Solving Relational Similarity Problems Using the Web as a Corpus

Preslav Nakov*
EECS, CS division
University of California at Berkeley
Berkeley, CA 94720, USA
nakov@cs.berkeley.edu

Marti A. Hearst
School of Information
University of California at Berkeley
Berkeley, CA 94720, USA
hearst@ischool.berkeley.edu

Abstract

We present a simple linguistically-motivated method for characterizing the semantic relations that hold between two nouns. The approach leverages the vast size of the Web in order to build lexically-specific features. The main idea is to look for verbs, prepositions, and coordinating conjunctions that can help make explicit the hidden relations between the target nouns. Using these features in instance-based classifiers, we demonstrate state-of-the-art results on various relational similarity problems, including mapping noun-modifier pairs to abstract relations like TIME, LOCATION and CONTAINER, characterizing noun-noun compounds in terms of abstract linguistic predicates like CAUSE, USE, and FROM, classifying the relations between nominals in context, and solving SAT verbal analogy problems. In essence, the approach puts together some existing ideas, showing that they apply generally to various semantic tasks, finding that verbs are especially useful features.

1 Introduction

Despite the tremendous amount of work on word similarity (see (Budanitsky and Hirst, 2006) for an overview), there is surprisingly little research on the important related problem of *relational similarity* – semantic similarity between pairs of words. Students who took the SAT test before 2005 or who are taking the GRE test nowadays are familiar with an instance of this problem – verbal analogy questions, which ask whether, e.g., the relationship between *ostrich* and *bird* is more similar to that between *lion* and *cat*, or rather between *primate* and *monkey*. These analogies are difficult, and the average test taker gives a correct answer 57% of the time (Turney and Littman, 2005).

Many NLP applications could benefit from solving relational similarity problems, including but not limited to question answering, information retrieval, machine translation, word sense disambiguation, and information extraction. For example, a relational search engine like TextRunner, which serves queries like "find all X such that X causes wrinkles", asking for all entities that are in a particular relation with a given entity (Cafarella et al., 2006), needs to recognize that *laugh wrinkles* is an instance of CAUSE−EFFECT. While there are not many success stories so far, measuring semantic similarity has proven its advantages for textual entailment (Tatu and Moldovan, 2005).

In this paper, we introduce a novel linguistically-motivated Web-based approach to relational similarity, which, despite its simplicity, achieves state-of-the-art performance on a number of problems. Following Turney (2006b), we test our approach on SAT verbal analogy questions and on mapping noun-modifier pairs to abstract relations like TIME, LOCATION and CONTAINER. We further apply it to (1) characterizing noun-noun compounds using abstract linguistic predicates like CAUSE, USE, and FROM, and (2) classifying the relation between pairs of nominals in context.

*After January 2008 at the Linguistic Modeling Department, Institute for Parallel Processing, Bulgarian Academy of Sciences, nakov@lml.bas.bg

2 Related Work

2.1 Characterizing Semantic Relations

Turney and Littman (2005) characterize the relationship between two words as a vector with coordinates corresponding to the Web frequencies of 128 fixed phrases like "X for Y" and "Y for X" instantiated from a fixed set of 64 joining terms like *for*, *such as*, *not the*, *is* *, etc. These vectors are used in a nearest-neighbor classifier to solve SAT verbal analogy problems, yielding 47% accuracy. The same approach is applied to classifying noun-modifier pairs: using the *Diverse* dataset of Nastase and Szpakowicz (2003), Turney&Littman achieve F-measures of 26.5% with 30 fine-grained relations, and 43.2% with 5 course-grained relations.

Turney (2005) extends the above approach by introducing the latent relational analysis (LRA), which uses automatically generated synonyms, learns suitable patterns, and performs singular value decomposition in order to smooth the frequencies. The full algorithm consists of 12 steps described in detail in (Turney, 2006b). When applied to SAT questions, it achieves the state-of-the-art accuracy of 56%. On the *Diverse* dataset, it yields an F-measure of 39.8% with 30 classes, and 58% with 5 classes.

Turney (2006a) presents an unsupervised algorithm for mining the Web for patterns expressing implicit semantic relations. For example, CAUSE (e.g., *cold virus*) is best characterized by "*Y * causes X*", and "*Y in * early X*" is the best pattern for TEMPORAL (e.g., *morning frost*). With 5 classes, he achieves F-measure=50.2%.

