Semantic Interpretation of Noun Compounds Using Verbal and Other Paraphrases

PRESLAV I. NAKOV, Qatar Computing Research Institute, Qatar Foundation
MARTI A. HEARST, University of California at Berkeley

We study the problem of semantic interpretation of noun compounds such as *bee honey*, *malaria mosquito*, *apple cake*, and *stem cell*. In particular, we explore the potential of using predicates that make explicit the hidden relation that holds between the nouns that form the noun compound. For example, *mosquito that carries malaria* is a paraphrase of the compound *malaria mosquito* in which the verb explicitly states the semantic relation between the two nouns. We study the utility of using such paraphrasing verbs, with associated weights, to build a representation of the semantics of a noun compound, e.g., *malaria mosquito* can be represented as follows: *carry* (23), *spread* (16), *cause* (12), *transmit* (9), etc. We also explore the potential of using multiple paraphrasing verbs as features for predicting abstract semantic relations such as CAUSE, and we demonstrate that using explicit paraphrases can help improve statistical machine translation.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Language parsing and understanding

General Terms: Algorithms, Languages.

Additional Key Words and Phrases: Lexical Semantics, Machine Translation, Multiword Expressions, Noun Compounds, Paraphrases, Web as a Corpus.

1. INTRODUCTION

> "*Butyrate activates the WAF1/Cip1 gene promoter through Sp1 sites in a p53-negative human colon cancer cell line*."[1]

An important characteristic of technical literature is the abundance of long sequences of nouns acting as a single noun, which are known as *noun compounds*. While eventually mastered by domain experts, noun compounds and their interpretation pose major challenges for automated analysis. For example, what is the internal syntactic structure of *human colon cancer cell line*: is it *human* [[[*colon cancer*] *cell*] *line*] or [*human* [[*colon cancer*] *cell*]] *line* or *human* [[*colon cancer*] [*cell line*]], etc.? Can *colon cancer* be paraphrased as *cancer that occurs in the colon*? Or as *cancer in the colon*? What is the relationship between *colon cancer* and *cell line*? Between *colon* and *cancer*? Is a *colon cancer cell line* a kind/type of *cell line*? Is it a kind/type of *line*?

[1]Nakaon K. et al., "Butyrate activates the WAF1/Cip1 gene promoter through Sp1 sites in a p53-negative human colon cancer cell line", *Journal of Biological Chemistry*, 272(35), pp. 22199–22206, 1997.

This work was partially supported by the National Science Foundation, under grant NSF DBI-0317510.
Author's addresses: Preslav Nakov, Qatar Computing Research Institute, Qatar Foundation, Tornado Tower, floor 10, P.O. Box 5825, Doha, Qatar; Marti Hearst, School of Information and EECS dept., Computer Science Division, 102 South Hall, Berkeley, CA 94720-4600, USA

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

Table I. Levi's recoverably deletable predicates (RDPs). Column 3 shows the modifier's function in the corresponding paraphrasing relative clause: when the modifier is the subject of that clause, the RDP is marked with the index 2.

RDP	Example	Subj/obj	Traditional Name
CAUSE$_1$	*tear gas*	object	causative
CAUSE$_2$	*drug deaths*	subject	causative
HAVE$_1$	*apple cake*	object	possessive/dative
HAVE$_2$	*lemon peel*	subject	possessive/dative
MAKE$_1$	*silkworm*	object	productive/composit.
MAKE$_2$	*snowball*	subject	productive/composit.
USE	*steam iron*	object	instrumental
BE	*soldier ant*	object	essive/appositional
IN	*field mouse*	object	locative
FOR	*horse doctor*	object	purposive/benefactive
FROM	*olive oil*	object	source/ablative
ABOUT	*price war*	object	topic

Noun compounds cannot just be ignored by natural language processing (NLP) applications since they are abundant in English written text. Baldwin and Tanaka [2004] find that 3-4% of the tokens in various corpora are part of noun compounds: 2.6% in the British National Corpus, 3.9% in the Reuters corpus.

Understanding the syntax and semantics of noun compounds is difficult but important for many natural language applications (NLP), including but not limited to question answering, machine translation, information retrieval, and information extraction. For example, a question answering system might need to determine whether *protein acting as a tumor suppressor* is a good paraphrase for *tumor suppressor protein*, and an information extraction system might need to decide whether *neck vein thrombosis* and *neck thrombosis* could possibly co-refer when used in the same document. Similarly, a machine translation system facing the unknown noun compound *WTO Geneva headquarters* might benefit from being able to paraphrase it as *Geneva headquarters of the WTO* or as *WTO headquarters located in Geneva*. Given a query like *migraine treatment*, an information retrieval system could use suitable paraphrasing verbs like *relieve* and *prevent* for page ranking and query refinement.

Below we focus on the task of finding suitable paraphrasing verbs and prepositions for a given target noun compound. We further discuss the utility of using such paraphrasing verbs as a representation of the semantics of a noun compound, or as features for predicting abstract semantic relations like LOCATION. We also demonstrate that using such explicit paraphrases can help improve statistical machine translation.

We should note that not all noun compounds are paraphrasable using verbs and prepositions only, and some cannot be paraphrased in terms of their constituent nouns at all, e.g., the semantics of *honeymoon* cannot be made explicit in terms of *honey* and *moon*. We will look into these issues in more detail in the discussion section.

2. NOUN COMPOUND SEMANTICS

Here we describe the most popular approaches to characterizing the semantics of noun compounds in theoretical and in computational linguistics.

2.1. Noun Compounds in Theoretical Linguistics

The dominant view in theoretical linguistics is that noun compound semantics can be expressed by a small set of abstract relations. For example, in the theory of Levi [1978], complex nominals – a general concept grouping the partially overlapping classes of nominal compounds (e.g., *peanut butter*), nominalizations (e.g., *dream analysis*), and non-predicate noun phrases (e.g., *electric shock*) – can be derived by two processes:

Table II. Levi's nominalization types with examples.

	Subjective	**Objective**	**Multi-modifier**
Act	*parental refusal*	*dream analysis*	*city land acquisition*
Product	*clerical errors*	*musical critique*	*student course ratings*
Agent	—	*city planner*	—
Patient	*student inventions*	—	—

(1) **Predicate Deletion** deletes the 12 abstract recoverably deletable predicates (RDPs) shown in Table I, e.g., *pie made of apples* → *apple pie*. 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;
(2) **Predicate Nominalization** produces nominals whose head is a nominalized verb, and whose modifier is derived from either the subject or the object of the underlying predicate, e.g., *the President refused general MacArthur's request* → *presidential refusal*. Multi-modifier nominalizations retaining both the subject and the object as modifiers are possible as well. Therefore, there are three types of nominalizations depending on the modifier, which are combined with four types of nominalizations the head can represent: *act*, *product*, *agent* and *patient*. See Table II for examples.

In the alternative linguistic theory of Warren [1978], noun compounds are organized into a four-level hierarchy, where the top level is occupied by the following six major semantic relations: `Possession`, `Location`, `Purpose`, `Activity-Actor`, `Resemblance`, and `Constitute`. `Constitute` is further sub-divided into finer-grained level-2 relations: `Source-Result`, `Result-Source` or `Copula`. Furthermore, `Copula` is sub-divided into the level-3 relations `Adjective-Like_Modifier`, `Subsumptive`, and `Attributive`. Finally, `Attributive` is divided into the level-4 relations `Animate_Head` (e.g., *girl friend*) and `Inanimate_Head` (e.g., *house boat*).

2.2. Noun Compounds in Computational Linguistics

In computational linguistics, most work on noun compound interpretation has focused on two-word noun compounds, or noun-noun compounds, which are the most common type of noun compounds. There have been two general lines of research: the first one derives the noun compound semantics from the semantics of the nouns it is made of [Rosario et al. 2002; Moldovan et al. 2004; Kim and Baldwin 2005; Ó Séaghdha 2009; Tratz and Hovy 2010], while the second one models the relationship between the nouns directly [Vanderwende 1994; Lapata 2002; Kim and Baldwin 2006; Nakov and Hearst 2006; 2008; Butnariu and Veale 2008].

In either case, the semantics of a noun compound is typically expressed by an abstract relation like `CAUSE` (e.g., *malaria mosquito*), `SOURCE` (e.g., *olive oil*), or `PURPOSE` (e.g., *migraine drug*), coming from a small fixed inventory. Some researchers, following the second line of research, however, have argued for a more fine-grained inventory [Downing 1977; Finin 1980]. Verbs are particularly useful in this respect and can capture elements of the semantics that the abstract relations cannot. For example, while most noun compounds expressing MAKE can be paraphrased by common patterns like *be made of* and *be composed of*, some compounds allow more specific patterns, e.g., *be squeezed from* for *orange juice*, and *be topped with* for *bacon pizza*.

Recently, the idea of using fine-grained paraphrasing verbs for noun compound interpretation has been gaining popularity [Butnariu and Veale 2008; Nakov 2008c]; there has also been a related shared task at SemEval-2010 [Butnariu et al. 2010] and at SemEval-2013 [Hendrickx et al. 2013]. This interest is partly driven by practicality: verbs are directly usable as paraphrases. Still, abstract relations remain dominant since they offer a more natural generalization.

In fact, this interest in using verbs to help the semantic interpretation of noun compounds represents a resurgence of an old idea. Many algorithms that perform semantic interpretation have placed heavy reliance on the appearance of verbs, since they are the predicates that act as the backbone of the assertion being made. Noun compounds are terse elisions of the predicate; their structure assumes that the reader knows enough about the nouns forming the noun compound and about the world at large to be able to infer what the relationship between the words is.

What is relatively new is the recent interest in trying to uncover the relationship between the two nouns by, in essence, rewriting or paraphrasing the noun compound in such a way as to be able to determine the predicate(s) holding between the nouns it is made of. Therefore, noun compound semantics is represented in terms of verbs, rather than a fixed number of abstract predicates [Levi 1978] (e.g., HAVE, MAKE, USE), relations [Girju et al. 2005] (e.g., LOCATION, INSTRUMENT, AGENT), or prepositions [Lauer 1995] (e.g., OF, FOR, IN), as is traditional in the literature. The idea is similar to the approach in [Finin 1980], who characterizes the implicit relation between the nouns forming a noun compound using an inventory of all possible verbs that can link the noun, e.g., *salt water* is interpreted using relations like *dissolved in*. See [Kim and Nakov 2011] for a further discussion on the relationship between using abstract relations and paraphrases for semantic interpretation of noun compounds.

Below we briefly describe some of the most popular abstract relation inventories that have been used for semantic interpretation of noun compounds in the NLP literature.

Vanderwende [1994] classified noun-noun compounds according to questions (mostly *Wh-questions*), which in turn correspond to 13 semantic relations:

> Subject (*Who/what?*), Object (*Whom/what?*), Locative (*Where?*), Time (*When?*), Possessive (*Whose?*), Whole-Part (*What is it part of?*), Part-Whole (*What are its parts?*), Equative (*What kind of?*), Instrument (*How?*), Purpose (*What for?*), Material (*Made of what?*), Causes (*What does it cause?*), Caused-by (*What causes it?*).

For example, *alligator shoes* is Material and answers the question *Made of what?*.

Barker and Szpakowicz [1998] extended the above relations to 20 relations:

> Agent, Beneficiary, Cause, Container, Content, Destination, Equative, Instrument, Located, Location, Material, Object, Possessor, Product, Property, Purpose, Result, Source, Time, Topic.

For example, the Possessor relation is defined as "the modifier has a head noun", e.g., *company car*.

Nastase and Szpakowicz [2003] further extended this latter inventory to 30 fine-grained elations, which they grouped into five coarse-grained super-relations (the corresponding fine-grained relations are shown in parentheses):

> CAUSALITY: cause, effect, detraction, purpose;
> PARTICIPANT: agent, beneficiary, instrument, object_property, object, part, possessor, property, product, source, whole, stative;
> QUALITY: container, content, equative, material, measure, topic, type;
> SPATIAL: direction, location_at, location_from, location;
> TEMPORALITY: frequency, time_at, time_through.

For example, *exam anxiety* is classified as effect and therefore also as CAUSALITY, and *blue book* is property and therefore also PARTICIPANT. This inventory is the most popular inventory of abstract semantic relations for noun compound interpretation. It is also a superset of the 19-relation noun compound inventory proposed by Kim and Baldwin [2006].

Similarly, **Girju et al. [2005]** proposed an inventory of 21 abstract relations

> Possession, Attribute-Holder, Agent, Temporal, Part-Whole, Is-a, Cause, Make/Produce, Instrument, Location/Space, Purpose, Source, Topic, Manner, Means, Theme, Accompaniment, Experiencer, Recipient, Measure, Result.

This inventory is a subset of a larger inventory proposed by Moldovan et al. [2004], which consists of 35 abstract semantic relations and was originally designed for the semantic interpretation of noun phrases in general.

Ó Séaghdha [2007] proposed a small inventory of just six semantic relations: BE, HAVE, IN, ACTOR, INST, ABOUT. While, on the surface, his core inventory is in the style of Levi [1978], each relation is further subdivided into sub-categories:

> BE: identity, substance-form, similarity;
> HAVE: possession, condition-experiencer, property-object, part-whole, group-member;
> IN: spatially located object, spatially located event, temporarily located object, temporarily located event;
> ACTOR: participant-event, participant-participant;
> INST: participant-event, participant-participant;
> ABOUT: topic-object, topic-collection, focus-mental activity, commodity-charge.

For example, *tax law* is TOPIC-OBJECT, *crime investigation* is FOCUS-MENTAL ACTIVITY, and they both are also ABOUT.

Tratz and Hovy [2010] proposed a large inventory of 43 semantic relations, organized in ten groups:

> CAUSAL GROUP: communicator of communication, performer of act/activity, creator/provider/cause of;
> PURPOSE/ACTIVITY GROUP: perform/engage in, create/provide/sell, obtain/access/seek, modify/process/change, mitigate/oppose/destroy, organize/supervise/authority, propel, protect/conserve, transport/transfer/trade, traverse/visit;
> OWNERSHIP, EXPERIENCE, EMPLOYMENT, AND USE GROUP: possessor + owned/possessed, experiencer + coginition/mental, employer + employee/volunteer, consumer + consumed, user/recipient + used/received, owned/possessed + possession, experience + experiencer, thing consumed + consumer, thing/means used + user;
> TEMPORAL GROUP: time [span] + X, X + time [span];
> LOCATION AND WHOLE+PART/MEMBER OF GROUP: location/geographic scope of X, whole + part/member of;
> COMPOSITION AND CONTAINMENT GROUP: substance/material/ingredient + whole, part/member + collection/config/series, X + spatial container/location/bounds;
> TOPIC GROUP: topic of communication/imagery/info, topic of plan/deal/arrangement/rules, topic of observation/study/evaluation, topic of cognition/emotion, topic of expert, topic of situation, topic of event/process;
> ATTRIBUTE GROUP: topic/thing + attribute, topic/thing + attribute value characteristic of;
> ATTRIBUTIVE AND COREFERENTIAL GROUP: coreferential, partial attribute transfer, measure + whole;
> OTHER GROUP: highly lexicalized / fixed pair, other.

While all the above relation inventories were intended to be suitable for broad-coverage text analysis, there have been some proposals tailored to a specific domain. For example, **Rosario et al. [2002]** defined 38 abstract relations for biomedical noun-noun compounds. However, due to data sparseness, only the following 18 of them were used for actual noun-noun compound interpretation experiments:

> Subtype, Activity/Physical_process, Produce_genetically, Cause, Characteristic, Defect, Person_Afflicted, Attribute_of_Clinical_Study, Procedure, Frequency/time_of, Measure_of, Instrument, Object, Purpose, Topic, Location, Material, and Defect_in_location.

Lauer [1995] proposed an unorthodox inventory, and defined the problem of noun compound interpretation as predicting which among eight prepositions paraphrases the noun compound best: of, for, in, at, on, from, with, about. For example, *olive oil* is *oil from olives*, *odor spray* is a *spray for odor*, and *night flight* is a *flight at night*.

Lauer's inventory is attractive since its paraphrases are directly usable by NLP applications. Moreover, it allows simple unsupervised interpretation since noun-preposition co-occurrences are easy to extract from a text corpus. However, it is also problematic since many noun compounds cannot be paraphrased adequately with prepositions, e.g., *woman professor* or *honey bee*. Moreover, prepositions do not align well with semantic relations, e.g., while both *morning milk* and *town house* are paraphrased using in, they express TIME and LOCATION, respectively. In fact, the correspondence is many-to-many since in, on, and at can all refer to both TIME and LOCATION.

Using abstract relations like CAUSE is problematic as well. First, it is unclear which relation inventory is best. Second, such relations capture only part of the semantics, e.g., classifying *malaria mosquito* as CAUSE obscures the fact that mosquitos do not directly cause malaria, but just transmit it. Third, in many cases, multiple relations are possible, e.g., in Levi's theory, *sand dune* is interpretable as both HAVE and BE.

Some of these issues are addressed by Finin [1980], who proposed to use a specific verb, e.g., *salt water* is interpreted as *dissolved in*. In a number of publications [Nakov and Hearst 2006; Nakov 2007; Nakov and Hearst 2008], we introduced and advocated an extension of this idea, where the set of all possible paraphrasing verbs, with associated weights, e.g., *malaria mosquito* can be found in text paraphrased as "mosquitos carry malaria" and "mosquitos spread malaria" and "mosquitos cause malaria"; the corresponding frequency of these paraphrases might be *carry (23)*, *spread (16)*, *cause (12)*, *transmit (9)*, etc. These verbs are fine-grained, directly usable as paraphrases, and using multiple of them for a noun compound approximates its semantics better.

