.00	78.00	6.00
	head noun	14.29	8.00	60.00	8.00
	both	14.29	7.00	69.00	7.00
bar	modifier	7.14	46.00	46.00	46.00
	head noun	7.14	30.00	24.00	28.00
	both	7.14	38.00	35.00	37.00
channel	modifier	12.50	24.00	24.00	22.00
	head noun	12.50	16.00	26.00	12.00
	both	12.50	20.00	25.00	17.00
child	modifier	25.00	72.00	72.00	60.00
	head noun	25.00	78.00	78.00	76.00
	both	25.00	75.00	75.00	68.00
circuit	modifier	16.67	68.00	68.00	66.00
	head noun	16.67	54.00	54.00	52.00
	both	16.67	61.00	61.00	59.00
day	modifier	10.00	18.00	68.00	14.00
	head noun	10.00	6.00	90.00	6.00
	both	10.00	12.00	79.00	10.00
nature	modifier	20.00	4.00	70.00	4.00
	head noun	20.00	34.00	14.00	32.00
	both	20.00	19.00	42.00	18.00
stress	modifier	20.00	2.00	48.00	2.00
	head noun	20.00	8.00	8.00	8.00
	both	20.00	5.00	28.00	5.00
Total	modifier	15.97	34.22	**60.22**	**30.33**
	head noun	15.97	32.00	**45.33**	**30.44**
	both	15.97	33.10	**52.78**	**30.22**

Fig. 6. Baseline WSD accuracy over each target noun in the modifier and head noun positions

sentential contexts from the British National Corpus for each NC and corresponding sense labeling. The classification accuracy for the output of each WSD method over each target noun, broken down across the modifier and head noun positions, is shown in Figures 6, 7 and 8. The overall best-performing method is indicated in boldface for each position in the NC.

The results show that the best-performing classifier overall was the supervised classifier with CORELEX noun features (54.78% accuracy overall), although some of results are outperformed by the equivalent classifier with WORDNET features (54.00% accuracy overall). Note, however, that the CORELEX results are predicated on considerable manual supplementation of the resource to boost its coverage. As such, despite its marginally lower accuracy, we consider the WORDNET-based classifier to be superior for open-text applications.

Comparing our methods with the benchmark, surprisingly SENSELEARNER performed poorly. We expected our methods to exceed SENSELEARNER since we have fine-tuned our methods to the task. However, considering the performance of SENSELEARNER over Senseval-2 and SEMCOR data, it shows that the one sense per collocation heuristic and word sense combinatorics are stronger predictors of noun sense in NCs than standard contextual features.

Target noun	Role in NC	CORELEX S+GR	CORELEX S+SR	CORELEX S+GR+SR	WORDNET S+GR	WORDNET S+SR	WORDNET S+GR+SR
art	modifier	64.00	60.94	64.00	70.48	67.14	65.24
	head noun	48.00	50.00	50.00	51.01	56.06	56.06
	both	56.00	64.00	57.00	60.75	61.76	60.65
authority	modifier	70.00	71.15	70.00	76.96	74.51	74.51
	head noun	52.00	50.00	50.00	53.85	49.52	49.52
	both	61.00	62.00	60.00	65.41	61.89	62.02
bar	modifier	54.00	47.17	50.00	47.09	47.09	49.03
	head noun	46.00	40.00	44.00	39.71	42.65	42.65
	both	50.00	47.00	47.00	43.40	44.88	45.84
channel	modifier	24.00	21.15	22.00	18.00	20.00	19.50
	head noun	28.00	37.5	40.00	24.39	35.61	31.71
	both	26.00	35.00	31.00	21.20	27.90	25.61
child	modifier	50.00	75.44	74.00	69.29	66.83	64.88
	head noun	76.00	71.15	76.00	76.38	76.37	76.37
	both	63.00	80.00	75.00	72.84	71.32	70.63
circuit	modifier	62.00	58.82	58.00	61.32	61.32	61.32
	head noun	48.00	43.14	50.00	56.50	57.00	57.00
	both	55.00	52.00	54.00	58.91	59.22	59.16
day	modifier	64.00	62.00	62.00	62.09	59.72	59.72
	head noun	88.00	88.68	88.00	88.89	89.39	89.39
	both	76.00	78.00	75.00	75.49	74.08	74.56
nature	modifier	70.00	61.02	58.00	69.61	67.65	65.20
	head noun	44.00	33.93	40.00	38.00	26.00	22.00
	both	57.00	55.00	49.00	53.81	46.53	43.60
stress	modifier	50.00	40.00	40.00	45.81	39.41	49.41
	head noun	24.00	22.22	26.00	26.51	30.23	28.37
	both	37.00	32.00	33.00	36.16	34.69	38.89
Total	modifier	**59.11**	55.74	55.33	58.01	56.12	55.58
	head noun	50.44	48.54	**51.56**	50.00	50.77	49.78
	both	**54.78**	52.14	53.44	54.00	53.48	52.68