2.2 Noun-Noun Compound Semantics

Lauer (1995) reduces the problem of noun compound interpretation to choosing the best paraphrasing preposition from the following set: *of*, *for*, *in*, *at*, *on*, *from*, *with* or *about*. He achieved 40% accuracy using corpus frequencies. This result was improved to 55.7% by Lapata and Keller (2005) who used Web-derived n-gram frequencies.

Barker and Szpakowicz (1998) use syntactic clues and the identity of the nouns in a nearest-neighbor classifier, achieving 60-70% accuracy.

Rosario and Hearst (2001) used a discriminative classifier to assign 18 relations for noun compounds from biomedical text, achieving 60% accuracy.

Rosario et al. (2002) reported 90% accuracy with a "descent of hierarchy" approach which characterizes the relationship between the nouns in a bioscience noun-noun compound based on the MeSH categories the nouns belong to.

Girju et al. (2005) apply both classic (SVM and decision trees) and novel supervised models (semantic scattering and iterative semantic specialization), using *WordNet*, word sense disambiguation, and a set of linguistic features. They test their system against both Lauer's 8 prepositional paraphrases and another set of 21 semantic relations, achieving up to 54% accuracy on the latter.

In a previous work (Nakov and Hearst, 2006), we have shown that the relationship between the nouns in a noun-noun compound can be characterized using verbs extracted from the Web, but we provided no formal evaluation.

Kim and Baldwin (2006) characterized the semantic relationship in a noun-noun compound using the verbs connecting the two nouns by comparing them to predefined seed verbs. Their approach is highly resource intensive (uses *WordNet*, *CoreLex* and *Moby's thesaurus*), and is quite sensitive to the seed set of verbs: on a collection of 453 examples and 19 relations, they achieved 52.6% accuracy with 84 seed verbs, but only 46.7% with 57 seed verbs.

2.3 Paraphrase Acquisition

Our method of extraction of paraphrasing verbs and prepositions is similar to previous paraphrase acquisition approaches. Lin and Pantel (2001) extract paraphrases from dependency tree paths whose ends contain semantically similar sets of words by generalizing over these ends. For example, given "*X solves Y*", they extract paraphrases like "*X finds a solution to Y*", "*X tries to solve Y*", "*X resolves Y*", "*Y is resolved by X*", etc. The approach is extended by Shinyama et al. (2002), who use named entity recognizers and look for anchors belonging to matching semantic classes, e.g., LOCATION, ORGANIZATION. The idea is further extended by Nakov et al. (2004), who apply it in the biomedical domain, imposing the additional restriction that the sentences from which the paraphrases are extracted cite the same target paper.

2.4 Word Similarity

Another important group of related work is on using syntactic dependency features in a vector-space model for measuring *word* similarity, e.g., (Alshawi and Carter, 1994), (Grishman and Sterling, 1994), (Ruge, 1992), and (Lin, 1998). For example, given a noun, Lin (1998) extracts verbs that have that noun as a subject or object, and adjectives that modify it.

3 Method

Given a pair of nouns, we try to characterize the semantic relation between them by leveraging the vast size of the Web to build linguistically-motivated lexically-specific features. We mine the Web for sentences containing the target nouns, and we extract the connecting verbs, prepositions, and coordinating conjunctions, which we use in a vector-space model to measure relational similarity.

The process of extraction starts with exact phrase queries issued against a Web search engine (*Google*) using the following patterns:

"$infl_1$ THAT $\star$ $infl_2$"
"$infl_2$ THAT $\star$ $infl_1$"
"$infl_1$ $\star$ $infl_2$"
"$infl_2$ $\star$ $infl_1$"

where: $infl_1$ and $infl_2$ are inflected variants of $noun_1$ and $noun_2$ generated using the *Java WordNet Library*[1]; THAT is a complementizer and can be *that*, *which*, or *who*; and $\star$ stands for 0 or more (up to 8) instances of *Google*'s star operator.

The first two patterns are subsumed by the last two and are used to obtain more sentences from the search engine since including e.g. *that* in the query changes the set of returned results and their ranking.