The idea that the semantics of most noun-noun compounds can be made explicit using paraphrases that involve a verb and/or a preposition recently gained popularity: it was the focus of SemEval-2010 task 9 [Butnariu et al. 2009; 2010]. The task assumed that candidate paraphrasing verbs and prepositions had already been identified by some hypothetical system, and asked the task participants to rank a long list of such candidates for a given noun-noun compound by relevance in decreasing order. For example, *cause* and *spread* are good paraphrases for *malaria mosquito*, and thus should be ranked high, while *be made up of* is bad and thus should be ranked low.

Following this line of research, below we describe the process of extracting such paraphrasing verbs automatically. We further build a lexicon of human-proposed paraphrasing verbs, and perform a number of experiments in assessing both the lexicon's quality and the feasibility of the idea of using paraphrasing verbs to characterize noun compounds semantics. Furthermore, we compare these paraphrasing verbs to various coarse-grained inventories proposed in the literature, as well as to human judgments. Finally, we show how these paraphrases can help improve machine translation.

3. USING VERBS TO CHARACTERIZE NOUN-NOUN RELATIONS

As we mentioned above, traditionally the semantics of a noun compound has been represented as an abstract relation drawn from a small closed inventory. We have mentioned that this is problematic because (1) it is unclear which inventory is best, and mapping between different inventories has proven challenging [Girju et al. 2005], (2) abstract relations capture only part of the semantics, (3) often multiple meanings are possible, and (4) sometimes none of the pre-defined meanings is suitable for a given example. Moreover, it is unclear how useful the proposed abstract inventories are, since researchers have often fallen short of demonstrating practical uses.

We believe that verbs have more expressive power and are better tailored for the task of semantic representation: (1) they are one of the most frequent open-class parts of speech in English and thus there is a very large number of them, and (2) they can capture fine-grained aspects of the meaning. For example, while *wrinkle treatment* and *migraine treatment* express `TREATMENT-FOR-DISEASE`, some fine-grained differences can be shown by specific verbs, e.g., *smooth* can paraphrase the former, but not the latter.

In many theories, verbs play an important role in the process of noun compound derivation [Levi 1978], and speakers frequently use verbs whenever it is necessary to make overt the hidden relation between the nouns in a noun-noun compound. This allows for simple extraction, but also for straightforward uses, of verbs and paraphrases in NLP tasks such as machine translation, information retrieval, etc.

We further believe that a single verb often is not enough and that the noun compound semantics is approximated better by a collection of verbs. For example, while *malaria mosquito* can very well be characterized as `CAUSE` (or *cause*), further aspects of the meaning, can be captured by adding some additional verbs, e.g., *carry*, *spread*, *be responsible for*, *be infected with*, *transmit*, *pass on*, etc.

In the following section, we describe our algorithm for discovering predicate relations that hold between the nouns in a noun-noun compound.

4. METHOD

In a typical noun-noun compound $noun_1$ $noun_2$, $noun_2$ is the head and $noun_1$ is a modifier, attributing a property to it. The main idea of the proposed method is to preserve the head-modifier relation by substituting the pre-modifier $noun_1$ with a suitable post-modifying relative clause, e.g., *tear gas* can be transformed into *gas that causes tears*, *gas that brings tears*, *gas which produces tears*, etc.

Using all possible inflections of $noun_1$ and $noun_2$ from WordNet [Fellbaum 1998], we issue exact phrase queries of the following type against a search engine:

```
"noun2 THAT * noun1"
```

where `THAT` is one of the following complementizers: *that*, *which*, or *who*. The search engine's $*$ operator stands for a wildcard substitution (we use up to 8 stars). The quotes are the search engine's way to specify that exact phrase matches are required.

We then collect the text snippets (summaries) from the search results pages (up to 1,000 per query) and we only keep the ones for which the sequence of words following `noun1` is non-empty and contains at least one non-noun, thus ensuring the snippet includes the entire noun phrase. In order to help part-of-speech tagging and shallow parsing the snippet, we further substitute the part before `noun2` by the fixed phrase "*We look at the*" (we add this phrase to the resulting snippet, not to the query). We then perform POS tagging [Toutanova and Manning 2000] and shallow parsing[2], and we extract all verb forms, and the following preposition, if any, between `THAT` and `noun1`.

[2]Using the OpenNLP tools: `http://opennlp.sourceforge.net`

We allow for adjectives and participles to fall between the verb and the preposition but not nouns; we ignore modal verbs and auxiliaries, but we retain the passive *be*, and we require exactly one verb phrase (thus disallowing complex paraphrases like *gas that makes the eyes fill with tears*). Finally, we lemmatize the main verb using WordNet.

The proposed method is similar to previous paraphrase acquisition approaches which look for similar endpoints and collect the intervening material. For example, Lin and Pantel [2001] extract paraphrases from dependency tree paths whose end points contain similar sets of words by generalizing over these ends, e.g., for "*X solves Y*", they extract paraphrases like "*X resolves Y*", "*Y is resolved by X*", "*X finds a solution to Y*", "*X tries to solve Y*", etc. The idea is extended by Shinyama et al. [2002], who use named entities of matching semantic classes as anchors, e.g., `LOCATION`, `ORGANIZATION`, etc. It 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 to be extracted cite the same target paper; this restriction yields higher accuracy. Unlike these approaches, whose goal is to create summarizing paraphrases, we look for verbs that can characterize noun compound semantics.

5. SEMANTIC INTERPRETATION

5.1. Verb-Based Vector-Space Model

As an illustration of the method, consider the paraphrasing verbs (the corresponding frequencies are shown in parentheses) extracted from 1000 search snippets for *cancer physician* and *cancer doctor*. Note the high proportion of shared verbs (underlined):

cancer doctor: *specialize in*(12), *treat*(12), *deal with*(6), *believe*(5), *cure*(4), *attack*(4), *get*(4), *understand*(3), *find*(2), *miss*(2), *remove*(2), *study*(2), *know about*(2), *suspect*(2), *use*(2), *fight*(2), *deal*(2), *have*(1), *suggest*(1), *track*(1), *diagnose*(1), *recover from*(1), *specialize*(1), *rule out*(1), *meet*(1), *be afflicted with*(1), *study*(1), *look for*(1), *die from*(1), *cut*(1), *mention*(1), *cure*(1), *die of*(1), *say*(1), *develop*(1), *contract*(1).

cancer physician: *specialize in*(11), *treat*(7), *have*(5), *diagnose*(4), *deal with*(4), *screen for*(4), *take out*(2), *cure*(2), *die from*(2), *experience*(2), *believe*(2), *include*(2), *study*(2), *misdiagnose*(1), *be treated for*(1), *work on*(1), *die of*(1), *survive*(1), *get*(1), *be mobilized against*(1), *develop*(1).

Now consider the following four different kinds of *treatments*:

cancer treatment: *prevent*(8), *treat*(6), *cause*(6), *irradiate*(4), *change*(3), *help eliminate*(3), *be*(3), *die of*(3), *eliminate*(3), *fight*(3), *have*(2), *ask for*(2), *be specific for*(2), *decrease*(2), *put*(2), *help fight*(2), *die from*(2), *keep*(2), *be for*(2), *contain*(2), *destroy*(2), *heal*(2), *attack*(2), *work against*(2), *be effective against*(2), *be allowed for*(1), *stop*(1), *work on*(1), *reverse*(1), *characterise*(1), *turn*(1), *control*(1), *see*(1), *identify*(1), *be successful against*(1), *stifle*(1), *advance*(1), *pinpoint*(1), *fight against*(1), *burrow into*(1), *eradicate*(1), *be advocated for*(1), *counteract*(1), *render*(1), *kill*(1), *go with*(1).

migraine treatment: *prevent*(5), *be given for*(3), *be*(3), *help prevent*(2), *help reduce*(2), *benefit*(2), *relieve*(1).

wrinkle treatment: *reduce*(5), *improve*(4), *make*(4), *smooth*(3), *remove*(3), *be on*(3), *tackle*(3), *work perfect on*(3), *help smooth*(2), *be super on*(2), *help reduce*(2), *fight*(2), *target*(2), *contrast*(2), *smooth out*(2), *combat*(1), *correct*(1), *soften*(1), *reverse*(1), *resist*(1), *address*(1), *eliminate*(1), *be*(1).

herb treatment: *contain*(19), *use*(8), *be concentrated with*(6), *consist of*(4), *be composed of*(3), *include*(3), *range from*(2), *incorporate*(1), *feature*(1), *combine*(1), *utilize*(1).

Table III. Componential analysis for *man*, *woman*, *boy*, and *bull*. The components are predefined.

	man	**woman**	**boy**	**bull**
ANIMATE	+	+	+	+
HUMAN	+	+	+	−
MALE	+	−	+	+
ADULT	+	+	−	+

Table IV shows a subset of these verbs found using the above extraction method for *cancer treatment*, *migraine treatment*, *wrinkle treatment* and *herb treatment*. As expected, *herb treatment*, which is quite different from the other compounds, shares no verbs with them: it *uses* and *contains* herb, but does not *treat* it. Moreover, while migraine and wrinkles cannot be *cured*, they can be *reduced*. Migraines can also be *prevented*, and wrinkles can be *smoothed*. Of course, these results are merely suggestive and should not be taken as ground truth, especially the absence of a verb. Still, they seem to capture interesting fine-grained semantic distinctions, which normally require deep knowledge of the semantics of the two nouns that form the noun compound and/or about the world in general.

The above examples suggest that paraphrasing verbs, and the corresponding frequencies, may be a good semantic representation from a computational linguistics point of view, e.g., they can be used in a vector space model in order to measure semantic similarity between noun-noun compounds.

We believe the paraphrasing verbs can be useful from a theoretical linguistics viewpoint as well (e.g., lexical semantics); we explore this idea below.

5.2. Componential Analysis

In lexical semantics, componential analysis is often used to represent the meaning of a word in terms of semantic primitives (features), thus reducing the word's meaning to series of binary components [Katz and Fodor 1963; Jackendoff 1983; Saeed 2003]. For example, *bachelor* is (i) human, (ii) male, and (iii) unmarried, which can be expressed as [+HUMAN] [+MALE] [−MARRIED]. Similarly, *boy* can be analyzed as [+ANIMATE] [+HUMAN] [+MALE] [−ADULT], while *man* is analyzable as [+ANIMATE] [+HUMAN] [+MALE] [+ADULT]. See Table III for more examples.

Componential analysis has been very successful in phonology, where the sound system is limited and the contrast between different sounds is very important. For example, /p/ is distinguished from /b/ by the role of the vocal chords, and this distinction can be represented as a feature, e.g., /p/ is [−VOICED], while /b/ is [+VOICED].

In lexical semantics, componential analysis is considered useful for making explicit important semantic relations such as hyponymy and incompatibility, but it has been criticized for the following reasons: (1) it is unclear how the analysis decides on the particular features/components to include, and (2) it cannot really capture the full meaning of a given word.

5.3. Dynamic Componential Analysis

Given the similarity between Tables III and IV, we propose to analyze the semantics of the relations that hold between the nouns in a noun-noun compound using a kind of componential analysis, which we call *dynamic componential analysis*. The components of the proposed model are paraphrasing verbs acquired dynamically from the Web in a principled manner; this addresses the major objection against the classic componential analysis, namely that it is inherently subjective.

Table IV. Dynamic componential analysis for different kinds of treatments.

	cancer treatment	**migraine treatment**	**wrinkle treatment**	**herb treatment**
treat	+	+	+	−
prevent	+	+	−	−
cure	+	−	−	−
reduce	−	+	+	−
smooth	−	−	+	−
cause	+	−	−	−
contain	−	−	−	+
use	−	−	−	+

6. COMPARISON TO ABSTRACT RELATIONS IN THE LITERATURE

We test the paraphrasing approach using noun compound examples from the literature: we extract corresponding verbal paraphrases for them, and we manually determine whether these verbs accurately reflect the expected abstract semantic relations.

6.1. Comparison to Girju *et al.* [2005]

First, we study how our paraphrasing verbs relate to the abstract semantic relations proposed by Girju *et al.* [2005] for the semantic classification of noun compounds. For this purpose, we try to paraphrase the 21 example noun-noun compounds provided in that article as illustrations of the 21 abstract relations.

Table V shows the target semantic relation, an example noun compound from that relation, and the top paraphrasing verbs, optionally followed by prepositions, that we generated for that example. The verbs expressing the target relation are in bold, those that are good paraphrases for the noun compound but not for the abstract relation are in italic, and the erroneous extractions are small. Frequencies are in parentheses.

We were able to extract paraphrases for 20 of these 21 examples: we could not extract any paraphrasing verbs for *girl mouth* (PART-WHOLE), querying for which has returned primarily pornographic texts, which did not match our predefined patterns.

Overall, the extracted verbs provide a good characterization of the noun compounds. While in one case the most frequent verb is the copula (*to be*), the following most frequent verbs are quite adequate. In the case of *malaria mosquito*, one can argue that the CAUSE relation assigned by Girju et al. [2005] is not entirely correct since the disease is only indirectly caused by the mosquitos (it is rather carried by them), and the proposed most frequent verbs *carry* and *spread* actually support a different abstract relation: AGENT. Still, *cause* appears as the third most frequent verb, indicating that it is common to consider indirect causation as a causal relation. In the case of *combustion gas*, the most frequent verb *support* is a good paraphrase of the noun compound, but is not directly applicable to the RESULT relation assigned by Girju et al. [2005]; however, the remaining verbs for that relation do support RESULT.

For the remaining noun-noun compounds, the most frequent verbs accurately capture the relation assigned by Girju *et al.* [2005]; in some cases, the less frequent verbs indicate other logical entailments for the noun combination.

6.2. Comparison to Barker & Szpakowicz [1998]

Table VI compares our paraphrasing verbs and the first 8 (out of 20) abstract relations from [Barker and Szpakowicz 1998]: the paper gives several examples per relation, and we show the results for each of them, omitting *charitable donation* (BENEFICIARY) and *overdue fine* (CAUSE) since the modifier in these cases is an adjective[3], and *composer arranger* (EQUATIVE), for which we could not extract suitable paraphrases.

[3]Barker and Szpakowicz [1998] allow for the modifier to be either a noun or an adjective.

Table V. Comparison to [Girju et al. 2005]: top paraphrasing verbs for an example from each of their 21 relations. Verbs expressing the target relation are in **bold**, those that are good paraphrases for the noun compound but not for the abstract relation are in *italics*, and errors are small.

Sem. Relation	Example	Extracted Verbs
POSSESSION	*family estate*	be in(29), **be held by(9)**, **be owned by(7)**
ATTRIBUTE-HOLDER	*quality sound*	*offer(19)*, **be(17)**, **have(14)**, *deliver(4)*, add(4)
AGENT	*crew investigation*	**involve(2)**, *accuse(1)*, *blame(1)*, **include(1)**
TEMPORAL	*night flight*	**arrive at(19)**, **leave at(16)**, **be at(6)**, **be conducted at(6)**, **occur at(5)**
DEPICTION-DEPICTED	*image team*	**create(12)**, fit(8), **enhance(5)**, *capture(5)*, **inspect(4)**, *have(3)*, **manage(3)**, *bring(2)*, **build (2)**, **embody(2)**, *personalize(2)*, *process(2)*,
PART-WHOLE	*girl mouth*	**–**
IS-A(HYPERNYMY)	*Dallas city*	include(9)
CAUSE	*malaria mosquito*	*carry(23)*, *spread(16)*, **cause(12)**, *transmit(9)*, *bring(7)*, *have(4)*, *be infected with(3)*, **infect with(3)**, *be needed for(3)*, **pass on(2)**, **give(2)**, **give out(2)**
MAKE/PRODUCE	*shoe factory*	**produce(28)**, **make(13)**, **manufacture(11)**
INSTRUMENT	*pump drainage*	**be controlled through(3)**, **use(2)**
LOCATION/SPACE	*Texas university*	be(5), **be in(4)**
PURPOSE	*migraine drug*	**treat(11)**, **be used for(9)**, **prevent(7)**, **work for(6)**, **stop(4)**, **help(4)**, work(4) **be prescribed for(3)**, **relieve(3)**, **block(3)**, *be for(3)*, **ward off(3)**, **end(3)**, **reduce(2)**, **cure(2)**
SOURCE	*olive oil*	**come from(13)**, **be obtained from(11)**, **be extracted from(10)**, **be made from(9)**, **be produced from(7)**, **be released from(4)**, *taste like(4)*, **be beaten from(3)**, **be produced with(3)**, **emerge from(3)**
TOPIC	*art museum*	**focus on(29)**, *display(16)*, **bring(14)**, **highlight(11)**, *house(10)*, *exhibit(9)* **demonstrate(8)**, **feature(7)**, *show(5)*, **tell about(4)**, **cover(4)**, **concentrate in(4)**
MANNER	*style performance*	**exude(11)**, **fit(5)**, **have(4)**, **mix(4)**, **capture(3)**, **constitute(3)**, **enhance(3)**, **match(3)**, **ooze(3)**
MEANS	*bus service*	**use(14)**, **operate(6)**, *include(6)*
EXPERIENCER	*disease victim*	**spread(12)**, **acquire(12)**, **suffer from(8)**, **die of(7)**, *develop(7)*, **contract(6)**, **catch(6)**, **be diagnosed with(6)**, *have(5)*, *beat(5)*, **be infected by(4)**, **survive(4)**, **die from(4)**, **get(4)**, **pass(3)**, **fall by(3)**, *transmit(3)*
RECIPIENT	*worker fatalities*	**involve(4)**, **happen to(2)**, **affect(1)** **occur to(1)**, **plague(1)**, **touch(1)**
MEASURE	*session day*	*be of(7)*, *have(5)*, *include(4)*, be after(3), *be(2)*
THEME	*car salesman*	**sell(38)**, mean inside(13), **buy(7)**, **travel by(5)**, **pay for(4)**, **deliver(3)**, **push(3)**, **demonstrate(3)**, **purr(3)**,
RESULT	*combustion gas*	*support(22)*, **result from(14)**, **be produced during(11)**,**be produced by(8)** **be formed from(8)**, **form during(8)**, **be created during(7)**, **originate from(6)**, **be generated by(6)**, **develop with(6)**

Table VI. Comparison to [Barker and Szpakowicz 1998]: top paraphrasing verbs for examples for their first 8 relations (out of 20). Verbs expressing the target relation are in **bold**, those that are good paraphrases for the noun compound but not for the abstract relation are in *italics*, and errors are small.