Fig. 7. Accuracy for the supervised WSD method over each target noun in the modifier and head noun positions

The majority-class baseline outperformed our WSD methods for nearly half of the target nouns. However, the majority-class baseline performed very poorly compared with our two supervised classifiers for a number of nouns. Note also that our supervised classifiers are hampered by a lack of training data (i.e. 50 instances for each target word). We also found that there was little difference in the performance for the modifier vs. head noun, but in secondary experimentation we confirmed that conditioning the disambiguation on the syntactic role improved accuracy.

Looking at the results for the supervised method, we observed that the grammatical role (GR) information had the greatest impact, and that the accuracy actually dropped slightly when SR information was included. This indicates that the grammatical role is a much stronger determinant of word sense, and that the correlation between SR and word sense is much smaller than we had expected.

The unsupervised method performed well below the supervised methods (both the majority class baseline and the WORDNET and CORELEX classifiers) and slightly below the first sense baseline, at the same combined accuracy as the minimally-supervised SENSELEARNER.

We additionally evaluated the impact of normalisation and weighting in Figure 8. First, we found that despite a slight improvement, it is not clear that score normalisation based on the number of

Target noun	Role in NC	Collocation w/o normal.	Collocation w/ normal.	Collocation normal.+weight	Word Overlap
art	modifier	52.00	44.00	62.00	14.00
	head noun	36.00	30.00	40.00	28.00
	both	44.00	37.00	48.00	51.00
authority	modifier	24.00	18.00	20.00	78.00
	head noun	36.00	36.00	32.00	56.00
	both	30.00	27.00	26.00	67.00
bar	modifier	22.00	20.00	38.00	14.00
	head noun	28.00	24.00	38.00	24.00
	both	25.00	22.00	38.00	19.00
channel	modifier	16.00	26.00	24.00	22.00
	head noun	26.00	30.00	26.00	24.00
	both	21.00	28.00	25.00	23.00
child	modifier	18.00	24.00	50.00	44.00
	head noun	30.00	38.00	56.00	54.00
	both	24.00	31.00	53.00	49.00
circuit	modifier	46.00	62.00	72.00	50.00
	head noun	38.00	42.00	50.00	40.00
	both	42.00	52.00	61.00	45.00
day	modifier	6.00	24.00	38.00	10.00
	head noun	24.00	16.00	16.00	20.00
	both	15.00	20.00	27.00	15.00
nature	modifier	36.00	30.00	26.00	36.00
	head noun	24.00	20.00	26.00	22.00
	both	30.00	25.00	26.00	29.00
stress	modifier	44.00	30.00	40.00	22.00
	head noun	24.00	28.00	20.00	28.00
	both	34.00	29.00	30.00	25.00
Total	modifier	29.33	30.89	**41.11**	32.22
	head noun	29.56	29.33	**33.78**	32.88
	both	29.44	30.11	**37.44**	32.56

Fig. 8. Accuracy for the unsupervised WSD method, with normalisation and weighting

candidates has much impact on results. The highest accuracy came from the weighting approach. Compared to the original method (i.e. Equation (7)), the weighting approach improves accuracy significantly (30.11% vs. 37.44%). We estimated this is due to the first sense heuristic and the introduction of corpus-inspecific sense distribution information. Further, we observed that although the word overlap method did not exceed the accuracy of that for the weighting method, it improved the overall accuracy by over 2% compared to that for the original method. We hypothesise that this is due to the re-weighting of infrequent words.

Finally, we used randomised estimation [Yeh 2000] to confirm that the observed differences between the baselines and our best-performing supervised system, and between SENSELEARNER and our best-performing unsupervised system, were statistically significant.

8. EVALUATING NOUN COMPOUND INTERPRETATION USING WSD

Word sense information used either directly (e.g. Moldovan et al. [2004]) or indirectly (e.g. Kim and Baldwin [2005], Nastase et al. [2006]) has been shown to be one of the highest-impact features when interpreting SRs. However, extracting gold-standard sense information would require excessive amounts of manual annotation. Since we have proposed an automatic method to extract

Method	WSD	CORELEX	WORDNET
Baseline	majority vote	.273	.273
Moldovan et al. [2004]	system-tagged	.402	**.426**
	first-sense	**.403**	.425
	hand-tagged	.447	.540
Kim and Baldwin [2005]	similarity	.346	.346

Fig. 9. Accuracy of interpreting semantic relations in NCs

word senses in this paper, in this section, we evaluate the impact of the word sense data on NC interpretation.