For each query, we collect the text snippets from the result set (up to 1,000 per query). We split them into sentences, and we filter out all incomplete ones and those that do not contain the target nouns. We further make sure that the word sequence following the second mentioned target noun is nonempty and contains at least one nonnoun, thus ensuring the snippet includes the entire noun phrase: snippets representing incomplete sentences often end with a period anyway. We then perform POS tagging using the *Stanford POS tagger* (Toutanova et al., 2003) and shallow parsing with the *OpenNLP tools*[2], and we extract the following types of features:

Freq.	Feature	POS	Direction
2205	of	P	$2 \rightarrow 1$
1923	be	V	$1 \rightarrow 2$
771	include	V	$1 \rightarrow 2$
382	serve on	V	$2 \rightarrow 1$
189	chair	V	$2 \rightarrow 1$
189	have	V	$1 \rightarrow 2$
169	consist of	V	$1 \rightarrow 2$
148	comprise	V	$1 \rightarrow 2$
106	sit on	V	$2 \rightarrow 1$
81	be chaired by	V	$1 \rightarrow 2$
78	appoint	V	$1 \rightarrow 2$
77	on	P	$2 \rightarrow 1$
66	and	C	$1 \rightarrow 2$
66	be elected	V	$1 \rightarrow 2$
58	replace	V	$1 \rightarrow 2$
48	lead	V	$2 \rightarrow 1$
47	be intended for	V	$1 \rightarrow 2$
45	join	V	$2 \rightarrow 1$
...	...	...	...
4	be signed up for	V	$2 \rightarrow 1$

Table 1: **The most frequent Web-derived features for *committee member*.** Here V stands for verb (possibly +preposition and/or +particle), P for preposition and C for coordinating conjunction; $1 \rightarrow 2$ means *committee* precedes the feature and *member* follows it; $2 \rightarrow 1$ means *member* precedes the feature and *committee* follows it.

Verb: We extract a verb if the subject NP of that verb is headed by one of the target nouns (or an inflected form), and its direct object NP is headed by the other target noun (or an inflected form). For example, the verb *include* will be extracted from "The *committee* includes many *members*." We also extract verbs from relative clauses, e.g., "This is a *committee* which includes many *members*." Verb particles are also recognized, e.g., "The *committee* must rotate off 1/3 of its *members*." We ignore modals and auxiliaries, but retain the passive *be*. Finally, we lemmatize the main verb using *WordNet*'s morphological analyzer *Morphy* (Fellbaum, 1998).

Verb+Preposition: If the subject NP of a verb is headed by one of the target nouns (or an inflected form), and its indirect object is a PP containing an NP which is headed by the other target noun (or an inflected form), we extract the verb and the preposi-

[1]JWNL: http://jwordnet.sourceforge.net

[2]OpenNLP: http://opennlp.sourceforge.net

tion heading that PP, e.g., "The thesis advisory *committee* consists of three qualified *members*." As in the verb case, we extract verb+preposition from relative clauses, we include particles, we ignore modals and auxiliaries, and we lemmatize the verbs.

Preposition: If one of the target nouns is the head of an NP containing a PP with an internal NP headed by the other target noun (or an inflected form), we extract the preposition heading that PP, e.g., "The *members* of the *committee* held a meeting."

Coordinating conjunction: If the two target nouns are the heads of coordinated NPs, we extract the coordinating conjunction.

In addition to the lexical part, for each extracted feature, we keep a direction. Therefore the preposition *of* represents two different features in the following examples "*member* of the *committee*" and "*committee* of *members*". See Table 1 for examples.

We use the above-described features to calculate relational similarity, i.e., similarity between pairs of nouns. In order to downweight very common features like *of*, we use TF.IDF-weighting:

$$w(x) = TF(x) \times \log\left(\frac{N}{DF(x)}\right) \qquad (1)$$

In the above formula, $TF(x)$ is the number of times the feature x has been extracted for the target noun pair, $DF(x)$ is the total number of training noun pairs that have that feature, and N is the total number of training noun pairs.

Given two nouns and their TF.IDF-weighted frequency vectors A and B, we calculate the similarity between them using the following generalized variant of the Dice coefficient:

$$Dice(A,B) = \frac{2 \times \sum_{i=1}^{n} \min(a_i, b_i)}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i} \qquad (2)$$

Other variants are also possible, e.g., Lin (1998).