Sem. Relation	Examples	Extracted Verbs
AGENT	*student protest*	**be led by(6)**, **be sponsored by(6)**, **pit(4)**, *be(4)*, **be organized by(3)**, **be staged by(3)**
	band concert	*feature(17)*, *capture(10)*, *include(6)*, **be given**, **by(6)**, *play of(4)*, *involve(4)*, be than(4) **be organized by(3)**, **be by(3)**, *start with(3)*
	military assault	**be initiated by(4)**, *shatter(2)*
BENEFICIARY	*student price*	*be(14)*, mean(4), differ from(4), **be for(3)**, **be discounted for(3)**, **be affordable for(3)**
CAUSE	*exam anxiety*	*be generated during(3)*
CONTAINER	*printer tray*	**hold(12)**, *come with(9)*, *be folded(8)*, *fit under(6)*, **be folded into(4)**, **pull from(4)**, **be inserted into(4)**, *be mounted on(4)*
	flood water	*cause(24)*, *produce(9)*, *remain after(9)*, *be swept by(6)*, *create(5)*, *bring(5)*, *reinforce(5)*
	film music	*fit(16)*, **be in(13)**, **be used in(11)**, **be heard**, **in(11)**, *play throughout(9)*, *be written for(9)*
	story idea	*tell(20)*, *make(19)*, *drive(15)*, *become(13)*, *turn into(12)*, *underlie(12)*, **occur within(8)**
CONTENT	*paper tray*	**feed(6)**, *be lined with(6)*, *stand up(6)*, **hold(4)**, **contain(4)**, *catch(4)*, **overflow with(3)**
	eviction notice	*result in(10)*, *precede(3)*, *make(2)*
DESTINATION	*game bus*	*be in(6)*, **leave for(3)**, *be like(3)*, *be(3)*, *make playing(3)*, *lose(3)*
	exit route	*be indicated by(4)*, **reach(2)**, *have(1)*, *do(1)*
	entrance stairs	*look like(4)*, *stand outside(3)*, *have(3)*, *follow from(3)*, *be at(3)*, be(3), *descend from(2)*
EQUATIVE	*player coach*	*work with(42)*, *recruit(28)*, **be(19)**, *have(16)*, *know(16)*, *help(12)*, *coach(11)*, *take(11)*
INSTRUMENT	*electron microscope*	**use(27)**, *show(5)*, **work with(4)**, **utilize(4)**, **employ(4)**, *beam(3)*
	diesel engine	*be(18)*, **operate on(8)**, *look like(8)*, **use(7)**, *sound like(6)*, **run on(5)**, **be on(5)**
	laser printer	**use(20)**, *consist of(6)*, *be(5)*

We obtained very good results for AGENT and INSTRUMENT, but other relations are problematic, probably due to the varying quality of the classifications: while *printer tray* and *film music* appear to be correctly assigned to CONTAINER, *flood water* and *story idea* are quite abstract and questionable; *entrance stairs* (DESTINATION) could be equally well analyzed as LOCATION or SOURCE; and *exam anxiety* (CAUSE) could refer to TIME. Finally, although Table VI shows the verb *to be* ranked third for *player coach*, in general the EQUATIVE relation poses a problem since the copula is not very frequent in the form of paraphrase we are looking for, e.g., *coach who is a player*.

Note that the EQUATIVE relation is symmetric, and thus a paraphrase like *coach who is a player* only captures part of its semantics: a better paraphrase would be *coach who is also a player*, where the adverb *also* indicates the symmetry of the relationship. Alternatively, we could paraphrase it as *. . . who is both a player and a coach*. This calls for the need to extend our patterns to capture adverbs in special cases such as this, where they provide important additional information.

Table VII. Top paraphrasing verbs for some relations from [Rosario et al. 2002].

Categ. Pair	Examples	Extracted Verbs
A01-A07 (Body Regions - Cardiovascular System)	ankle artery foot vein forearm vein finger artery neck vein head vein leg artery thigh vein	*feed*(133), *supply*(111), *drain*(100), *be in*(44), *run*(37), *appear on*(29), *be located in*(22), *be found in*(20), *run through*(19), *be behind*(19), *run from*(18), *serve*(15), *be felt with*(14), *enter*(14), *pass through*(12), *pass by*(12), *show on*(11), *be visible on*(11), *run along*(11), *nourish*(10), *be seen on*(10), *occur on*(10), *occur in*(9), *emerge from*(9), *go into*(9), ...
A01-M01.643 (Body Regions - Disabled Persons)	arm patient eye outpatient abdomen patient	*be*(54), *lose*(40), *have*(30), *be hit in*(11), *break*(9), *gouge out*(9), *injure*(8), *receive*(7), *be stabbed in*(7), *be shot in*(7), *need*(6), ...
A01-M01.150 (Body Regions - Disabled Persons)	leg amputee arm amputee knee amputee	*lose*(13), *grow*(6), *have cut off*(4), *miss*(2), *need*(1), *receive*(1), *be born without*(1)
A01-M01.898 (Body Regions - Donors)	eye donor skin donor	*give*(4), *provide*(3), *catch*(1)
D02-E05.272 (Organic Chemicals - Diet)	choline diet methionine diet carotene diet saccharin diet	*be low in*(18), *contain*(13), *be deficient in*(11), *be high in*(7), *be rich in*(6), *be sufficient in*(6), *include*(4), *be supplemented with*(3), *be in*(3), *be enriched with*(3), *contribute*(2), *miss*(2), ...

6.3. Comparison to Rosario *et al.* [2002]

Rosario et al. [2002] characterize the semantics of biomedical noun-noun compounds based on the semantic categories of the constituent nouns in the MeSH lexical hierarchy. For example, all compounds whose first noun falls under the A01 subhierarchy (*Body Regions*), and whose second noun is under A07 (*Cardiovascular System*), e.g., *mesentery artery*, *leg vein*, *finger capillary*, are hypothesized to express the same semantic relation. If the relation is heterogeneous for some of the categories, they descend 1-2 levels down, e.g., A01-M01 (*Body Regions – Persons*) is decomposed to A1-M01.643, A1-M01.150, and A1-M01.898. They call this *the descent of hierarchy*.

We studied the ability of our method to generate verbs that can characterize the abstract semantic relation that is expressed by noun-noun compounds belonging to a particular pair of MeSH categories, e.g., A1-A7.

We first extracted a large number of noun-noun compounds from a collection of 1.4 million *MEDLINE* abstracts, which we then mapped to pairs of MeSH categories. We extracted a total of 228,702 noun-noun pairs, 40,861 of which were unique, which corresponds to 35,205 unique MeSH category pairs of various generalization levels (see [Nakov 2007] for details).

Given a category pair, such as A01-A07 and A01-M01.643, we considered all noun-noun compounds whose elements are in the corresponding MeSH sub-hierarchies, and we acquired paraphrasing verbs (+prepositions) for each of them from the Web. We then aggregated the results in order to obtain a set of characterizing paraphrasing verbs for the target category pair.

As Table VII shows, the results are quite good for A01-A07, for which we have a lot of examples, and for D02-E05.272, which seems relatively unambiguous, but they are not as good for A01-M01.*, which is both more ambiguous and has fewer examples: generalizing verbal paraphrases for a category seems to work best for categories represented by multiple relatively unambiguous examples.

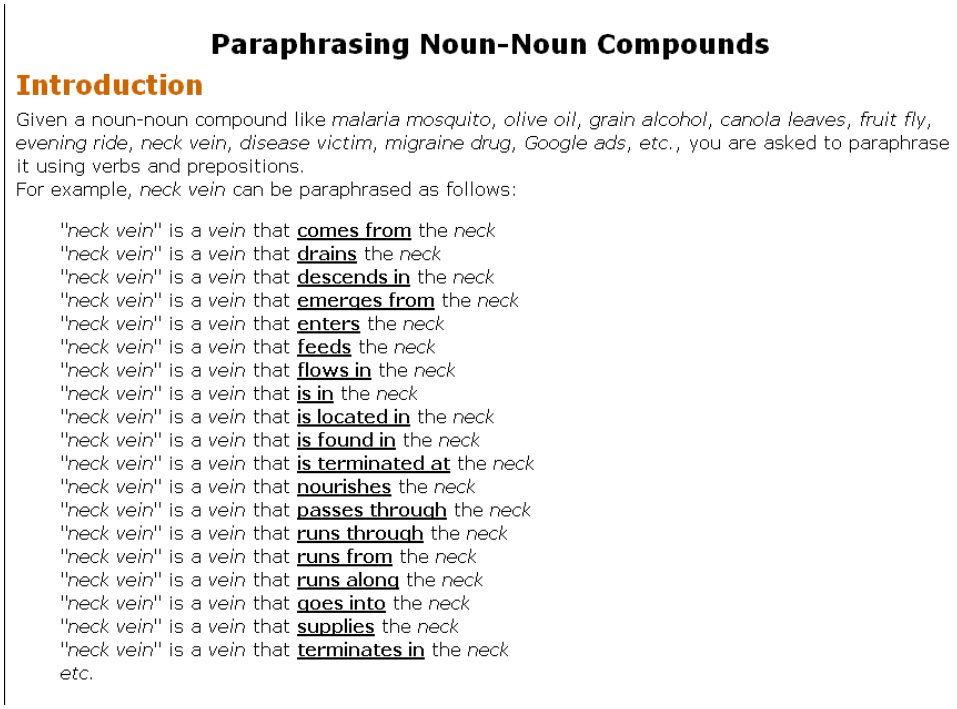

Fig. 1. The noun-noun compound paraphrasing task in Amazon's Mechanical Turk: Introduction.

7. COMPARISON TO HUMAN-GENERATED VERBS

In order to evaluate the verb-based semantic relations we obtained when applying our method, we conducted an experiment in which we gathered paraphrases for noun-noun compounds from human judges. For this purpose, we defined a special noun-noun compound paraphrasing task asking human judges to propose verbal paraphrases of the kind we generate: We asked for verbs, possibly followed by prepositions, that could be used in a paraphrase involving *that*. For example, *nourish*, *run along* and *come from* are good paraphrasing verbs for the noun-noun compound *neck vein* since they can be used in paraphrases like *a vein that nourishes the neck*, *a vein that runs along the neck*, or *a vein that comes from the neck*.

In an attempt to make the task as clear as possible and to ensure high-quality results, we provided detailed instructions, we stated explicit restrictions, and we gave several example paraphrases. We instructed the participants to propose at least three paraphrasing verbs per noun-noun compound, if possible. The instructions we provided and the actual interface the human judges used are shown in Figures 1 and 2; this is the user interface of the *Amazon Mechanical Turk* Web service.[4]

Tables VIII, IX X and XI compare human- and program-generated paraphrasing verbs for the noun-noun compounds *malaria mosquito*, *olive oil*, *disease victim* and *night flight*, respectively. The human-generated paraphrasing verbs, obtained from ten judges, are shown on the left sides of the tables, sorted by frequency in descending order, while the right sides list the program-generated verbs; the verbs appearing on both sides are underlined.

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

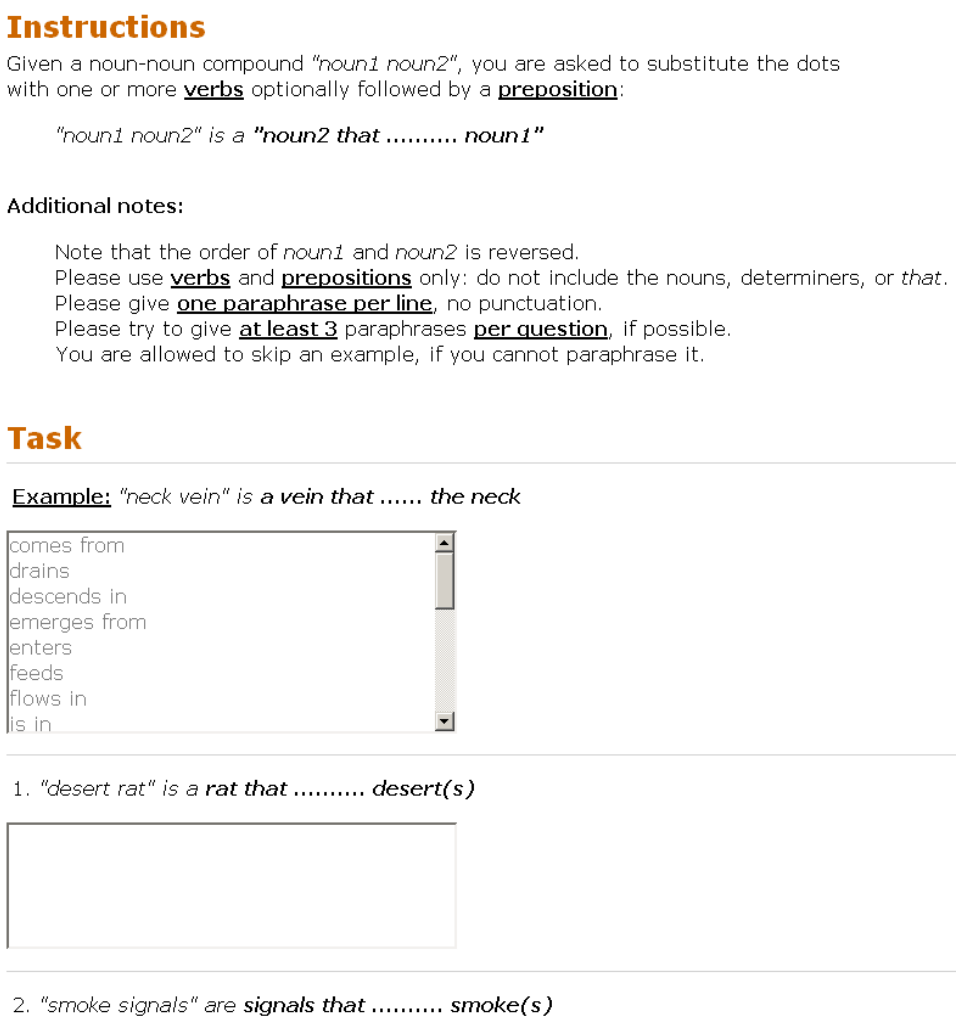

Fig. 2. The noun-noun compound paraphrasing task in Amazon's Mechanical Turk: Instructions, example, and sample questions.

We can see in these tables a sizable overlap between the human- and the program-generated paraphrasing verbs for *malaria mosquito*, which is relatively unambiguous and expresses an indirect causation, and a smaller overlap for the more ambiguous *night flight*, *disease victim*, and *olive oil*. For example, the latter can refer to multiple abstract relations, e.g., CONTAINER (*oil that is inside the olive*), SOURCE or ORIGIN (*oil that comes from olives*), PRODUCT (*oil that is produced from olives*), QUALITY (*oil that tastes like olive*), etc. Still, for all four given examples, there is a general tendency for the most frequent human-proposed and the top program-generated verbs to overlap.

We further compared the human and the program-generated paraphrases in a bigger study using the complex nominals listed in the appendix of [Levi 1978]. We had to exclude the examples with an adjectival modifier, which are allowed by Levi's theory. Moreover, some of the noun compounds were spelled as a single word, which, according to our definition of a noun compound, represents a single noun.

Table VIII. Human- and program-generated verbs for *malaria mosquito*. Verbs appearing on both sides are underlined.

#	Human Judges	#	Program
8	carries	23	carries
4	causes	16	spreads
2	transmits	12	causes
2	is infected with	9	transmits
2	infects with	7	brings
1	has	4	has
1	gives	3	is infected with
1	spreads	3	infects with
1	propagates	2	gives
1	supplies	2	is needed for

Table IX. Human- and program-generated verbs for *olive oil*. Verbs appearing on both sides are underlined: both full and partial overlaps.

#	Human Judges	#	Program
5	is pressed from	13	comes from
4	comes from	11	is obtained from
4	is made from	10	is extracted from
2	is squeezed from	9	is made from
2	is found in	7	is produced from
1	is extracted from	4	is released from
1	is in	4	tastes like
1	is produced out of	3	is beaten from
1	is derived from	3	is produced with
1	is created from	3	emerges from
1	contains		
1	is applied to		

Table X. Human- and program-generated verbs for *disease victim*. Verbs appearing on both sides are underlined: both full and partial overlaps.

#	Human Judges	#	Program
6	has	12	spreads
3	suffers from	12	acquires
3	is infected with	8	suffers from
2	dies of	7	dies of
2	exhibits	7	develops
2	carries	6	contracts
1	is diagnosed with	6	is diagnosed with
1	contracts	6	catches
1	is inflicted with	5	has
1	is ill from	5	beats
1	succumbs to	4	is infected by
1	is affected by	4	survives
1	presents	4	dies from
		4	gets
		3	passes
		3	falls by
		3	transmits

Table XI. Human- and program-generated verbs for *night flight*. Verbs appearing on both sides are underlined: both full and partial overlaps.