To find the usability of system-tagged word senses, we designed a supervised learning method using the output of a WSD system as sense features to interpret NCs. As sense features, we used hand-tagged sense data which is 100% correct (*hand-tagged* in Figure 9) as well as system-tagged sense data which contains errors (*system-tagged* and *first-sense* in Figure 9). Then we built two supervised classifiers for NC interpretation using NC interpretation method introduced in Moldovan et al. [2004] and Kim and Baldwin [2005]. In this experiment, we used semantic features from CORELEX and WORDNET as before. We compare this directly to a first-sense disambiguation method, trained over the full British National Corpus (i.e. the same as used for the first-sense baseline in WSD). The output of the first sense classifier is combined with the CORELEX and WORDNET features of the collocating noun as above, producing a fully comparable classifier. As a benchmark, we tested the similarity method presented in Kim and Baldwin [2005] which indirectly uses word sense information. This method is based on nearest-neighbour matching over the union of senses of the modifier and head noun, with distance defined by word-level similarity in WORDNET. That is, this method makes use of word sense information but does not attempt to perform explicit WSD.

In evaluation, we used the same dataset we used to evaluate our NC interpretation method described in Section 6. As the baseline, we assigned all NCs the majority class interpretation, namely TOPIC.

Figure 9 shows the classification accuracy of NC interpretation using CORELEX and WORDNET semantic features for the collocating noun. The “system-tagged” senses are those predicted by the WORDNET-based supervised method.

The accuracy of the WORDNET-based supervised classifier and the first sense method are almost identical using both WORDNET and CORELEX, but the upper bound classifier based on hand-tagged data is better than both of these, particularly for the CORELEX representation of the collocating noun. This suggests that the features of the collocating nouns are weighted higher than the noisy word sense features of the target noun, and that to approach the upper bound accuracy, significantly higher WSD performance is required. Both automatic WSD-based methods clearly outperform both the baseline and the benchmark interpretation method which show the impact of word sense for NC interpretation. Note that since all of the NCs in our dataset contain polysemous nouns, the performance of our similarity method is considerably lower than that reported in the original paper.

To answer our original question about whether WSD can contribute to NC interpretation, the answer appears to be a resounding yes. This is significant both in documenting a task where WSD makes a positive impact, and in opening up a new research direction in the field of NC interpretation.

9. CONCLUSION

In this paper, we have investigated word sense distributions in noun compounds. First, we described a method to automatically disambiguate the word sense of polysemous component nouns in NCs, and second, we used the predicted word sense information to interpret NCs. To disambiguate word sense, we used sense collocation and lexical substitution, and built supervised and unsupervised WSD classifiers, achieving an accuracy of 54.78% and 37.44% accuracy, respectively. We also reproduced the methodology of Moldovan et al. [2004] and Kim and Baldwin [2005], and demonstrated that word sense information for component nouns is vital in NC interpretation.

Acknowledgements

This research was supported in part by the Australian Research Council, under grants FT120100658 and DP110101934.

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)*. Columbus, USA, 317–325.

AGIRRE, E., BENGOETXEA, K., GOJENOLA, K., AND NIVRE, J. 2011. Improving dependency parsing with semantic classes. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*. Portland, USA, 699–703.

AGIRRE, E. AND EDMONDS, P. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer, Dordrecht, Netherlands.

AGIRRE, E. AND MARTINEZ, D. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of COLING workshop on Semantic Annotation and Intelligent Content*. Saarbrucken, Germany, 11–19.

ATTERER, M. AND SCHÜTZE, H. 2007. Prepositional phrase attachment without oracles. *Computational Linguistics 33,* 4, 469–476.

BANERJEE, S. AND PEDERSEN, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. Acapulco, Mexico, 805–810.

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, 96–102.

BUITELAAR, P. 1989. Corelex: Systematic polysemy and underspecification. Ph.D. thesis, Brandeis University.

BURNARD, L. 1995. User guide for the British National Corpus.

BUTNARIU, C. AND VEALE, T. 2008. A concept-centered approach to noun-compound interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester, UK, 81–88.

CARPUAT, M. AND WU, D. 2005. Word sense disambiguation vs. staitstical machine translation. In *Proceedings of 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, USA, 387–394.

COPESTAKE, A. AND LASCARIDES, A. 1997. Integrating symbolic and statistical representations: The lexicon pragmatics interface. In *Proceedings of the 35th Annual Meeting of the Association of Coomputational Linguistics and 8th Conference of the European Chapter of Association of Computational Linguistics*. Madrid, Spain, 136–143.

DAELEMANS, W., ZAVREL, J., VAN DER SLOOT, K., AND VAN DEN BOSCH, A. 2004. TiMBL: Tilburg memory based learner, version 5.1, reference guide.

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

FELLBAUM, C., Ed. 1998. *WordNet, An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, USA.