4 Relational Similarity Experiments

4.1 SAT Verbal Analogy

Following Turney (2006b), we use *SAT verbal analogy* as a benchmark problem for relational similarity. We experiment with the 374 SAT questions collected by Turney and Littman (2005). Table 2 shows two sample questions: the top word pairs are called *stems*, the ones in italic are the *solutions*, and the remaining ones are *distractors*. Turney (2006b) achieves 56% accuracy on this dataset, which matches the average human performance of 57%, and represents a significant improvement over the 20% random-guessing baseline.

	ostrich:**bird**		**palatable**:**toothsome**
(a)	*lion*:*cat*	(a)	rancid:fragrant
(b)	goose:flock	(b)	chewy:textured
(c)	ewe:sheep	*(c)*	*coarse*:*rough*
(d)	cub:bear	(d)	solitude:company
(e)	primate:monkey	(e)	no choice

Table 2: **SAT verbal analogy: sample questions.** The stem is in **bold**, the correct answer is in *italic*, and the distractors are in plain text.

Note that the righthand side example in Table 2 is missing one distractor; so do 21 questions. The dataset also mixes different parts of speech: while *solitude* and *company* are nouns, all remaining words are adjectives. Other examples contain verbs and adverbs, and even relate pairs of different POS. This is problematic for our approach, which requires that both words be nouns[3]. After having filtered all examples containing nonnouns, we ended up with 184 questions, which we used in the evaluation.

Given a verbal analogy example, we build six feature vectors – one for each of the six word pairs. We then calculate the relational similarity between the stem of the analogy and each of the five candidates, and we choose the pair with the highest score; we make no prediction in case of a tie.

The evaluation results for a leave-one-out cross-validation are shown in Table 3. We also show 95%-confidence intervals for the accuracy. The last line in the table shows the performance of Turney's LRA when limited to the 184 noun-only examples. Our best model $v + p + c$ performs a bit better, 71.3% vs. 67.4%, but the difference is not statistically significant. However, this "inferred" accuracy could be misleading, and the LRA could have performed better if it was restricted to solve *noun-only* analogies, which seem easier than the general ones, as demonstrated by the significant increase in accuracy for LRA when limited to nouns: 67.4% vs. 56%.

[3]It can be extended to handle adjective-noun pairs as well, as demonstrated in section 4.2 below.

Model	✓	×	∅	Accuracy	Cover.
$v+p+c$	129	52	3	**71.3±7.0**	**98.4**
v	122	56	6	68.5±7.2	96.7
$v+p$	119	61	4	66.1±7.2	97.8
$v+c$	117	62	5	65.4±7.2	97.3
$p+c$	90	90	4	50.0±7.2	97.8
p	84	94	6	47.2±7.2	96.7
baseline	37	147	0	20.0±5.2	100.0
LRA	122	59	3	67.4±7.1	98.4

Table 3: **SAT verbal analogy: 184 noun-only examples.** v stands for verb, p for preposition, and c for coordinating conjunction. For each model, the number of correct (✓), wrong (×), and nonclassified examples (∅) is shown, followed by accuracy and coverage (in %s).

Model	✓	×	∅	Accuracy	Cover.
$v+p$	240	352	8	**40.5±3.9**	**98.7**
$v+p+c$	238	354	8	40.2±3.9	98.7
v	234	350	16	40.1±3.9	97.3
$v+c$	230	362	8	38.9±3.8	98.7
$p+c$	114	471	15	19.5±3.0	97.5
p	110	475	15	19.1±3.0	97.5
baseline	49	551	0	8.2±1.9	100.0
LRA	239	361	0	39.8±3.8	100.0

Table 4: **Head-modifier relations, 30 classes:** evaluation on the *Diverse* dataset, micro-averaged (in %s).

4.2 Head-Modifier Relations

Next, we experiment with the *Diverse* dataset of Barker and Szpakowicz (1998), which consists of 600 head-modifier pairs: noun-noun, adjective-noun and adverb-noun. Each example is annotated with one of 30 fine-grained relations, which are further grouped into the following 5 coarse-grained classes (the fine-grained relations are shown in parentheses): `CAUSALITY` (*cause, effect, purpose, detraction*), `TEMPORALITY` (*frequency, time_at, time_through*), `SPATIAL` (*direction, location, location_at, location_from*), `PARTICIPANT` (*agent, beneficiary, instrument, object, object_property, part, possessor, property, product, source, stative, whole*) and `QUALITY` (*container, content, equative, material, measure, topic, type*). For example, *exam anxiety* is classified as *effect* and therefore as `CAUSALITY`, and *blue book* is *property* and therefore also `PARTICIPANT`.