#	Human Judges	#	Program
5	occurs at	19	arrives at
5	is at	16	leaves at
4	happens at	6	is at
2	takes off at	6	is conducted at
1	arrives by	5	occurs at
1	travels through		
1	runs through		
1	occurs during		
1	is taken at		
1	is performed at		
1	is flown during		
1	departs at		
1	begins at		

Therefore, we had to exclude the following concatenated words that appeared in Levi's dataset: *whistleberries*, *gunboat*, *silkworm*, *cellblock*, *snowball*, *meatballs*, *windmill*, *needlework*, *textbook*, *doghouse*, and *mothballs*. Some other examples contained a modifier that is a concatenated noun compound, e.g., *wastebasket category*, *hairpin turn*, *headache pills*, *basketball season*, *testtube baby*. These examples are noun-noun compounds under our definition, and thus we retained them.

However, we found them inconsistent with the other examples in the collection from Levi's theory point of view: the dataset was supposed to contain noun-noun compounds only. Even more problematic (but not for our definition), is *beehive hairdo*, where both the modifier and the head are concatenations; we retained that example as well. As a result, we ended up with 250 good noun-noun compounds out of the original 387 complex nominals.

We randomly distributed these 250 noun-noun compounds into groups of 5 as shown in Figure 2, which yielded 50 Mechanical Turk tasks known as HITs (Human Intelligence Tasks), and we requested 25 different human judges (workers) per HIT. We had to reject some of the submissions, which were empty or did not follow the instructions, in which cases we requested additional judges in order to guarantee at least 25 good submissions per HIT. Each human subject was allowed to work on any number of HITs (between 1 and 50), but was not permitted to do the same HIT twice. A total of 174 different human judges worked on the 50 HITs, producing 19,018 different verbs. After removing the empty and the bad submissions, and after normalizing the verbs (see below), we ended up with a total of 17,821 verbs, which means 71.28 verbs per noun-noun compound on average, not necessarily distinct.

Since many judges did not strictly follow the instructions, we performed some automatic cleaning of the results, followed by a manual check and correction, when it was necessary. First, some judges included the target nouns, the complementizer *that*, or determiners like *a* and *the*, in addition to the paraphrasing verb, in which cases we removed this extra material. For example, *star shape* was paraphrased as *shape that looks like a star* or as *looks like a* instead of just *looks like*. Second, the instructions required that a paraphrase be a sequence of one or more verb forms possibly followed by a preposition (complex prepositions like *because of* were allowed), but in many cases the proposed paraphrases contained words belonging to other parts of speech, e.g., nouns (*is in the shape of*, *has responsibilities of*, *has the role of*, *makes people have*, *is part of*, *makes use of*) or predicative adjectives (*are local to*, *is full of*); we filtered out such paraphrases. In case a paraphrase contained an adverb, e.g., *occur only in*, *will eventually bring*, we removed the adverb and kept the paraphrase.

We also normalized the verbal paraphrases by removing the leading modals (e.g., *can cause* becomes *cause*), perfect tense *have* and *had* (e.g., *have joined* becomes *joined*), or progressive tense *be* (e.g., *is donating* becomes *donates*). We converted complex verbal construction of the form '<*raising verb*> *to be*' (e.g., *appear to be*, *seems to be*, *turns to be*, *happens to be*, *is expected to be*) to just *be*. We further removed present participles introduced by *by*, e.g., *are caused by peeling* becomes *are caused*.[5] We also filtered out any paraphrase that involved *to* as part of the infinitive of a verb different from *be*, e.g., *is willing to donate* or *is painted to appear like* are not allowed. We also added *be* when it was missing in passive constructions, e.g., *made from* became *be made from*. Finally, we lemmatized the conjugated verb forms using WordNet, e.g., *comes from* becomes *come from*, and *is produced from* becomes *be produced from*. We also fixed some occasional spelling errors that we noticed, e.g., *bolongs to*, *happens becasue of*, *is mmade from*.

The resulting lexicon of human-proposed paraphrasing verbs with corresponding frequencies, and some other lexicons, e.g., a lexicon of the first verbs proposed by each judge only, and a lexicon of paraphrasing verbs automatically extracted from the Web as described in [Nakov and Hearst 2008], are released under the Creative Commons License Attribution 3.0 Unported, and can be downloaded from the *Multiword Expressions Website*: `http://multiword.sf.net`. See [Nakov 2008d] for additional details.

We performed a number of experiments in order to assess both the quality of the created lexicon and the feasibility of the idea of using paraphrasing verbs to characterize noun compounds semantics.

For each noun-noun compound from the *Levi-250 dataset*, we constructed two frequency vectors $\overrightarrow{h}$ (human) and $\overrightarrow{p}$ (program). The former is composed of the above-described human-proposed verbs (after lemmatization) and their corresponding frequencies, and the latter contains verbs and frequencies that were automatically extracted from the Web, as described in [Nakov and Hearst 2008]. We then calculated the cosine similarity between $\overrightarrow{h}$ and $\overrightarrow{p}$ as follows:

$$cos(\overrightarrow{h}, \overrightarrow{p}) = \frac{\sum_{i=1}^{n} h_i p_i}{\sqrt{\sum_{i=1}^{n} h_i^2} \sqrt{\sum_{i=1}^{n} p_i^2}} \quad (1)$$

Table XII shows human- and program-proposed vectors for sample noun-noun compounds together with the corresponding cosine. The average cosine similarity (in %) for all 250 noun-noun compounds is shown in Table XIII. Since the judges were instructed to provide at least three paraphrasing verbs per noun-noun compound, and they tried to comply, some bad verbs were generated as a result. In such cases, the very first verb proposed by a judge for a given noun-noun compound is likely to be the best one. We tested this hypothesis by calculating the cosine using these first verbs only. As the last two columns of the table show, using all verbs yields consistently higher cosine similarity, which suggests that there are many additional good human-generated verbs among those that follow the first one. However, the differences are 1-2% only and are not statistically significant according to the two-tailed Pearson's χ^2 test, $p < 0.05$.

A limitation of the Web-based verb-generating method is that it could not provide paraphrasing verbs for 14 compounds (which is not a high number out of 250; it only constitutes 5.6%); in these cases, the cosine score is zero.

[5]This could cause problems in some cases, e.g., if the original was *peelings are caused by peeling potato*, dropping *by peeling* would yield *peelings are caused potato*, which is not an appropriate paraphrase.

Table XII. Human- and program-proposed vectors, and cosines for sample noun-noun compounds: the shared verbs are underlined.

0.96 "blood donor" NOMINALIZATION:AGENT
Human: give(30), donate(16), supply(8), provide(6), share(2), contribute(1), volunteer(1), offer(1), choose(1), hand over(1), . . .
Progr.: give(653), donate(395), receive(74), sell(41), provide(39), supply(17), be(13), match(11), contribute(10), offer(9), . . .

0.93 "city wall" HAVE$_2$
Human: surround(24), protect(10), enclose(8), encircle(7), encompass(3), be in(3), contain(2), snake around(1), border(1), go around(1), . . .
Progr.: surround(708), encircle(203), protect(191), divide(176), enclose(72), separate(49), ring(41), be(34), encompass(25), defend(25), . . .

0.91 "disease germ" CAUSE$_1$
Human: cause(20), spread(5), carry(4), create(4), produce(3), generate(3), start(2), promote(2), lead to(2), result in(2), . . .
Progr.: cause(919), produce(63), spread(37), carry(20), propagate(9), create(7), transmit(7), be(7), bring(5), give(4), . . .

0.89 "flu virus" CAUSE$_1$
Human: cause(19), spread(4), give(4), result in(3), create(3), infect with(3), contain(3), be(2), carry(2), induce(1), . . .
Progr.: cause(906), produce(21), give(20), differentiate(17), be(16), have(13), include(11), spread(7), mimic(7), trigger(6), . . .

0.89 "gas stove" USE
Human: use(20), run on(9), burn(8), cook with(6), utilize(4), emit(3), be heated by(2), need(2), consume(2), work with(2), . . .
Progr.: use(98), run on(36), burn(33), be(25), be heated by(10), work with(7), be used with(7), leak(6), need(6), consume(6), . . .

0.89 "collie dog" BE
Human: be(12), look like(8), resemble(2), come from(2), belong to(2), be related to(2), be called(2), be classified as(2), be made from(1), be named(1), . . .
Progr.: be(24), look like(14), resemble(8), be border(5), feature(3), come from(2), tend(2), be bearded(1), include(1), betoken(1), . . .

0.87 "music box" MAKE$_1$
Human: play(19), make(12), produce(10), emit(5), create(4), contain(4), provide(2), generate(2), give off(2), include(1), . . .
Progr.: play(104), make(34), produce(18), have(16), provide(14), be(13), contain(9), access(8), say(7), store(6), . . .

0.87 "cooking utensils" FOR
Human: be used for(17), be used in(9), facilitate(4), help(3), aid(3), be required for(2), be used during(2), be found in(2), be utilized in(2), involve(2), . . .
Progr.: be used for(43), be used in(11), make(6), be suited for(5), replace(3), be used during(2), facilitate(2), turn(2), keep(2), be for(1), . . .

Table XIII. Average cosine similarity (in %) between the human- and the program-generated verbs for 250 noun-noun compounds from [Levi 1978]. Shown are the results for different limits on the minimum number of program-generated Web verbs. The last column shows the cosine when only the first verb proposed by each judge is used.

Min # of Web Verbs	Number of Compounds	Cosine Sim. w/ Humans *All Verbs*	Cosine Sim. w/ Humans *First Only*
0	250	31.81%	30.60%
1	236	33.70%	32.41%
3	216	35.39%	34.07%
5	203	36.85%	35.60%
10	175	37.31%	35.53%

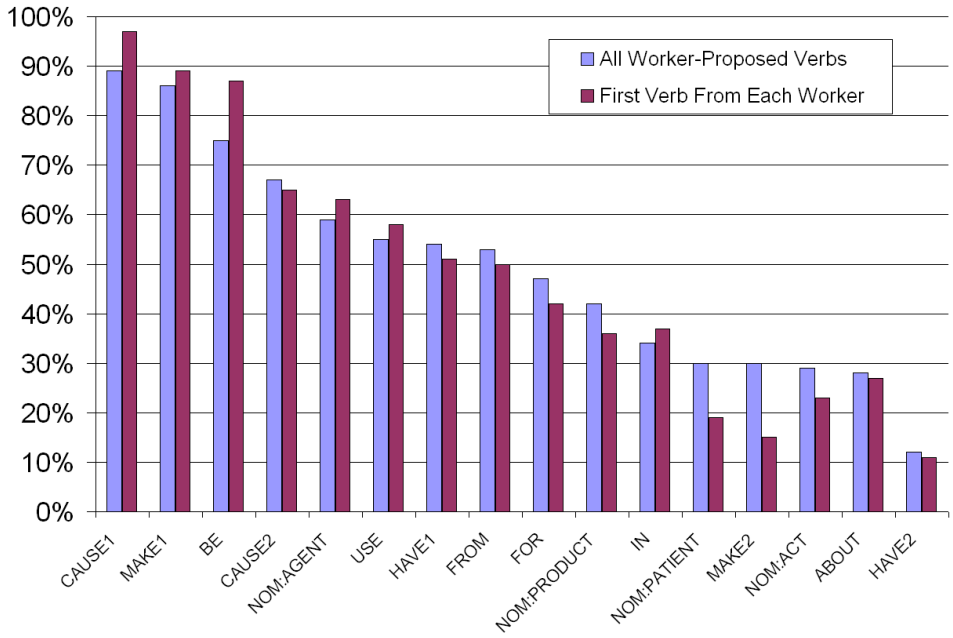

Fig. 3. Cosine similarity (in %) between human- and program- generated verbs by relation: using all human-proposed verbs vs. the first verb from each judge.

This includes the following compounds: *beard trim* (NOMINALIZATION:PRODUCT), *blanket excuse* (BE), *coffee nerves* (CAUSE_2), *communist tenet* (IN), *factory rejects* (NOMINALIZATION:PATIENT), *financing dilemma* (ABOUT), *lemon peel* (HAVE_2), *midnight snack* (IN), *morphology lecture* (ABOUT), *pedal extremities* (BE), *pork suet* (FROM), *testtube baby* (FROM), *vapor lock* (CAUSE_2), and *wastebasket category* (BE). We can see that this list includes some three-word noun compounds in disguise, *testtube baby* and *wastebasket category*, which cause data sparseness partly because of the variation in spelling: *test-tube* vs. *test tube*, and *wastebasket* vs. *waste basket*. It also includes some rare words such as *tenet* and *suet*, which are hard to find paraphrases for on the Web. Three are instances of BE, which are arguably hard to paraphrase with verbs other than *to be* (which was also not extracted in these cases); other researchers have also reported problems with using verbs to paraphrase equative or BE relations [Kim and Baldwin 2006]. Finally, *blanket excuse* is an idiomatic expression, which is not paraphrasable in the sense required here.

When the calculation is performed for the remaining 236 compounds only, the cosine increases by 2%. Table XIII shows the results when the cosine calculations are limited to compounds with at least 1, 3, 5 or 10 different verbs. We can see that the cosine similarity increases with the minimum number of required verbs, which means that the extracted verbs are generally good, and part of the low cosines are due to an insufficient number of extracted verbs.

We further compared the human- and the program-generated verbs aggregated by relation. Given a relation like HAVE_1, 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 then compared the corresponding pairs of summed vectors separately for each relation.

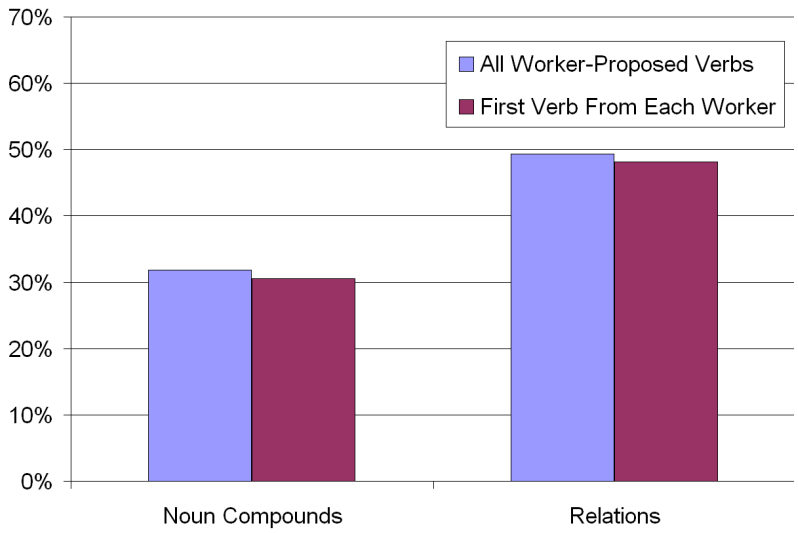

Fig. 4. Average cosine similarity (in %) between the human- and the program-generated verbs for 250 noun-noun compounds from [Levi 1978] calculated for each noun compound and aggregated by relation: using all human-proposed verbs vs. the first verb from each judge.

Figure 3 shows the cosine similarity for each of the 16 relations using all human-proposed verbs and only the first verb from each judge. We can see a very high cosine similarity (mid-70% to mid-90%) for relations like $\texttt{CAUSE}_1$, $\texttt{MAKE}_1$, BE, but low cosine similarity of 11-30% for reverse relations like $\texttt{HAVE}_2$ and $\texttt{MAKE}_2$, and for most nominalizations (except for NOM:AGENT).

Interestingly, using only the first verb improves the results for highly-correlated relations, but damages low-correlated ones. This suggests that when a relation is more homogeneous, the first verbs proposed by the judges are good enough, and the following verbs only introduce noise. However, when the relation is more heterogeneous, the extra verbs are more likely to be useful.

As Figure 4 shows, overall the average cosine similarity is slightly higher when all judge-proposed verbs are used vs. when only the first verbs from each judge are used: this is true both when comparing the individual noun-noun compounds and when the comparison is performed for the 16 relations. The figure also shows that while the cosine similarity for individual noun-noun compounds is in the low-30%, for relations it is almost 50%.

7.1. Comparison to FrameNet

The idea to approximate noun compounds semantics by a collection of verbs is related to the approach taken in the Berkeley FrameNet project[6] [Baker et al. 1998], which builds on the ideas for case frames of Fillmore [1968]. According to Fillmore [1982], p.115, frames are "*characterizing a small abstract 'scene' or 'situation', so that to understand the semantic structure of the verb it is necessary to understand the properties of such schematized scenes*". In FrameNet, a lexical item evokes a number of relations and concepts representative of the context in which the word applies; together they represent the speaker's understanding of the word.

[6]http://www.icsi.berkeley.edu/~framenet/

Table XIV. The *Causation* frame in FrameNet: comparing with the top program- and the human-generated verbs for $\texttt{CAUSE}_1$.

Frames	Verbs	Program	Humans
Causation	(be) because of		✓
Causation	bring	✓	
Causation	bring about	✓	✓
Causation	bring on		✓
Causation	induce	✓	✓
Causation	lead to		✓
Causation	make	✓	✓
Causation	mean		
Causation	precipitate		✓
Causation	put		
Causation	raise		
Causation	result in	✓	✓
Causation	render		
Causation	send		
Causation	wreak		
Overlap:		**5/15 (33.3%)**	**8/15 (53.3%)**

Table XV. The *Using* frame in FrameNet: comparing with the top program- and the human-generated verbs for USE.

Frame	Verb	Program	Humans
Using	apply	✓	✓
Using	avail oneself		
Using	employ	✓	✓
Using	operate	✓	
Using	utilise	✓	✓
Using	use	✓	✓
Overlap:		**5/6 (83.3%)**	**4/6 (66.7%)**

Table XVI. The *Possession* frame in FrameNet: comparing with the top program- and the human-generated verbs for $\texttt{HAVE}_1$.