FININ, T. W. 1980. The semantic interpretation of compound nominals. Ph.D. thesis, University of Illinois, Urbana Champaign.

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*. Prague, Czech Republic, 13–18.

HERMJAKOB, U., HOVY, E., AND LIN, C.-Y. 2002. Automated question answering in Webclopedia. In *Proceedings of the ACL-02 Demonstrations Session*. Philadelphia, USA, 98–99.

IDE, N. AND VERONIS, J. 1998. Word sense disambiguation : The state of the art. *Computational Linguistics 24,* 1, 1–40.

KIM, S. N. AND BALDWIN, T. 2005. Automatic interpretation of compound nouns using WordNet similarity. In *Proceedings of 2nd International Joint Conference on Natual Language Processing*. Jeju, Korea, 945–956.

KIM, S. N. AND BALDWIN, T. 2006a. Automatic extraction of verb-particles using linguistic features. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*. Trento, Italy, 65–72.

KIM, S. N. AND BALDWIN, T. 2006b. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*. Sydney, Australia, 491–498.

KIM, S. N. AND BALDWIN, T. 2007. Disambiguating noun compounds. In *Proceedings of 22nd AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 901–906.

LANDES, S., LEACOCK, C., AND TENGI, R. I. 1998. Building semantic concordances. In *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed. MIT Press, Cambridge, USA.

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*. Boston, USA, 121–128.

LEACOCK, C. AND CHODOROW, N. 1998. *Combining local context and WordNet similarity for word sense identification*. MIT Press, Cambridge, USA.

LEVI, J. 1979. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York, New York, USA.

MACKINLAY, A., DRIDAN, R., MCCARTHY, D., AND BALDWIN, T. 2012. The effects of semantic annotations on precision parse ranking. In *Proc. of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*. Montréal, Canada, 228–236.

MARCUS, M. P., SANTORINI, B., AND MARCINKIEWICZ, M. A. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics 19,* 2, 313–330.

MCCARTHY, D., KOELING, R., WEEDS, J., AND CARROLL, J. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*. Barcelona, Spain, 280–287.

MIHALCEA, R. AND FARUQUE, E. 2004. Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of the ACL/SIGLEX Senseval-3*. Barcelona, Spain, 155–158.

MIHALCEA, R. AND MOLDOVAN, D. 1999. An automatic method for generating sense taggged corpora. In *Proceedings of the 16th Conference of the American Association of Aritificial Intelligence*. Orlando, USA, 461–466.

MINNEN, G., CARROLL, J., AND PEARCE, D. 2001. Applied morphological processing of English. *Natural Language Engineering 7,* 3, 207–223.

MOLDOVAN, D., BADULESCU, A., TATU, M., ANTOHE, D., AND GIRJU, R. 2004. Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL 2004: Workshop on Computational Lexical Semantics*. Boston, USA, 60–67.

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*. Bularia, 233–244.

NAKOV, P. AND HEARST, M. 2008. Solving relational similarity problems using the web as a corpus. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Columbus, OH, USA, 452–460.

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*. Boston, USA, 781–787.

NAVIGLI, R. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys 41,* 2.

NG, H. T. AND LEE, H. B. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, USA, 40–47.

Ó SÉAGHDHA, D. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proceedings of the 4th Corpus Linguistics Conference*.

PALMER, M., FELLBAUM, C., COTTON, S., DELFS, L., AND DANG, H. T. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Toulouse, France, 21–24.

PRAGER, J. AND CHU-CARROLL, J. 2001. Use of WordNet hypernyms for answering what-is questions.

PUSTEJOVSKY, J. 1995. *The Generative Lexicon*. MIT press, Cambridge, USA.

ROSARIO, B. AND MARTI, H. 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*. Pittsburgh, Pennsylvania, USA, 82–90.

SANDERSON, M. 1996. Word sense disambiguation and information retrieval. Ph.D. thesis, University of Glasgow.

SPARCK JONES, K. 1983. *Compound noun interpretation problems*. Prentice-Hall, Englewood Cliffes, 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*. Uppsala, Sweden, 678–687.

VANDERWENDE, L. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th Conference on Computational linguistics*. Kyoto, Japan, 782–788.

VICKREY, D., BIEWALD, L., TEYSSIER, M., AND KOLLER, D. 2005. Word-sense disambiguation for machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, Canada, 771–778.

YAROWSKY, D. 1993. One sense per collocation. In *Proceedings of the ARPA Human Language Technology Workshop*. Plainsboro, New Jerey, USA, 266–271.

YAROWSKY, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association of Computational Linguistics*. Cambridge, USA, 189–196.

YEH, A. 2000. More accurate tests for the statistical significance of result differences. In *International Conference on Computational Linguistics (COLING)*. 947–953.