Some examples in the dataset are problematic for our method. First, in three cases, there are two modifiers, e.g., *infectious disease agent*, and we had to ignore the first one. Second, seven examples have an adverb modifier, e.g., *daily exercise*, and 262 examples have an adjective modifier, e.g., *tiny cloud*. We treat them as if the modifier was a noun, which works in many cases, since many adjectives and adverbs can be used predicatively, e.g., '*This exercise is performed daily.*' or '*This cloud looks very tiny.*'

For the evaluation, we created a feature vector for each head-modifier pair, and we performed a leave-one-out cross-validation: we left one example for testing and we trained on the remaining 599 ones, repeating this procedure 600 times so that each example be used for testing. Following Turney and Littman (2005) we used a 1-nearest-neighbor classifier. We calculated the similarity between the feature vector of the testing example and each of the training examples' vectors. If there was a unique most similar training example, we predicted its class, and if there were ties, we chose the class predicted by the majority of tied examples, if there was a majority.

The results for the 30-class *Diverse* dataset are shown in Table 4. Our best model achieves 40.5% accuracy, which is slightly better than LRA's 39.8%, but the difference is not statistically significant.

Table 4 shows that the verbs are the most important features, yielding about 40% accuracy regardless of whether used alone or in combination with prepositions and/or coordinating conjunctions; not using them results in 50% drop in accuracy.

The reason coordinating conjunctions do not help is that head-modifier relations are typically expressed with verbal or prepositional paraphrases. Therefore, coordinating conjunctions only help with some infrequent relations like *equative*, e.g., finding *player and coach* on the Web suggests an equative relation for *player coach* (and for *coach player*).

As Table 3 shows, this is different for SAT verbal analogy, where verbs are still the most important feature type and the only whose presence/absence makes a statistical difference. However, this time coordinating conjunctions (with prepositions) do help a bit (the difference is not statistically significant) since SAT verbal analogy questions ask for a broader range of relations, e.g., antonymy, for which coordinating conjunctions like *but* are helpful.

Model	Accuracy
$v + p + c + sent + query$ (type C)	**68.1±4.0**
v	67.9±4.0
$v + p + c$	67.8±4.0
$v + p + c + sent$ (type A)	**67.3±4.0**
$v + p$	66.9±4.0
$sent$ (sentence words only)	59.3±4.2
p	58.4±4.2
Baseline (majority class)	57.0±4.2
$v + p + c + sent + query$ (C), 8 stars	67.0±4.0
$v + p + c + sent$ (A), 8 stars	65.4±4.1
Best type C on *SemEval*	67.0±4.0
Best type A on *SemEval*	66.0±4.1

Table 5: **Relations between nominals:** evaluation on the *SemEval* dataset. Accuracy is macro-averaged (in %s), up to 10 *Google* stars are used unless otherwise stated.

4.3 Relations Between Nominals

We further experimented with the *SemEval'07* task 4 dataset (Girju et al., 2007), where each example consists of a sentence, a target semantic relation, two nominals to be judged on whether they are in that relation, manually annotated *WordNet* senses, and the Web query used to obtain the sentence:

```
"Among the contents of the
<e1>vessel</e1> were a set of
carpenter's <e2>tools</e2>, several
large storage jars, ceramic utensils,
ropes and remnants of food, as well
as a heavy load of ballast stones."
WordNet(e1) = "vessel%1:06:00::",
WordNet(e2) = "tool%1:06:00::",
Content-Container(e2, e1) = "true",
Query = "contents of the * were a"
```

The following nonexhaustive and possibly overlapping relations are possible: `Cause-Effect` (e.g., *hormone-growth*), `Instrument-Agency` (e.g., *laser-printer*), `Theme-Tool` (e.g., *work-force*), `Origin-Entity` (e.g., *grain-alcohol*), `Content-Container` (e.g., *bananas-basket*), `Product-Producer` (e.g., *honey-bee*), and `Part-Whole` (e.g., *leg-table*). Each relation is considered in isolation; there are 140 training and at least 70 test examples per relation.