Frames	Verbs	Program	Humans
Possession	belong		
Possession	have got		
Possession	have	✓	✓
Possession	lack	✓	
Possession	own	✓	✓
Possession	possess	✓	
Possession	want	✓	
Overlap:		**5/7 (71.4%)**	**2/7 (28.6%)**

Table XVII. The *Intentionally_create* and *Manufacturing* frames in FrameNet: comparing with the top program- and the human-generated verbs for $\texttt{MAKE}_1$.

Frames	Verbs	Program	Humans
Intentionally_create	create	✓	✓
Intentionally_create	establish		
Intentionally_create	found		
Intentionally_create	generate	✓	✓
Intentionally_create, Manufacturing	make	✓	✓
Intentionally_create, Manufacturing	produce	✓	✓
Intentionally_create	setup		
Intentionally_create	synthesise		✓
Manufacturing	fabricate		
Manufacturing	manufacture		✓
Overlap:		**4/10 (40%)**	**6/10 (60%)**

For example, the meaning of a sentence like '*Sara faxed Jeremy the invoice.*' is not derived from the meaning of the verb *fax* alone, but also from speaker's knowledge about situations where somebody gives something to somebody else [Goldberg 1995; Petruck 1996; Baker and Ruppenhofer 2002].

Our program- and human-generated paraphrasing verbs represent similar world knowledge: a dynamically constructed semantic frame in terms of which the target noun compound is to be understood. Therefore, we could expect similarities of the relations between the entities in the manually created frames of FrameNet and the ones we generate automatically: in particular, there should be some overlap between our automatically generated verbs and the ones listed in the corresponding FrameNet frame. If so, given a sentence, our verbs can help automatically select the FrameNet frame that best characterizes the situation described in that sentence.

As a preliminary investigation of the potential of these ideas, we compared the verbs we generated for four of Levi's relations $\texttt{CAUSE}_1$, USE, $\texttt{HAVE}_1$, and $\texttt{MAKE}_1$, and the verbs listed in FrameNet for the frames that we found to correspond to these relations. We also tried to compare the FrameNet verbs to the human-proposed ones. In both cases, we used the top 150 verbs for the target Levi relation.

The results are shown in Tables XIV, XV, XVI, and XVII, respectively. The results vary across relation, but the overall performance of the human- and program-proposed verbs is comparable, and the average overlap is over 50%. However, we believe that this percent is an underestimation. Most verbs that we are missing are in our view very weakly associated with the general relation expressed by the frame. For example, for the *Causation* frame (see Table XIV), both the humans and the program miss the verbs *mean*, *put*, *raise*, *render*, *send*, and *wreak*, which we believe do not express a $\texttt{CAUSE}_1$ relation in general, even though they might do so in a particular context.

Similarly, for the *Using* frame (see Table XV), we miss one verb, *avail oneself*, which does not necessarily express a USE relation; it also contains a pronoun, and thus cannot possibly have been extracted by our patterns. For the *Possession* frame (see Table XVI), we miss one verb *have got*, but we do have *have*. For the *Intentionally_create* frame (see Table XVII), we miss three verbs: *establish*, *found* and *setup*, which do not necessarily express the $\texttt{MAKE}_1$ relation, i.e., here we have only a partial overlap between the frame and our abstract relations. Finally, again in Table XVII, for the *Manufacturing* frame, we miss the verb *fabricate*, which is a legitimate verb that we have failed to extract.

8. APPLICATION TO RELATIONAL SIMILARITY

Here, we extend the above method to measuring the semantic similarity between pairs of words, i.e., to relational similarity. This is an important but understudied problem. Despite the tremendous amount of computational linguistics publications on word similarity (see [Budanitsky and Hirst 2006] for an overview), there is surprisingly little work on relational similarity. Students taking the SAT examination are familiar with verbal analogy questions, where they need to decide whether, e.g., the relation between *ostrich* and *bird* is more similar to the one between *lion* and *cat*, or rather between *primate* and *monkey*. These kinds of questions are hard; the average test taker achieves about 56.8% on the average [Turney and Littman 2005].

Given a pair of words, we mine the Web for sentences containing these words and then we extract verbs, prepositions, and coordinating conjunctions that connect them; we then use these lexical features in instance-based classifiers. We apply the approach to several relational similarity problems, including solving SAT verbal analogy (which is the benchmark problem for relational similarity), classifying head-modifier relations, and extracting relations between complex nominals and between noun compounds without context, as well as between pairs of nominals in a sentential context. We report results from queries executed in the period 2005-2008.

8.1. Method

8.1.1. Feature Extraction. Given a pair of nouns $noun_1$ and $noun_2$, we mine the Web for sentences containing them and we extract connecting verbs, prepositions, and coordinating conjunctions, which we will later use as lexical features in a vector-space model to measure semantic similarity between pairs of nouns.

The extraction process starts with a set of exact phrase queries generated using the following patterns:

"$infl_1$ THAT * $infl_2$"
"$infl_2$ THAT * $infl_1$"
"$infl_1$ * $infl_2$"
"$infl_2$ * $infl_1$"

where:

$infl_1$ and $infl_2$ are inflected variants of $noun_1$ and $noun_2$;
THAT can be *that*, *which*, or *who*;
and * stands for 0 or more (up to 8) stars, representing the * operator.

Note that, unlike before, we issue queries not only with $noun_2$ preceding $noun_1$, but also with $noun_1$ preceding $noun_2$; moreover, in addition to queries containing THAT, we also issue queries without THAT. This yields four general query patterns instead of just one, thus increasing the number of snippets that could be extracted and processed, and ultimately yielding more potential features.

For each query, we collect the text snippets (summaries) from the result set (up to 1,000 per query) and we split them into sentences. We then filter out the incomplete sentences and the ones that do not contain the target nouns, as well as duplicates. We POS-tag the sentences using the OpenNLP tagger, and we extract three features:

Verbs: We extract a verb, if the subject NP of that verb is headed by one of the target nouns (or an inflected form of a target noun), 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]. 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 preposition heading that PP, e.g., "The thesis advisory *committee* consists of three qualified *members*." We also extract verb+preposition from relative phrases, we include particles, we ignore modals and auxiliaries, and we lemmatize the verbs.

Prepositions: If one of the target nouns is the head of an NP that contains a PP inside which there is an NP headed by the other target noun, we extract the preposition heading that PP, e.g., "The *members* of the *committee* held a meeting."

Coordinating conjunctions: 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 XVIII for examples.

After having extracted the linguistic features, we use them in instance-based classifiers based on the vector-space model. This vector representation is similar to previous approaches, e.g., [Alshawi and Carter 1994; Grishman and Sterling 1994; Ruge 1992; Lin 1998].

Table XVIII. 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.

Frequency	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	$2 \rightarrow 1$
45	rotate off	V	$2 \rightarrow 1$
44	be signed up for	V	$2 \rightarrow 1$
43	notify	V	$1 \rightarrow 2$
40	provide that	V	$2 \rightarrow 1$
39	need	V	$1 \rightarrow 2$
37	stand	V	$2 \rightarrow 1$
36	be	V	$2 \rightarrow 1$
36	vote	V	$1 \rightarrow 2$
36	participate in	V	$2 \rightarrow 1$
35	allow	V	$1 \rightarrow 2$
33	advise	V	$2 \rightarrow 1$
32	inform	V	$1 \rightarrow 2$
31	form	V	$2 \rightarrow 1$
...	...	...	...

For example, Lin [1998] measures *word* similarity using triples extracted from a dependency parser. In particular, given a noun, he finds all verbs that have it as a subject or an object, and all adjectives that modify it, together with frequencies. In contrast, here we are interested in relational rather than word similarity.

Our method is also related to the idea of Devereux and Costello [2006] that the meaning of a noun-noun compound can be characterized by a distribution over several dimensions, as opposed to being expressed by a single relation. However, their relations are fixed and abstract, while ours are dynamic and based on verbs, prepositions and coordinating conjunctions.

It is also similar to the work of Kim and Baldwin [2006], who characterize the relation using verbs. However they use a fixed set of seed verbs for each relation; they also use the grammatical roles of the noun-noun compound's head and modifier. In contrast, our verbs, prepositions and coordinating conjunctions are dynamically extracted, and we do not try to generalize the head and the modifier.

Our approach is also similar to that of Séaghdha and Copestake [2007], who use grammatical relations as features to characterize a noun-noun compound; however, we use verbs, prepositions and coordinating conjunctions instead.

Table XIX. Example of an SAT verbal analogy question. The stem is in **bold**, the correct answer is in *italics*, and the distractors are in plain text.

	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

8.1.2. Similarity Measure. The above-described features are used in the calculation of the similarity between noun pairs. We use TF.IDF-weighting in order to downweight very common features like *of*:

$$w(x) = TF(x) \times \log\left(\frac{N}{DF(x)}\right) \tag{2}$$

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 this feature, and N is the total number of training noun pairs.

We use these weights in a variant of the Dice coefficient. The classic Dice coefficient for two sets A and B is defined as follows:

$$Dice(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{3}$$

This definition applies to Boolean vectors as well since they are equivalent to discrete sets, but it does not apply to numerical vectors in general. Therefore, we use the following generalized definition:[7]

$$Dice(A, B) = \frac{2 \times \sum_{i=1}^{n} \min(a_i, b_i)}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i} \tag{4}$$

A bigger value for the Dice coefficient indicates higher similarity. Therefore, we take $\min(a_i, b_i)$ in order to avoid giving unbalanced weight to a feature when the weights are lopsided. For example, if the noun pair (*committee*, *members*) has the feature *include* as a verb in the forward direction 1,000 times, and the noun pair (*ant*, *hill*) has it only twice, we do not want to give a lot of weight for overlapping on that feature.

8.2. Solving SAT Verbal Analogy Problems

Following Turney [2006], we use *SAT verbal analogy* as a benchmark problem. We experiment with *Turney's dataset*, which consists of 374 SAT questions from various sources, including 190 from actual SAT tests, 80 from SAT guidebooks, 14 from the ETS Web site, and 90 from other SAT test preparation Web sites. Table XIX shows two example problems: the top pairs are called *stems*, the ones in italics are the *solutions*, and the remaining ones are *distractors*. Turney [2006] achieved 56% accuracy, on the full set of 374 analogy problems, which matches the 56.8% average human performance, and is a significant improvement over the 20% random-guessing baseline (there are five candidate pairs to choose from).

Note that the righthand example in Table XIX is missing one distractor; so do 21 examples in *Turney's dataset*. It also mixes different parts of speech: while *solitude* and *company* are nouns, all remaining words are adjectives. Other examples in *Turney's dataset* contain verbs and adverbs, and even relate pairs of different part of speech.

[7] Other researchers have proposed different generalizations, e.g., [Lin 1998].

Table XX. SAT verbal analogy: evaluation on 184 noun-only questions. For each model, the number of correctly classified, incorrectly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (% of examples for which the model makes prediction). The letters v, p and c indicate the kinds of features used: v stands for verb (possibly +preposition), p for preposition, and c for coordinating conjunction.

Model	**Correct**	**Incorrect**	**N/A**	**Accuracy**	**Coverage**
$v+p+c$	129	52	3	**71.27±6.98**	**98.37**
v	122	56	6	68.54±7.15	96.74
$v+p$	119	61	4	66.11±7.19	97.83
$v+c$	117	62	5	65.36±7.23	97.28
$p+c$	90	90	4	50.00±7.23	97.83
p	84	94	6	47.19±7.20	96.74
baseline	37	147	0	20.00±5.15	100.00
LRA: [Turney 2006]	122	59	3	67.40±7.13	98.37

Table XXI. Predicting Levi's RDP on the *Levi-214 dataset* using verbs v, prepositions p, and coordinating conjunctions c as features: leave-one-out cross-validation. Shown are micro-averaged accuracy and coverage in %, 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.

	Results When Using `THAT`				**Results When Not Using** `THAT`			
Model	**Accuracy**	**Cover.**	**ANF**	**ASF**	**Accuracy**	**Cover.**	**ANF**	**ASF**
Human: all v	78.4±6.0	99.5	34.3	70.9	–		–	–
Human: first v only	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	–	–	–	–	–	–

This is problematic for our approach, which requires that both words be nouns,[8] and thus we limit our evaluation to examples where all 12 words are nouns. After having filtered all examples containing non-nouns, we ended up with 184 questions, which we use in our experiments.

Given an SAT verbal analogy example, we build six feature vectors – one for each of the six word pairs. We calculate the similarity between the stem of the analogy and each of the five candidates using the Dice coefficient with TF.IDF-weighting, and we choose the pair with the highest score. We make no prediction if two or more pairs tie for the highest score.

The evaluation results are shown in Table XX: We use leave-one-out cross validation since we need to set the TF.IDF weights. The last line shows the performance of Turney's Latent Relational Analysis (LRA) when limited to the 184 noun-only dataset. Our best model $v+p+c$ performs a bit better, 71.27% vs. 67.40%, but the difference is not statistically significant (Pearson's Chi-square test). Note that this "inferred" accuracy could be misleading, and the LRA would have performed better if it was trained to solve *noun-only* analogies, which seem easier, as demonstrated by the significant increase in accuracy for LRA when limited to nouns: 67.4% vs. 56.8% for 184 and 374 questions, respectively. The learning process for LRA is probably harder on the full dataset, e.g., its pattern discovery step needs to learn patterns for many POS as opposed to nouns only, etc.

8.3. Predicting Levi's RDPs

Next, we experimented with trying to predict Levi's recoverably deletable predicates (see Table I) using verbs, prepositions and coordinating conjunctions that were automatically extracted from the Web as features. In these experiments, we only used those noun-noun compounds that are not nominalizations, i.e., for which Levi has an RDP provided; this left us with 214 examples (*Levi-214 dataset*) and 12 classes; this is a reduction from 250 examples and 16 classes.

The results are shown in Table XXI. Using prepositions alone only yields about 33% accuracy, which is a very statistically significant improvement over the majority-class baseline of 19.6% (two-tailed Pearson's χ^2 test, $p < 0.0015$), but is well below the classifier performance of 45.8% accuracy when using verbs.

Overall, the most important Web-derived features are the verbs: they yield 45.8% accuracy when used alone, and 50% when used 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 is statistically significant.

The righthand side of Table XXI reports the results when the query patterns involving `THAT` (see Section 8.1.1) were not used. We can observe a small 1-3% drop in accuracy for all models involving verbs, which is not statistically significant.

We further tried to make the predictions based on the human-proposed verbs instead. These results are shown in the same Table XXI. We achieved 78.4% accuracy using all human-proposed verbs, and 72.3% with the first verb from each judge. This result is very strong for a 12-way classification problem, and supports the hypothesis that the paraphrasing verbs are very important features for the task of noun-noun compound interpretation.

The difference between using Web-derived verbs and using human-proposed verbs (78.4% vs. 50%) is very statistically significant (according to a two-tailed Pearson's χ^2 test, $p < 0.0001$), and suggests that the human-proposed verbs might be an upper bound on the accuracy that could be achieved with automatically extracted features.

Table XXI also shows the average number of distinct features and the sum of feature counts per example. As we can see, for Web-derived features, there is a strong positive correlation between the number of extracted features and the classification accuracy, the best result being achieved with more than 200 features per example. Note, however, that using human-proposed verbs yields very high accuracy while using about seven times less features.

8.4. Predicting Head-Modifier Relations for the *Nastase & Szpakowicz Dataset*

Next, we experiment with the head-modifier dataset of Nastase and Szpakowicz [2003], which contains head-modifier relations between noun-noun and adjective-noun pairs. The dataset contains 600 head-modifier examples, each annotated with 30 fine-grained relations, grouped into 5 coarse-grained classes; see Section 2.2 for a list of these relations.

There are some problematic examples in this dataset. First, in three cases, there are two modifiers rather than one, e.g., *infectious disease agent*. In these cases, we ignore the first modifier. Second, 269 examples have an adjective as a modifier, e.g., *tiny cloud*. We treat them as if the modifier was a noun, which works in many cases, since many adjectives can be used predicatively, e.g., '*This cloud looks very tiny.*', which means that paraphrases could still be extracted.

[8]The approach can be extended to handle adjective-noun pairs, as demonstrated in section 8.4 below.

Table XXII. Head-modifier relations, 30 classes: evaluation on the *Nastase & Szpakowicz dataset*. For each model, the number of correctly classified, incorrectly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (% of examples for which the model makes prediction). Accuracy and coverage are micro-averaged.

Model	Correct	Incorrect	N/A	Accuracy	Coverage
$v+p$	240	352	8	**40.54±3.88**	**98.67**
$v+p+c$	238	354	8	40.20±3.87	98.67
v	234	350	16	40.07±3.90	97.33
$v+c$	230	362	8	38.85±3.84	98.67
$p+c$	114	471	15	19.49±3.01	97.50
p	110	475	15	19.13±2.98	97.50
baseline	49	551	0	8.17±1.93	100.00
LRA (Turney)	239	361	0	39.83±3.84	100.00

Table XXIII. Head-modifier relations, 5 classes: evaluation on the *Nastase & Szpakowicz dataset*. For each model, the number of correctly classified, incorrectly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (% of examples for which the model makes prediction). Accuracy and coverage are micro-averaged.

Model	Correct	Incorrect	N/A	Accuracy	Coverage
$v+p$	328	264	8	**55.41±4.03**	**98.67**
$v+p+c$	324	269	7	54.64±4.02	98.83
v	317	267	16	54.28±4.06	97.33
$v+c$	310	280	10	52.54±4.03	98.33
$p+c$	240	345	15	41.03±3.91	97.50
p	237	341	22	41.00±3.94	96.33
baseline	260	340	0	43.33±3.91	100.00
LRA (Turney)	348	252	0	58.00±3.99	100.00

For the evaluation, we create a feature vector for each head-modifier pair, and we perform a leave-one-out cross-validation: we leave one example for testing and we train on the remaining 599; we repeat this procedure 600 times, so that each example gets used for testing. Following Turney & Littman [2005], we use a 1-nearest-neighbor classifier. We calculate the similarity between the feature vector of the testing example and the vectors of the training examples using the Dice coefficient with TF.IDF-weighting. If there is a single highest-scoring training example, we predict its class for that test example. Otherwise, if there are ties for the top rank, we assume the class predicted by the majority of the tied examples, if there is a majority.