Given an example, we reduced the target entities e_1 and e_2 to single nouns by retaining their heads only. We then mined the Web for sentences containing these nouns, and we extracted the above-described feature types: verbs, prepositions and coordinating conjunctions. We further used the following problem-specific contextual feature types:

Sentence words: after stop words removal and stemming with the Porter (1980) stemmer;

Entity words: lemmata of the words in e_1 and e_2;

Query words: words part of the query string.

Each feature type has a specific prefix which prevents it from mixing with other feature types; the last feature type is used for type C only (see below).

The *SemEval* competition defines four types of systems, depending on whether the manually annotated *WordNet* senses and the *Google* query are used: A (WordNet=no, Query=no), B (WordNet=yes, Query=no), C (WordNet=no, Query=yes), and D (WordNet=yes, Query=yes). We experimented with types A and C only since we believe that having the manually annotated *WordNet* sense keys is an unrealistic assumption for a real-world application.

As before, we used a 1-nearest-neighbor classifier with TF.IDF-weighting, breaking ties by predicting the majority class on the training data. The evaluation results are shown in Table 5. We studied the effect of different subsets of features and of more *Google* star operators. As the table shows, using up to ten *Google* stars instead of up to eight (see section 3) yields a slight improvement in accuracy for systems of both type A (65.4% vs. 67.3%) and type C (67.0% vs. 68.1%). Both results represent a statistically significant improvement over the majority class baseline and over using sentence words only, and a slight improvement over the best type A and type C systems on *SemEval'07*, which achieved 66% and 67% accuracy, respectively.[4]

4.4 Noun-Noun Compound Relations

The last dataset we experimented with is a subset of the 387 examples listed in the appendix of (Levi, 1978). Levi's theory is one of the most important linguistic theories of the syntax and semantics of *complex nominals* – a general concept grouping

[4]The best type B system on *SemEval* achieved 76.3% accuracy using the manually-annotated *WordNet* senses in context for each example, which constitutes an additional data source, as opposed to an additional resource. The systems that used *WordNet* as a resource only, i.e., ignoring the manually annotated senses, were classified as type A or C. (Girju et al., 2007)

Model	USING THAT				NOT USING THAT			
	Accuracy	Cover.	ANF	ASF	Accuracy	Cover.	ANF	ASF
Human: all v	78.4±6.0	99.5	34.3	70.9	–		–	–
Human: first v from each worker	72.3±6.4	99.5	11.6	25.5	–	–	–	–
$v + p + c$	50.0±6.7	99.1	216.6	1716.0	49.1±6.7	99.1	206.6	1647.6
$v + p$	50.0±6.7	99.1	208.9	1427.9	47.6±6.6	99.1	198.9	1359.5
$v + c$	46.7±6.6	99.1	187.8	1107.2	43.9±6.5	99.1	177.8	1038.8
v	45.8±6.6	99.1	180.0	819.1	42.9±6.5	99.1	170.0	750.7
p	33.0±6.0	99.1	28.9	608.8	33.0±6.0	99.1	28.9	608.8
$p + c$	32.1±5.9	99.1	36.6	896.9	32.1±5.9	99.1	36.6	896.9
Baseline	19.6±4.8	100.0	–	–	–	–	–	–

Table 6: **Noun-noun compound relations, 12 classes:** evaluation on *Levi-214* dataset. Shown are micro-averaged accuracy and coverage in %s, followed by average number of features (ANF) and average sum of feature frequencies (ASF) per example. The righthand side reports the results when the query patterns involving THAT were not used. For comparison purposes, the top rows show the performance with the human-proposed verbs used as features.

together the partially overlapping classes of nominal compounds (e.g., *peanut butter*), nominalizations (e.g., *dream analysis*), and nonpredicate noun phrases (e.g., *electric shock*).

In Levi's theory, complex nominals can be derived from relative clauses by removing one of the following 12 abstract predicates: $\texttt{CAUSE}_1$ (e.g., *tear gas*), $\texttt{CAUSE}_2$ (e.g., *drug deaths*), $\texttt{HAVE}_1$ (e.g., *apple cake*), $\texttt{HAVE}_2$ (e.g., *lemon peel*), $\texttt{MAKE}_1$ (e.g., *silkworm*), $\texttt{MAKE}_2$ (e.g., *snowball*), USE (e.g., *steam iron*), BE (e.g., *soldier ant*), IN (e.g., *field mouse*), FOR (e.g., *horse doctor*), FROM (e.g., *olive oil*), and ABOUT (e.g., *price war*). In the resulting nominals, the modifier is typically the object of the predicate; when it is the subject, the predicate is marked with the index 2. The second derivational mechanism in the theory is nominalization; it produces nominals whose head is a nominalized verb.

Since we are interested in noun compounds only, we manually cleansed the set of 387 examples. We first excluded all concatenations (e.g., *silkworm*) and examples with adjectival modifiers (e.g., *electric shock*), thus obtaining 250 noun-noun compounds (*Levi-250* dataset). We further filtered out all nominalizations for which the dataset provides no abstract predicate (e.g., *city planner*), thus ending up with 214 examples (*Levi-214* dataset).

As in the previous experiments, for each of the 214 noun-noun compounds, we mined the Web for sentences containing both target nouns, from which we extracted paraphrasing verbs, prepositions and coordinating conjunctions. We then performed leave-one-out cross-validation experiments with a 1-nearest-neighbor classifier, trying to predict the correct predicate for the testing example. The results are shown in Table 6. As we can see, using prepositions alone yields about 33% accuracy, which is a statistically significant improvement over the majority-class baseline. Overall, the most important features are the verbs: they yield 45.8% accuracy when used alone, and 50% together with prepositions. Adding coordinating conjunctions helps a bit with verbs, but not with prepositions. Note however that none of the differences between the different feature combinations involving verbs are statistically significant.

The righthand side of the table reports the results when the query patterns involving THAT (see section 3) were not used. We can observe a small 1-3% drop in accuracy for all models involving verbs, but it is not statistically significant.

We also show the average number of distinct features and sum of feature counts per example: as we can see, there is a strong positive correlation between number of features and accuracy.

5 Comparison to Human Judgments

Since in all above tasks the most important features were the verbs, we decided to compare our Web-derived verbs to human-proposed ones for all noun-noun compounds in the *Levi-250* dataset. We asked human subjects to produce verbs, possibly

followed by prepositions, that could be used in a paraphrase involving *that*. For example, *olive oil* can be paraphrased as '*oil that comes from olives*', '*oil that is obtained from olives*' or '*oil that is from olives*'. Note that this implicitly allows for prepositional paraphrases – when the verb is to *be* and is followed by a preposition, as in the last paraphrase.

We used the *Amazon Mechanical Turk* Web service[5] to recruit human subjects, and we instructed them to propose at least three paraphrasing verbs per noun-noun compound, if possible. We randomly distributed the noun-noun compounds into groups of 5 and we requested 25 different human subjects per group. Each human subject was allowed to work on any number of groups, but not on the same one twice. A total of 174 different human subjects produced 19,018 verbs. After filtering the bad submissions and normalizing the verbs, we ended up with 17,821 verbs. See (Nakov, 2007) for further details on the process of extraction and cleansing. The dataset itself is freely available (Nakov, 2008).

We compared the human-proposed and the Web-derived verbs for *Levi-214*, aggregated by relation. Given a relation, we collected all verbs belonging to noun-noun compounds from that relation together with their frequencies. From a vector-space model point of view, we summed their corresponding frequency vectors. We did this separately for the human- and the program-generated verbs, and we compared the resulting vectors using Dice coefficient with TF.IDF, calculated as before. Figure 1 shows the cosine correlations using all human-proposed verbs and the first verb from each judge. We can see a very-high correlation (mid-70% to mid-90%) for relations like $\texttt{CAUSE}_1$, $\texttt{MAKE}_1$, BE, but low correlations of 11-30% for reverse relations like $\texttt{HAVE}_2$ and $\texttt{MAKE}_2$. Interestingly, using the first verb only improves the results for highly-correlated relations, but negatively affects low-correlated ones.

Finally, we repeated the cross-validation experiment with the *Levi-214* dataset, this time using the human-proposed verbs[6] as features. As Table 6 shows, we achieved 78.4% accuracy using all verbs (and and 72.3% with the first verb from each worker), which is a statistically significant improvement over the 50% of our best Web-based model. This result is strong for a 12-way classification problem, and confirms our observation that verbs and prepositions are among the most important features for relational similarity problems. It further suggests that the human-proposed verbs might be an upper bound on the accuracy that could be achieved with automatically extracted features.