The results for the 30-class and 5-class *Nastase & Szpakowicz dataset* are shown in Tables XXII and XXIII, respectively. For the 30-way classification, our best model achieves 40.54% accuracy, which is comparable to the accuracy of Turney's LRA: 39.83%. For the 5-way classification, we achieve 55.41% vs. 58.00% for Turney's LRA. In either case, the differences are not statistically significant (tested with Pearson's Chi-square test). Given that Turney's algorithm requires substantial resources, synonym expansion, and costly computation over multiple machines, we believe that our simple approach is preferable.

Overall, we can see that it is best to use verbs and prepositions ($v+p$); adding coordinating conjunctions ($v+p+c$) lowers the accuracy. Using prepositions in addition to verbs ($v+p$) is better than using verbs only (v), but combining verbs and coordinating conjunctions ($v+c$) lowers the accuracy. Coordinating conjunctions only help when combined with prepositions ($p+c$). Overall, verbs are the most important features, followed by prepositions.

Table XXIV. Noun-noun compound relations, 19 classes: evaluation on the *Kim & Baldwin dataset*. For each model, the number of correctly classified, incorrectly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (% of examples for which the model makes prediction). Accuracy and coverage are micro-averaged.

Model	Correct	Incorrect	N/A	Accuracy	Coverage
$v+p+c$	43	45	0	48.86	100.00
$v+p$	43	45	0	48.86	100.00
$v+c$	43	45	0	48.86	100.00
v	43	45	0	48.86	100.00
p	37	51	0	42.05	100.00
$p+c$	36	52	0	40.91	100.00
baseline	36	52	0	40.90	100.00
Kim & Baldwin	46	42	0	52.60	100.00

8.5. Predicting Noun-Noun Compound Relations for the *Kim & Baldwin Dataset*

Another dataset we experimented with is the one created by Kim and Baldwin [2006], which consists of 453 noun-noun compounds. Unlike the *Nastase & Szpakowicz dataset*, here the modifier is always a noun, and there are only 19 classes, which are a subset of the 30 classes of Nastase and Szpakowicz [2003]: `agent`, `beneficiary`, `cause`, `container`, `content`, `equative`, `instrument`, `located`, `location`, `material`, `object`, `possessor`, `product`, `property`, `purpose`, `result`, `source`, `time`, `topic`. The dataset is split into 365 training and 88 testing examples. The baseline majority class classifier, which always chooses *topic*, yields 40.9% accuracy; the inter-annotator agreement is 52.3%.

Since the approach of Kim & Baldwin's algorithm had difficulties with the `time` and `equative` relations, they also experimented without them. For this 17-class dataset, there are 355 training and 85 testing examples, and the accuracy for the baseline majority class classifier is 42.3%.

It is important to note that this dataset has multiple labels for many of the examples. Following Kim and Baldwin [2006], we consider predicting any of the possibly multiple classes for a given example as a match.

For the evaluation, we first build weighted feature vectors for each of the 453 training examples. Then, for each testing example, we find the most similar training example, where the similarity is calculated using the Dice coefficient, as described above. We consider it a match if the two class sets have a non-empty overlap.

The results are shown in Table XXIV and Table XXV for the 19-class and 17-class datasets, respectively. Our results are significantly above the baseline: by 7-8% absolute. They are a bit worse but not statistically significantly different from those of Kim and Baldwin [2006], whose approach is more complicated and makes use of manually selected seed verbs as well as of various resources such as WordNet, CoreLex and Moby's thesaurus.

We can see once again that verbs are the most important features; moreover, prepositions and coordinating conjunctions have practically no impact when combined with verbs: v, $v+p$, $v+c$ and $v+p+c$ yield the same accuracy, while combining prepositions with coordinating conjunctions ($p+c$) decreases the performance compared to using prepositions only. The reason coordinating conjunctions do not help increase the accuracy is that noun-noun compound relations are best expressed with verbal or prepositional paraphrases; coordinating conjunctions only help with some infrequent relations such as `equative`, e.g., finding *(both) player and coach* suggests an `equative` relation for *player coach* or *coach player*.

Table XXV. Noun-noun compound relations, 17 classes: evaluation on the *Kim & Baldwin dataset*. For each model, the number of correctly classified, incorrectly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (% of examples for which the model makes prediction). Accuracy and coverage are micro-averaged.

Model	Correct	Incorrect	N/A	Accuracy	Coverage
$v+p+c$	42	43	0	49.41	100.00
$v+p$	42	43	0	49.41	100.00
$v+c$	42	43	0	49.41	100.00
v	42	43	0	49.41	100.00
p	36	49	0	42.35	100.00
$p+c$	35	50	0	41.18	100.00
baseline				42.30	100.00
Kim & Baldwin				52.60	100.00

Table XXVI. Relations between nominals: evaluation on the *SemEval-2007 Task 4* dataset. Accuracy is macro-averaged (in %), up to 10 search engine stars are used unless otherwise stated.

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

8.6. Predicting Relations Between Nominals as for SemEval-2007 Task 4

We further experimented with the SemEval-2007 Task 4 dataset [Girju et al. 2007; 2009], 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. For example:

```
"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, approximately 50% of which are positive.

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: words from the context sentence, after stopword removal and stemming with the Porter stemmer [Porter 1980];

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 as described below.

The SemEval-2007 Task 4 competition defined four types of systems, depending on whether the manually annotated WordNet senses and the query against the search engine 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. Regardless of classifier's prediction, if e_1's and e_2's heads had the same lemma, we classified the example as negative, e.g., the following training sentence would be considered a negative example for `Origin-Entity`: "*This surgical operation isolates most of the <e1>right hemisphere</e1> from the <e2>left hemisphere</e2>.*" The rational behind this is that semantic relations hold between different entities.

The results are shown in Table XXVI. Once again, we can see that verbs are the most important features, while the impact of prepositions, coordinating conjunctions and context words is rather limited; see [Nakov and Hearst 2007a] for additional details.

We also studied the effect of different subsets of features and of more search engine star operators. As Table XXVI shows, using up to ten stars instead of up to eight (see Section 8.1.1) 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 at SemEval-2007 Task 4, which achieved 66% and 67% accuracy, respectively.[9]

In future work, we plan experiments with the related but much larger, multi-way classification dataset of SemEval-2010 Task 8 [Hendrickx et al. 2009; 2010].

8.7. Discussion

We have seen that verbs are the single most important feature for predicting semantic relations, among the features that we have considered, followed by the prepositions and coordinating conjunctions. While prepositions were typically helpful when combined with verbs, the impact of conjunctions was minor, and sometimes even negative.

The reason is that semantic relations in noun-noun compounds, or between a head and a modifier, or in SemEval-2007 Task 4 sentences, are typically expressed with verbal or prepositional paraphrases; coordinating conjunctions only helped with infrequent relations such as *equative*, e.g., finding *player and coach* suggests an equative relation for *player coach* or *coach player*.

This is different for SAT verbal analogy, where the best model is $v + p + c$, as Table XX shows. Verbs are still the most important feature, and also the only one whose presence/absence makes a statistical difference. However, this time using c does help. The reason is that SAT verbal analogy questions ask for a broader range of relations, such as antonymy, for which coordinating conjunctions such as *but* can be helpful.

[9]The best type B system on SemEval-2007 Task 4 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. See [Girju et al. 2007] for details.

We have further seen that using more Web queries helps. Table XXI shows that having queries with an explicit THAT in addition to queries without it yields consistent improvements in paraphrasing verbs extraction (but it did not impact prepositions and coordinating conjunctions), and ultimately in classification accuracy. Table XXVI also shows improvements when using up to ten stars in queries, instead of eight.

Finally, Table XXI shows that the human-proposed paraphrasing verbs are very good features for predicting Levi's RDPs, yielding 78.4% accuracy, which is very high for a 12-way classification task; at the same time, using Web-derived paraphrasing verbs and prepositions only yielded 50% accuracy. These results suggest that the general idea of using paraphrasing verbs is very useful, even if data sparseness issues prevented automatic verb extraction from the Web to match the performance when using human-proposed verbs. The 78.4% accuracy suggests that, from a representation viewpoint, a frequency distribution over verbs is indeed very useful for capturing the semantics of abstract relations.

Overall, we have demonstrated that using Web-derived paraphrasing verbs can yield results that are competitive to the state-of-the-art on a number of datasets and semantic classification problems.

9. APPLICATION TO MACHINE TRANSLATION

Next, we show how our verb-based explicit paraphrases can help improve statistical machine translation (SMT). In these experiments, we restrict our focus to paraphrases involving prepositions, which can be also seen as a special kind of verbal paraphrases, e.g., *juice from apples* can be paraphrased as *juice that is from apples*.

Most modern SMT systems rely on aligned bilingual corpora, called bi-texts, from which they learn how to translate small pieces of text: individual words and short phrases, typically, up to seven tokens long. In many cases, these pieces are semantically equivalent but syntactically different from what can be matched in the test-time input, and thus the potential for high-quality translation can be missed: this is because translating using longer phrases usually yields better translation quality because phrases are internally fluent, and consistent, e.g., in terms of number/gender/person agreement, and have the correct internal word order; if long matches are not possible, the SMT system has to resort to word-for-word translation, which is typically very risky.

Below we try to increase the average length of the phrases used in the actual process of machine translation by expanding the training bi-text using paraphrases that are nearly equivalent semantically but different syntactically. In particular, we apply sentence-level paraphrasing on the source-language side, focusing on noun compounds: starting with a syntactic tree, we recursively generate new sentence variants where noun compounds are paraphrased using suitable prepositions (derived using Web frequencies), and vice versa – preposition-containing noun phrases are turned into noun compounds. We then pair each paraphrased sentence (we only paraphrase source-side sentences) with the original target-side translation, assuming that our paraphrases are meaning-preserving. Thus, we augment our training data "for free": by creating new data from data that is already available rather than having to create more aligned data, which would be costly.

It might appear somewhat inconsistent that we have argued above that using paraphrasing verbs is superior to using prepositions as verbs capture semantics better, but, yet, here we opt to use prepositions instead of verbs. While it is true that prepositions are inferior to verbs for semantic representation, they have some advantages when used as explicit paraphrases.

First, they are much easier to extract: given a list of prepositions, it is enough to instantiate them in exact-phrase queries (e.g., "*juice that is from apples*", "*juice from apples*", "*juice which is for apples*", "*juice for apples*", "*juice that is in apples*", "*juice in apples*", etc.) and then just to get the page hit for the target pattern directly, without the need for any syntactic analysis of Web snippets. Second, this also reduces the number of queries since one does not have to iterate over the results of a query to get 1000 snippets, nor does one need to include a variable number of stars in the query to be able to get more contexts. This is an important consideration since paraphrasing an entire training corpus involves orders of magnitude more compounds than just the few hundred that one finds in standard noun compound interpretation datasets. Third, since prepositions are used much more frequently than verbs, a paraphrase involving a preposition is much more likely to match test-time input than one involving a verb. Finally, we do not only paraphrase compounds, but we also turn explicit paraphrases into compounds; this latter operation is relatively easy with prepositions, but it can get complicated with verbs as they can get involved in more complex syntactic structures.

This is why below we restrict ourselves to prepositions and we do not use verbs. We also use genitives; technically, they are not compounds (they involve the genitive clitic *'s*, which is not a noun; it is an element of syntax), but still have a very similar structure. However, despite our choice not to consider paraphrasing verbs, we believe that they would be useful for SMT, and we plan to explore this direction in future work.

9.1. Sentence-Level Paraphrases

Given a sentence from the source (English) side of the training corpus, we generate conservative meaning-preserving syntactic paraphrases of that sentence. Each paraphrase is paired with the foreign (Spanish) translation that is associated with the original source sentence in the training bi-text. This augmented training corpus is then used to train an SMT system. Note that we only paraphrase the source (English) side of the training bi-text.

We further introduce a variation on this idea that can be used with a *phrase-based* SMT. In this alternative, the source-language *phrases* from the phrase table are paraphrased, but again using the target source-language phrase only, as opposed to requiring a third parallel pivot language as in [Callison-Burch et al. 2006].

Given a sentence like "*I welcome the Commissioner's statement about the progressive and rapid lifting of the beef import ban.*", we parse it using the Stanford parser [Klein and Manning 2003], and we recursively apply the following syntactic transformations:

(1) $\mathbf{[_{NP}\ NP_1\ P\ NP_2]} \Rightarrow \mathbf{[_{NP}\ NP_2\ NP_1]}$
the lifting of the beef import ban ⇒ *the beef import ban lifting*
(2) $\mathbf{[_{NP}\ NP_1}$ ***of*** $\mathbf{NP_2]} \Rightarrow \mathbf{[_{NP}\ NP_2\ gen\ NP_1]}$
the lifting of the beef import ban ⇒ *the beef import ban's lifting*
(3) $\mathbf{NP}_{gen} \Rightarrow \mathbf{NP}$
Commissioner's statement ⇒ *Commissioner statement*
(4) $\mathbf{NP}_{gen} \Rightarrow \mathbf{NP}_{PP_{of}}$
Commissioner's statement ⇒ *statement of (the) Commissioner*
(5) $\mathbf{NP}_{NC} \Rightarrow \mathbf{NP}_{gen}$
inquiry committee chairman ⇒ *inquiry committee's chairman*
(6) $\mathbf{NP}_{NC} \Rightarrow \mathbf{NP}_{PP}$
the beef import ban ⇒ *the ban on beef import*

where: **gen** is a genitive marker: ' or 's; **P** is a preposition; $\mathbf{NP}_{PP}$ is an NP with an internal PP-attachment; $\mathbf{NP}_{PP_{of}}$ is an NP with an internal PP headed by *of*; $\mathbf{NP}_{gen}$ is an NP with an internal genitive marker; $\mathbf{NP}_{NC}$ is an NP that is a noun compound.

The resulting paraphrases are shown in Table XXVIII. In order to prevent transformations (1) and (2) from constructing awkward noun phrases (NPs), we impose certain limitations on NP_1 and NP_2. They cannot span a verb, a preposition or a quotation mark (although they can contain some kinds of nested phrases, e.g., an ADJP in case of coordinated adjectives, as in *the progressive and controlled lifting*). Therefore, the phrase *reduction in the taxation of labour* is not transformed into *taxation of labour reduction* or *taxation of labour's reduction*. We further require the head to be a noun and we do not allow it to be an indefinite pronoun like *anyone*, *everybody*, and *someone*.

Transformations (1) and (2) are more complex than they may look. In order to be able to handle some hard cases, we apply additional restrictions. First, some determiners, pre-determiners and possessive adjectives must be eliminated in case of conflict between NP_1 and NP_2, e.g., *the lifting of this ban* can be paraphrased as *the ban lifting*, but not as *this ban's lifting*.[10] Second, in case both NP_1 and NP_2 contain adjectives, these adjectives have to be put in the right order, e.g., *the first statement of the new commissioner* can be paraphrased as *the first new commissioner's statement*, but not *the new first commissioner's statement*. There is also the option of not re-ordering them, e.g., *the new commissioner's first statement*.

Further complications are due to scope ambiguities of modifiers of NP_1. For example, in *the first statement of the new commissioner*, the scope of the adjective *first* is not *statement* alone, but *statement of the new commissioner*. This is very different for the NP *the biggest problem of the whole idea*, where the adjective *biggest* applies to *problem* only, and therefore it cannot be transformed to *the biggest whole idea's problem* (although we do allow for *the whole idea's biggest problem*).

Finally, a special problem is caused by measurements such as *spoon of salt*, *cup of coffee*, *glass of water*, *bowl of rice*, and *basket of apples*. For example, *cup of coffee* (measurement) cannot be paraphrased as *coffee cup* (kind of cup, purpose of the cup) or *coffee's cup* (possession) and vice versa. We opted not to handle these cases in any special way for the following reasons: (i) doing so requires a manual list of measurement words, which we did not have; (ii) some examples can have a non-measurement interpretation, e.g., *basket of apples* can refer to the object depicted on a still life oil painting by the French artist Paul Cézanne, and (iii) we paraphrase only the source side of a training bi-text, where over-generation is not very harmful – it will simply generate phrases that will not match the supposedly grammatical test-time input.

The first four transformations are syntactic, but (5) and (6) are not. The algorithm must determine whether a genitive marker is feasible for (5) and must choose the correct preposition for (6). In either case, for noun compounds of length three or more, we also need to choose the correct position to modify, e.g., *inquiry's committee chairman* vs. *inquiry committee's chairman*.

In order to improve the paraphrase accuracy, we use the Web as a corpus, generating and testing the paraphrases in the context of the words in the sentence. First, we split the noun compound into two sub-parts N_1 and N_2 in all possible ways, e.g., *beef import ban lifting* would be split as: (a) N_1="*beef*", N_2="*import ban lifting*", (b) N_1="*beef import*", N_2="*ban lifting*", and (c) N_1="*beef import ban*", N_2="*lifting*". For each split, we issue exact phrase queries against a search engine using the following patterns:

"lt N_1 gen N_2 rt"
"lt N_2 prep det N_1' rt"
"lt N_2 that be det N_1' rt"
"lt N_2 that be prep det N_1' rt"

[10] Some of these restrictions are not clear-cut: for example, one could argue that *this ban's lifting* could be an acceptable, even if unusual, paraphrase of *the lifting of this ban*.

Table XXVII. Example English phrases from the phrase table and corresponding automatic paraphrases.

1	**% of members of the irish parliament**
	% of irish parliament members
	% of irish parliament 's members
2	**universal service of quality .**
	universal quality service .
	quality universal service .
	quality 's universal service .
3	**action at community level**
	community level action
4	**, and the aptitude for communication and**
	, and the communication aptitude and
5	**to the fall-out from chernobyl .**
	to the chernobyl fall-out .
6	**flexibility in development - and quick**
	development flexibility - and quick
7	**, however , the committee on transport**
	, however , the transport committee
8	**and the danger of infection with aids**
	and the danger of aids infection
	and the aids infection danger
	and the aids infection 's danger

where: N_1' can be a singular or a plural form of N_1; `lt` is the word preceding N_1 in the original sentence, if any; `rt` is the word following N_2 in the original sentence, if any; `gen` is a genitive marker ('s or '); `that` is *that*, *which* or *who*; `be` is *is* or *are*; `det` is *the*, *a*, *an*, or none; and `prep` is one of the prepositions used by Lauer [1995] for noun compound (NC) interpretation: *about*, *at*, *for*, *from*, *in*, *of*, *on*, and *with*.

Given a particular split, we find the number of page hits for each instantiation of the above paraphrase patterns, filtering out the ones whose page hit counts are less than ten. We then calculate the total number of page hits H for all paraphrases (for all splits and all patterns), and we retain the ones whose page hits counts are at least 10% of H, which allows for multiple paraphrases (possibly corresponding to different splits) for a given noun compound. If no paraphrases are retained, we repeat the above procedure with `lt` set to the empty string. If there are still no good paraphrases, we set `rt` to the empty string. If this does not help either, we make a final attempt, by setting both `lt` and `rt` to the empty string. For example, *EU budget* is paraphrased as *EU's budget* and *budget of the EU*; also *environment policy* becomes *policy on environment*, *policy on the environment*, and *policy for the environment*; *UN initiatives* is paraphrased as *initiatives of the UN*, *initiatives at the UN*, and *initiatives in the UN*, and *food labelling* becomes *labelling of food* and *labelling of foods*.

We apply the same algorithm to paraphrase English *phrases* from the phrase table, but without transformations (5) and (6). See Table XXVII for sample paraphrases.

9.2. Experiments

We trained and evaluated several English→Spanish phrase-based statistical machine translation systems using the *Europarl corpus* [Koehn 2005] and the standard splits.

First, we built English→Spanish and Spanish→English directed word alignments using IBM model 4 [Brown et al. 1993], we combined them using the *intersect+grow heuristic* [Och and Ney 2003], and we extracted phrase-level translation pairs. We thus obtained a *phrase table* where each translation pair is associated with the following five standard parameters: forward phrase translation probability, reverse phrase translation probability, forward lexical translation probability, reverse lexical translation probability, and phrase penalty.

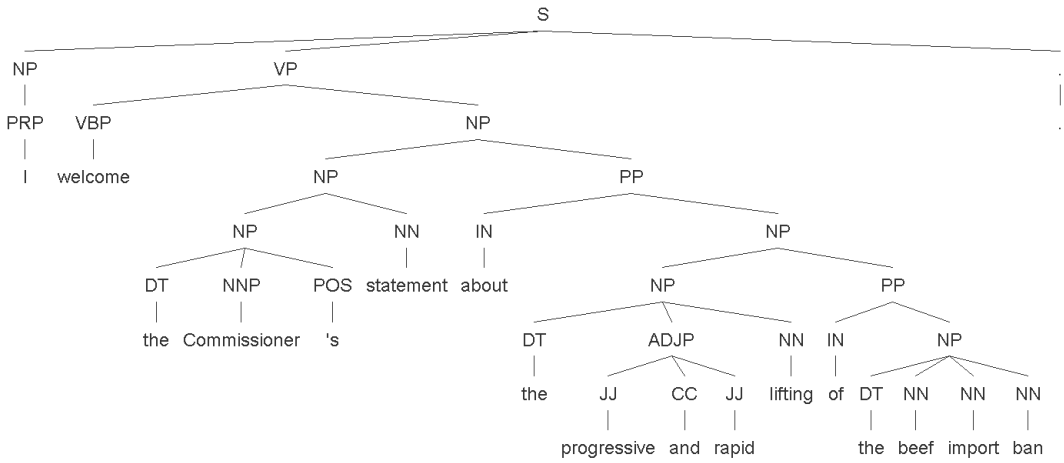

Fig. 5. Example parse tree generated by the Stanford parser. We transform noun compounds into NPs with an internal PP-attachment; we turn NPs with an internal PP-attachment into noun compounds, or into NPs with an internal possessive marker; and we remove possessive markers whenever possible, or substitute them with *of*. All these transformations are applied recursively.

We then trained a log-linear model using the following feature functions: language model probability, word penalty, distortion cost, and the above-mentioned parameters from the phrase table. We set the feature weights by optimizing the BLEU score directly using *minimum error rate training* (MERT) [Och 2003] on the first 500 sentences from the development set. We then used these weights in a beam search decoder [Koehn et al. 2007] to translate the 2,000 test sentences, and we compared the translations to the gold standard using BLEU [Papineni et al. 2001].

Baseline. Our baseline system S is trained on the original training corpus.

Sentence-Level Paraphrasing. We further built S_{pW}, which uses a version of the training corpus augmented with syntactic paraphrases of the English side *sentences* paired with their Spanish translations. In order to see the effect of not breaking NCs and not using the Web, we built S_p, which does not use transformations (5) and (6).

Phrase Table Paraphrasing. System $S^\star$ augments the *phrase table* of the baseline system S using syntactic transformations (1)-(4), as in S_p, i.e., without noun compound paraphrases. Similarly, $S^\star_{pW}$ is obtained by paraphrasing the *phrase table* of S_{pW}.

Combined Systems. Finally, we merged the phrase tables for some of the above systems, which we designate with a "+", e.g., $S + S_{pW}$ and $S^\star + S^\star_{pW}$. In these merges, the phrases from the first phrase table are given priority over those from the second one in case a phrase pair is present in both phrase tables. This is important since the parameters estimated from the original corpus are more reliable.

Following [Bannard and Callison-Burch 2005], we also performed an experiment with an additional feature F_{pW} for each phrase: its value is 1 if the phrase is in the phrase table of S, and 0.5 if it comes from the phrase table of S_{pW}. As before, we optimized the weights using MERT. For $S^\star + S^\star_{pW}$, we also tried using two features: in addition to F_{pW}, we introduced $F_\star$, whose value is 0.5 if the phrase comes from paraphrasing a phrase table entry, and 1 if it was in the original phrase table.

Table XXVIII. Example sentences and their automatically generated paraphrases. Paraphrased noun compounds are in italics.

I welcome the Commissioner 's statement about the progressive and rapid beef import ban lifting .
I welcome the progressive and rapid beef import ban lifting Commissioner 's statement .
I welcome the Commissioner 's statement about the beef import ban 's progressive and rapid lifting .
I welcome the beef import ban 's progressive and rapid lifting Commissioner 's statement .
I welcome the Commissioner 's statement about the progressive and rapid lifting of the *ban on beef imports* .
I welcome the Commissioner statement about the progressive and rapid lifting of the beef import ban .
I welcome the Commissioner statement about the progressive and rapid beef import ban lifting .
I welcome the progressive and rapid beef import ban lifting Commissioner statement .
I welcome the Commissioner statement about the beef import ban 's progressive and rapid lifting .
I welcome the beef import ban 's progressive and rapid lifting Commissioner statement .
I welcome the Commissioner statement about the progressive and rapid lifting of the *ban on beef imports* .
I welcome the statement of Commissioner about the progressive and rapid lifting of the beef import ban .
I welcome the statement of Commissioner about the progressive and rapid beef import ban lifting .
I welcome the statement of Commissioner about the beef import ban 's progressive and rapid lifting .
I welcome the statement of Commissioner about the progressive and rapid lifting of the *ban on beef imports* .
I welcome the statement of the Commissioner about the progressive and rapid lifting of the beef import ban .
I welcome the statement of the Commissioner about the progressive and rapid beef import ban lifting .
I welcome the statement of the Commissioner about the beef import ban 's progressive and rapid lifting .
I welcome the statement of the Commissioner about the progressive and rapid lifting of the *ban on beef imports* .
The EU budget , as an instrument of economic policy , amounts to 1.25 % of European GDP .
The EU budget , as an economic policy instrument , amounts to 1.25 % of European GDP .
The EU budget , as an economic policy 's instrument , amounts to 1.25 % of European GDP .
The *EU 's budget* , as an instrument of economic policy , amounts to 1.25 % of European GDP .
The *EU 's budget* , as an economic policy instrument , amounts to 1.25 % of European GDP .
The *EU 's budget* , as an economic policy 's instrument , amounts to 1.25 % of European GDP .
The *budget of the EU* , as an instrument of economic policy , amounts to 1.25 % of European GDP .
The *budget of the EU* , as an economic policy instrument , amounts to 1.25 % of European GDP .
The *budget of the EU* , as an economic policy 's instrument , amounts to 1.25 % of European GDP .
We must cooperate internationally , and this should include UN initiatives .
We must cooperate internationally , and this should include *initiatives of the UN* .
We must cooperate internationally , and this should include *initiatives at the UN* .
We must cooperate internationally , and this should include *initiatives in the UN* .
Both reports on economic policy confirm the impression that environment policy is only a stepchild .
Both reports on economic policy confirm the impression that *policy on the environment* is only a stepchild .
Both reports on economic policy confirm the impression that *policy on environment* is only a stepchild .
Both reports on economic policy confirm the impression that *policy for the environment* is only a stepchild .
Both economic policy reports confirm the impression that environment policy is only a stepchild .
Both economic policy reports confirm the impression that *policy on the environment* is only a stepchild .
Both economic policy reports confirm the impression that *policy on environment* is only a stepchild .
Both economic policy reports confirm the impression that *policy for the environment* is only a stepchild .
To the contrary , what is needed now - absolutely needed - is a reduction in the taxation of labour .
To the contrary , what is needed now - absolutely needed - is a reduction in the labour taxation .
To the contrary , what is needed now - absolutely needed - is a labour taxation reduction .
To the contrary , what is needed now - absolutely needed - is a reduction in the labour 's taxation .
To the contrary , what is needed now - absolutely needed - is a labour 's taxation reduction .

9.3. Evaluation Results

We report evaluation results in terms of BLEU, which is the standard automatic evaluation measure for machine translation.[11] We also use BLEU for tuning with MERT.

BLEU measures the similarity between a machine translation's output and one or more gold-standard human translations; it is defined as follows:

$$\text{BLEU} = BP \cdot \left(\prod_{n=1}^{N} p_n \right)^{\frac{1}{N}} \tag{5}$$

[11] Some popular alternatives include METEOR [Lavie and Denkowski 2009] and TER [Snover et al. 2006]. However, BLEU remains the de-facto standard.

Table XXIX. Notation for the experimental runs.

S	baseline, trained on the original corpus;
S_p	original corpus, augmented with sentence-level paraphrases, no transformations (5) and (6) (i.e. without using the Web);
S_{pW}	original corpus with sentence paraphrases, all transformations;
$\star$	means paraphrasing the phrase table;
$+$	means merging the phrase tables;
†	using an extra feature: F_{pW};
‡	using two extra features: $F_\star$, F_{pW}.

Table XXX. BLEU scores and n-gram precisions for 10k training sentences. The last two columns show the total number of entries in the phrase table and the number of phrases that were usable at testing time, respectively.

		n-gram precision					# of phrases	
System	**BLEU**	**1-gr.**	**2-gr.**	**3-gr.**	**4-gr.**	**BP**	**gener.**	**used**
S (baseline)	22.38	55.4	27.9	16.6	10.0	0.995	181k	41k
S_p	21.89	55.7	27.8	16.5	10.0	0.973	193k	42k
S_{pW}	22.57	55.1	27.8	16.7	10.2	1.000	202k	43k