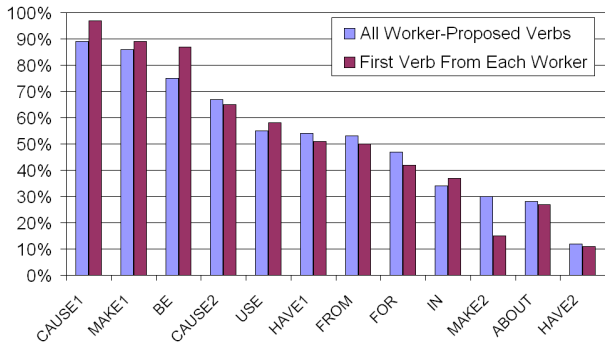

Figure 1: **Cosine correlation (in %s) between the human- and the program- generated verbs by relation:** using all human-proposed verbs vs. the first verb.

6 Conclusions and Future Work

We have presented a simple approach for characterizing the relation between a pair of nouns in terms of linguistically-motivated features which could be useful for many NLP tasks. We found that verbs were especially useful features for this task. An important advantage of the approach is that it does not require knowledge about the semantics of the individual nouns. A potential drawback is that it might not work well for low-frequency words.

The evaluation on several relational similarity problems, including SAT verbal analogy, head-modifier relations, and relations between complex nominals has shown state-of-the-art performance. The presented approach can be further extended to other combinations of parts of speech: not just noun-noun and adjective-noun. Using a parser with a richer set of syntactic dependency features, e.g., as proposed by Padó and Lapata (2007), is another promising direction for future work.

Acknowledgments

This research was supported in part by NSF DBI-0317510.

[5] http://www.mturk.com

[6] Note that the human subjects proposed their verbs without any context and independently of our Web-derived sentences.

References

Hiyan Alshawi and David Carter. 1994. Training and scaling preference functions for disambiguation. *Computational Linguistics*, 20(4):635–648.

Ken Barker and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proc. of Computational linguistics*, pages 96–102.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Michael Cafarella, Michele Banko, and Oren Etzioni. 2006. Relational Web search. Technical Report 2006-04-02, University of Washington, Department of Computer Science and Engineering.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Journal of Computer Speech and Language - Special Issue on Multiword Expressions*, 4(19):479–496.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of SemEval*, pages 13–18, Prague, Czech Republic.

Ralph Grishman and John Sterling. 1994. Generalizing automatically generated selectional patterns. In *Proceedings of the 15th conference on Computational linguistics*, pages 742–747.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 491–498.

Mirella Lapata and Frank Keller. 2005. Web-based models for natural language processing. *ACM Trans. Speech Lang. Process.*, 2(1):3.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Dept. of Computing, Macquarie University, Australia.

Judith Levi. 1978. *The Syntax and Semantics of Complex Nominals.* Academic Press, New York.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(4):343–360.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of ICML*, pages 296–304.

Preslav Nakov and Marti Hearst. 2006. Using verbs to characterize noun-noun relations. In *AIMSA*, volume 4183 of *LNCS*, pages 233–244. Springer.

Preslav Nakov, Ariel Schwartz, and Marti Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of SIGIR'04 Workshop on Search and Discovery in Bioinformatics*, pages 81–88, Sheffield, UK.

Preslav Nakov. 2007. *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. Ph.D. thesis, EECS Department, University of California, Berkeley, UCB/EECS-2007-173.

Preslav Nakov. 2008. Paraphrasing verbs for noun compound interpretation. In *Proceedings of the LREC'08 Workshop: Towards a Shared Task for Multiword Expressions (MWE'08)*, Marrakech, Morocco.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 285–301, Tilburg, The Netherlands.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Martin Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of EMNLP*, pages 82–90.

Barbara Rosario, Marti Hearst, and Charles Fillmore. 2002. The descent of hierarchy, and selection in relational semantics. In *Proceedings of ACL*, pages 247–254.

Gerda Ruge. 1992. Experiment on linguistically-based term associations. *Inf. Process. Manage.*, 28(3):317–332.

Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of HLT*, pages 313–318.

Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of HLT*, pages 371–378.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259.

Peter Turney and Michael Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning Journal*, 60(1-3):251–278.

Peter Turney. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of IJCAI*, pages 1136–1141.

Peter Turney. 2006a. Expressing implicit semantic relations without supervision. In *Proceedings of ACL*, pages 313–320.

Peter Turney. 2006b. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.