$S^\star$	22.58	55.4	28.0	16.7	10.1	1.000	207k	41k
$S + S_p$	22.73	55.8	28.3	16.9	10.3	0.994	262k	54k
$S + S_{pW}$	**23.05**	55.8	28.5	17.1	10.6	0.995	280k	56k
$S + S_{pW}$†	**23.13**	55.8	28.5	17.1	10.5	1.000	280k	56k
$S^\star + S^\star_{pW}$	**23.09**	56.1	28.7	17.2	10.6	0.993	327k	56k
$S^\star + S^\star_{pW}$‡	**23.09**	55.8	28.4	17.1	10.5	1.000	327k	56k

BLEU has two components: (1) a brevity penalty (BP), and (2) a precision component, the geometric mean of n-gram precisions p_n, $1 \leq n \leq N$. The BP is defined as follows:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases} \quad (6)$$

where c is the length of the candidate, and r is the effective reference corpus length.

The evaluation results are shown in Tables XXX and XXXI. Table XXX shows not only the BLEU score, but also the individual components of BLEU. The differences between the baseline and the remaining systems shown in Table XXX are statistically significant, which was tested using bootstrapping [Zhang and Vogel 2004].

Gain of 33%–50% compared to doubling the training data. As Table XXXI shows, neither paraphrasing the training sentences, S_{pW}, nor paraphrasing the phrase table, $S^\star$, yields any notable improvements. For 10k training sentences, the systems are comparable and improve BLEU by .3, while for 40k sentences, $S^\star$ matches the baseline, and S_{pW} even drops below it. However, merging the phrase tables of S and S_{pW}, yields improvements of almost .7 for 10k and 20k sentences (both are statistically significant), and about .3 for 40k sentences (not statistically significant). While this improvement might look small, it is comparable to that of [Bannard and Callison-Burch 2005], who achieved .7 improvement for 10k sentences, and 1.0 for 20k (translating in the reverse direction: Spanish→English). Note also that the .7 improvement in BLEU for 10k and 20k sentences is about 1/3 of the 2 BLEU point improvement achieved by the baseline system by doubling the training size; it is also statistically significant. Similarly, the .3 gain on BLEU for 40k sentences is equal to half of what would have been gained if we had trained on 80k sentences.

Improved precision for all n-grams. Table XXX compares different systems trained on 10k sentences. Comparing the baseline with the last four systems, we can see that all n-gram precisions are improved by about .4-.7 points.

Importance of noun compound splitting. S_p is trained on the training corpus augmented with paraphrased sentences, where the noun compound splitting rules (5) and (6) are not used. We can see that the results for this system go below the baseline:

Table XXXI. BLEU scores for different number of training sentences.

	# of training sentences			
System	10k	20k	40k	80k
S (baseline)	**22.38**	**24.33**	**26.48**	**27.05**
S_{pW}	22.57	24.41	25.96	
$S^{\star}$	22.58	25.00	26.48	
$S + S_{pW}$	**23.05**	**25.01**	**26.75**	

while there is a .3 gain on unigram precision, bigram and trigram precision go down by about .1. BP decreases as well: since the sentence-level paraphrases (except for genitives, which are infrequent) convert NPs into NCs, the resulting sentences are shorter, and thus the translation model learns to generate shorter sentences. This is different in S_{pW}, where transformations (5) and (6) counter-weight (1)-(4), thus balancing BP. A different kind of argument applies to $S + S_p$, which is worse than $S + S_{pW}$, but not because of BP. In this case, there is no improvement for unigrams, but a consistent .2-.3 drop for bigrams, trigrams and fourgrams. The reason is shown in the last column of Table XXXI: omitting rules (5) and (6) results in fewer training sentences, which means smaller phrase table and thus fewer phrase pairs usable at translation time.

More usable phrases. The last two columns of Table XXX show that, in general, having more phrases in the phrase table implies more usable phrases at translation time. A notable exception is $S^{\star}$, whose phrase table is bigger than those of S_p and S_{pW}, but yields lower utility phrases. Therefore, we can conclude that the additional phrases extracted from paraphrased sentences are more likely to be usable at test time than the ones generated by paraphrasing the phrase table.

Paraphrasing sentences vs. paraphrasing the phrase table. As Tables XXX and XXXI show, paraphrasing the phrase table, as in $S^{\star}$ (BLEU score 22.58), cannot compete against paraphrasing the training corpus followed by merging the resulting phrase table with the phrase table for the original corpus[12], as in $S + S_{pW}$ (BLEU score 23.05). We also tried to paraphrase the phrase table of $S + S_{pW}$, but the resulting system $S^{\star} + S^{\star}_{pW}$ yielded little improvement: 23.09 BLEU score. Adding the two extra features, $F_{\star}$ and F_{pW}, also did not yield improvements: $S^{\star} + S^{\star}_{pW}$‡ achieved the same BLEU score as $S^{\star} + S^{\star}_{pW}$. This shows that extracting additional phrases from the augmented corpus is a better idea than paraphrasing the phrase table, which can result in erroneous splitting of noun phrases. Paraphrasing whole sentences as opposed to paraphrasing the phrase table could potentially improve the approach of [Callison-Burch et al. 2006] as well: while low probability and context dependency could be problematic, a language model could help filter the bad sentences out. Such filtering could potentially improve our results as well. Finally, note that different paraphrasing strategies could be used when paraphrasing phrases vs. sentences. For example, paraphrasing the phrase table can be done more aggressively: if an ungrammatical phrase is generated in the phrase table, it would most likely have no negative effect on translation quality since it would be unlikely to be observed at translation time.

Quality of the paraphrases and comparison to [Callison-Burch et al. 2006]. An important difference between our syntactic paraphrasing and the multilingual approach of Callison-Burch et al. [2006] is that their paraphrases are only contextually synonymous and often depart significantly from the original meaning. As a result, they could not achieve improvements by simply augmenting the phrase table: this introduced too much noise and the accuracy was below the baseline by 3-4 BLEU points.

[12]Note that $S^{\star}$ does not use rules (5) and (6). However, as $S + S_p$ shows, the claim holds even if these rules are excluded when paraphrasing whole sentences: the BLEU score for $S + S_p$ is 22.73 vs. 22.58 for $S^{\star}$.

Table XXXII. BLEU scores on the *News Commentary* data (64k sentences).

Model	LMs Trained On: News Only	LMs Trained On: News+Euro
S^{news}	32.27	**33.99**
$S^{news} \prec S^{news}_{pW}$	32.09	**34.42**
$S^{news} \prec S^{euro}$		**34.05**
$S^{news} \prec S^{news}_{pW} \prec S^{euro}$		34.25
$S^{news} \prec S^{euro} \prec S^{news}_{pW}$		**34.69**

In order to achieve an improvement, they had to introduce an extra feature penalizing the low probability paraphrases and promoting the original phrase table entries. In contrast, our paraphrases are meaning-preserving and less context-dependent. For example, introducing feature F_{pW} which penalises phrases coming from the paraphrased corpus in system $S + S_{pW}\dagger$ yielded a tiny improvement in BLEU score (23.13 vs. 23.05), i.e., the phrases extracted from our augmented corpus are almost as good as the ones from the original corpus. Finally, note that our paraphrasing method is *complementary* to that of [Callison-Burch et al. 2006] and therefore the two can be combined: the strength of our approach is in improving the *coverage of longer phrases* using syntactic paraphrases, while the strength of theirs is in improving the *vocabulary coverage* with words extracted from additional corpora (although they do get some gain from using longer phrases as well).

Paraphrasing the target side. We also tried paraphrasing the target language side, i.e., translating into English, which resulted in decreased performance. This is not surprising: the set of available source phrases remains the same, and a possible improvement could only come from producing a more fluent translation, e.g., from transforming an NP with an internal PP into a noun compound. However, unlike the original translations, the extra ones are a priori less likely to be judged correct since they were not observed on training.

News Commentary & Domain Adaptation. We further applied our paraphrasing to domain adaptation using the data from the ACL'07 Workshop on SMT: 1.3M words (64k sentences) of *News Commentary* data and 32M words of *Europarl* data. We used the standard training/tuning/testing splits, and we tested on *News Commentary* data.

This time we used two additional features with MERT (indicated with the $\prec$ operation): for the original and for the augmented phrase table, which allows extra weight to be given to phrases appearing in both. With the default distance reordering, for 10k sentences we had 28.88 BLEU for $S + S_{pW}$ vs. 28.07 for S, and for 20k we had 30.65 vs. 30.34. However, for 64k sentences, there was almost no difference: 32.77 vs. 32.73. Using a different tokenizer and a lexicalized reordering model, we got 32.09 vs. 32.34, i.e., the results were worse. However, as Table XXXII shows, using a second language model (LM) trained on *Europarl*, we were able to improve BLEU to 34.42 (for $S + S_{pW}$) from 33.99 (for S). Using S_{pW} lead to even bigger improvements (0.64 BLEU) when added to $S^{news} \prec S^{euro}$, where an additional phrase table from *Europarl* was used. See [Nakov 2008b] for further details.

9.4. Problems, Limitations, and Possible Extensions

Error analysis has revealed that the biggest problems for the proposed method are incorrect PP-attachments in the parse tree, and, less frequently, wrong POS tags (e.g., JJ instead of NN). Using a syntactic parser further limits the applicability of the approach to languages for which such parsers are available. In fact, for our purposes, it might be enough to use a shallow parser or just a POS tagger. This would cause problems with PP-attachment, but these attachments are often assigned incorrectly by parsers anyway.

The main target of our paraphrases are noun compounds – we turn NPs into noun compounds and vice versa – which limits the applicability of the approach to languages where noun compounds are a frequent phenomenon, e.g., Germanic, but not Romance or Slavic. From a practical viewpoint, an important limitation is that the size of the phrase table and/or of the training corpus increases, which slows down both training and translation, and limits the applicability to relatively small corpora for computational reasons. Last but not least, as Table XXXI shows, the improvements get smaller for bigger training corpora, which suggests that it becomes harder to generate useful paraphrases that are not already in the corpus.

While in our experiments we used phrase-based SMT, any machine translation approach that learns from parallel corpora could potentially benefit from the idea of syntactic corpus augmentation. At present, our paraphrasing rules are English-specific, but they could be easily adapted to other Germanic languages, which make heavy use of noun compounds; the general idea of automatically generating nearly equivalent source-side syntactic paraphrases can in principle be applied to any language.

We have seen that the benefits from paraphrasing diminish as the size of the training bi-text increases. This is common in SMT: the effect of smart improvements over a baseline typically decrease as the size of the training data increases. Thus, the proposed approach is worth considering primarily when training SMT systems on small corpora, as in the case of resource-poor language pairs. Note that most of the 6,500+ world languages are resource-poor from an SMT viewpoint, and this situation is unlikely to change in the near future; this number is even more striking when looking at *language pairs*. Moreover, resource-rich pairs can be resource-poor in specific domains.

Better use of the Web could be made for paraphrasing noun compounds (e.g., using verbal paraphrases), and other syntactic transformations could be tried (e.g., adding/removing complementisers like *that* and commas from nonmandatory positions). Moreover, a language model could be used to filter out the bad paraphrases.

Even more promising would be to use a tree-to-tree syntax-based SMT system and to learn suitable syntactic transformations that can make the source-language trees structurally closer to the target-language ones. For example, the English sentence "*Remember the guy who you are with!*" would be transformed into "*Remember the guy with whom you are!*", whose word order is closer to the Spanish "*¡Recuerda al individuo con quien estás!*", which might facilitate the translation process.

Finally, the process could be made part of the decoding, which would eliminate the need for the costly paraphrasing of the training corpus and might allow dynamically generating paraphrases both for the phrase table entries and for the target sentence that is being translated.

10. DISCUSSION

Below we discuss the potential applications that could benefit from paraphrasing verbs, prepositions and coordinating conjunctions. We also analyze the classes of compounds that can be handled as well as potential shortcomings and limitations.

10.1. Applications

As discussed above, semantic interpretation of noun-noun compounds can be used as a set of or a distribution over fine-grained paraphrasing verbs, or directly in paraphrases of the target noun-noun compounds tasks, e.g., for noun compound translation in isolation [Baldwin and Tanaka 2004; Grefenstette 1999; Tanaka and Baldwin 2003], for paraphrase-augmented machine translation [Callison-Burch et al. 2006; Nakov 2008a; Nakov and Hearst 2007b; Nakov 2008b], for machine translation evaluation [Russo-Lassner et al. 2005; Kauchak and Barzilay 2006], and for summarization evaluation [Zhou et al. 2006], among others.

As we have shown above, assuming annotated training data, the paraphrasing verbs can be used as features to predict abstract relations like CAUSE, USE, and MAKE. Such coarse-grained relations can in turn be helpful for other applications, e.g., for recognizing textual entailment as shown by Tatu and Moldovan [2005]. Note however, that, for this task, it is possible to use our noun compound paraphrasing verbs directly; for details, see [Nakov 2013] or Appendix B of [Nakov 2007].

In information retrieval, the paraphrasing verbs can be used for index normalization [Zhai 1997], query expansion, query refinement, results re-ranking, etc. For example, when querying for *migraine treatment*, search results containing good paraphrasing verbs like *relieve* or *prevent* could be preferred.

In text mining, the paraphrasing verbs can be used to seed a Web search that looks for particular classes of NPs such as diseases, drugs, etc. For example, after having found that *prevent* is a good paraphrasing verb for *migraine treatment*, we can use the query `"* which prevents migraines"` to obtain different treatments/drugs for migraine, e.g., *feverfew*, *Topamax*, *natural treatment*, *magnesium*, *Botox*, *Glucosamine*, etc. Using a different paraphrasing verb, e.g., using `"* reduces migraine"` can produce additional results: *lamotrigine*, *PFO closure*, *Butterbur Root*, *Clopidogrel*, *topamax*, *anticonvulsant*, *valproate*, *closure of patent foramen ovale*, *Fibromyalgia topamax*, *plant root extract*, *Petadolex*, *Antiepileptic Drug Keppra (Levetiracetam)*, *feverfew*, *Propranolol*, etc. This is similar to the idea of a relational Web search of Cafarella et al. [2006], whose system TextRunner serves four types of relational queries, among which there is one asking for all entities that are in a particular relation with a given target entity, e.g., "*find all X such that X prevents migraines*".

10.2. Classes of Noun Compounds

We should note that there exist several different classes of noun compounds and not all of them are paraphrasable using verbs and prepositions only. Still, most noun compounds are paraphrasable in some way, which is often done, e.g., in order to explain their meaning to somebody.

Let us start with *endocentric compounds* (known as *tatpurusa* in Sanskrit), which can be schematized as "AB is a type/kind of B", e.g., *orange juice*, *apple pie*, *malaria mosquito*. In these compounds, the modifier attributes some property to the head. This property can be made explicit by means of a paraphrase involving verbs and/or prepositions only, e.g., *be squeezed from*, *be extracted from*, and *from* for *orange juice*, *be made of*, *contain*, and *from* for *apple pie*, and *cause*, *carry*, and *with* for *malaria mosquito*.

Such kinds of paraphrases do not work well for most *exocentric compounds* (*bahuvrihi* in Sanskrit), which lack an overtly expressed semantic head, e.g., *birdbrain* is a kind of person, not a kind of brain. Still, most exocentric compounds can be paraphrased, e.g., "*birdbrain* is a stupid person, i.e., one whose *brain* is the size of a *bird*'s brain", similarly, *ladyfinger* can be analyzed as "a pastry that resembles a *lady finger*". However, as these examples show, such paraphrases need to be more complex and do not only have to include the nouns forming the noun compound, but also to make explicit the actual head of the noun compound, here *person* and *pastry*, respectively. Thus, while these compounds are, in principle, interpretable by paraphrases, in practice, they cannot be handled by the approach presented here.

There is a third category of compounds, known as *appositional*, where each of the nouns forming the compound expresses a different aspect of the whole that the compound represents, e.g., *coach-player* is somebody who is both a coach and a player. Such compounds are paraphrasable as "AB *is both* A *and* B" or "AB *is* A *and also* B", which involve adverbs like *both* and *also*, and thus cannot be adequately handled by the method presented here; still, they are paraphrasable as *be*, e.g., *coach who is a player*, which is not ideal since it does not fully capture the semantics, but is acceptable.

A related category of compounds, known as *copulative* or *coordinative* (*dvandva* in Sanskrit), consists of compounds that form an entity that is the sum of the nouns it is made of and is also distinct from any of them, e.g., *Bosnia-Herzegovina* or *gerund-participle*. Just like the previous category, compounds belonging to this category cannot be paraphrased using verbs and prepositions.

Two other categories of compounds that the proposed method does not cover are (a) those formed by reduplication and (b) portmanteaux. There are various kinds of *reduplication* in English such as exact (e.g., *bye-bye*), ablaut (e.g., *chit-chat*), rhyming (e.g., *walkie-talkie*, *hokey-pokey*), sh-reduplication (e.g., *baby-shmaby*), contrastive (e.g., *I'll make the tuna salad, and you make the salad-salad.*), etc. Despite this variety, reduplication is not very common nor is it very productive in English, except probably for the last two categories. *Portmanteaux* compounds, are composed by blending the sounds of two or more words, while also combining their meanings, e.g., *brunch*, formed from *breakfast* + *lunch*, *Eurasia* which blends *Europe* and *Asia*, and *Merkozy*, made of *Merkel* and *Sarkozy*. Portmanteaux are fairly rare in English.[13]

Overall, the proposed method covers adequately the endocentric compounds only. However, they are by far the most frequent and the most productive in English. Exocentric compounds are mostly idiomatic, which makes them rare and almost a closed class, and thus potentially listable in a dictionary. The other kinds of compounds are even more rare in English.

The method would also have problems paraphrasing right-headed compounds in English, since they do not follow the pattern "AB is a kind of B", which is typical for left-headed endocentric compounds, e.g., *vitamin D* is a kind of *vitamin*, not a kind of *D*. Right-headed compounds in English often have as a second word an identifying name or a number, e.g., *interferon alpha*, *vitamin D*, *exit 15*, *Route 66*, *Linguistics 101*, *Cafe Viena*, which makes them hard to paraphrase with verbs and prepositions.

Other, equally non-paraphrasable, right-headed noun-noun compounds include borrowings from languages like French, which is mostly left-headed, e.g., *beef Julienne*, and compounds where the first noun is a classifier, e.g., *Mount Whitney*, *planet Earth*, and *President Obama*.

Whether the method would be applicable or not depends also on the transparency of the compound. Levi [1978] arranges compounds in a transparency scale as follows: (1) transparent, e.g., *mountain village*, *orange peel*, (2) partly opaque, e.g., *grammar school*, *brief case*, (3) exocentric, e.g., *birdbrain*, *ladybird*, (4) partly idiomatic, e.g., *monkey wrench*, *flea market*, and (5) completely idiomatic, e.g., *honeymoon*, *duck soup*. Paraphrasing using verbs and prepositions works best for the first category, e.g., *mountain village* can be adequately paraphrased as *a village that is surrounded by/is nested between or amid/sits among/is in (the) mountains*. It also works to some extent for the second category, e.g., *grammar school* is more than just *a school that teaches grammar*, even though that would be an arguably valid paraphrase; it is *a school teaching classical languages*, and more recently, *an academically-oriented secondary school*. Finally, verbs and prepositions have little chance, if any, to paraphrase noun compounds from the last three categories.

10.3. Potential Shortcomings

Context. Note that the proposed approach interprets noun compounds outside of context, but the interpretation of many noun compounds can be context-dependent. For example, *museum book* can mean *a book about a museum*, *a book bought in a museum*, *a book on display in a museum*, etc. Our paraphrasing approach does not model context at all and thus it could mix several of these interpretations.

[13]The interested reader can find more about portmanteaux is English in [Cook and Stevenson 2010].

A more fundamental problem is that compounds could get completely novel interpretations in a certain context, e.g., *museum book* could mean *a book that I am planning to take with me when I go to the museum tomorrow*; such interpretations are out of reach for the presented approach, as well as for all alternative computational approaches we are aware of.

Data sparseness. A general problem with our Web-based paraphrasing approach is that it suffers from data sparseness issues in the case of rare words or of rarely co-occurring terms. This can be alleviated by combining our relational features (verbs, prepositions and coordinating conjunctions) with attributional features (e.g., hypernyms and co-hyponyms of the relation arguments which can be automatically extracted with lexico-syntactic patterns [Hearst 1992]). The potential of such a combination has already been demonstrated in [Séaghdha and Copestake 2009] and [Nakov and Kozareva 2011]. For example, for SemEval-2007 Task 4, Nakov and Kozareva [2011] report a state-of-the-art accuracy of 71.3% for the combination of relational and attributional features, compared to 68.1% when using relational features only.

Corpus size. Another interesting research question is how our paraphrases extracted from the Web by shallow pattern matching compare to paraphrases extracted from a smaller but properly parsed corpus. There are many indicators in the literature that size is the thing that matters most in corpus-based linguistics, and thus using the Web should be preferable because it is orders of magnitude larger than other existing corpora [Halevy et al. 2009]. For example, Lapata and Keller [2005] have shown that using the Web (in their case, Altavista as of 2004) outperforms using a fixed corpus such as the British National Corpus (BNC) for a number of NLP problems, including noun compound interpretation using the eight paraphrasing prepositions of Lauer [1995]: using the Web yielded 55.71% accuracy vs. only 27.85% for BNC. While using a corpus with proper parsing might help reduce noise of extraction, accumulating statistics over a corpus that is orders of magnitude larger, e.g., the Web is about a millon times larger than the 100 million words BNC, might be even more efficient at reducing noise. A major problem with using a fixed corpus is data sparseness: we have seen that this is a problem for us even when using the Web, and it only gets worse with orders of magnitude smaller corpora.

Nakov [2007] reports that extracting prepositional paraphrases for biomedical noun compounds from a properly POS-tagged 420 million words subset of MEDLINE yields higher accuracy than using Web snippets (93.3% vs. 89.7%), but this comes at the expense of lower coverage (83.6% vs. 99.3%); experimenting on this same corpus with general noun compounds from Levi's dataset yielded much more extreme degradation in performance – less than 50% accuracy (vs. 82.1% with Web statistics) and less than 10% coverage (vs. 86.9% with Web statistics). Thus, we believe that smaller corpora, even if better processed linguistically, do not appear to be a viable alternative to the orders of magnitude larger Web.

11. CONCLUSIONS AND FUTURE WORK

We have presented a simple, lightly supervised approach to noun-noun compound interpretation which uncovers the predicates that can be used to paraphrase the hidden relations between the two nouns. We have shown the potential for this approach for several NLP tasks. We have explored and experimentally tested the idea that, in general, the semantics of a noun-noun compound can be characterized by the set of paraphrasing verbs that connect the target nouns, with associated weights. These verbs are fine-grained, directly usable as paraphrases, and the use of multiple verbs for a given noun-noun compound allows for better approximating its semantics.

An important advantage of the approach is that it does not require knowledge about the meaning of the constituent nouns in order to correctly assign relations. A potential drawback is that it might not work well for low-frequency words.

We have also shown that the explicit paraphrases we generate can help improve statistical machine translation, even when the paraphrases are limited to prepositions; in future work, we plan to experiment with paraphrasing verbs as well. Using a language model to filter unreliable sentence-level paraphrases is another possible extension.

While verbs and prepositions are the most frequent ways to paraphrase a noun-noun compound, there are many other ways to express the explicit relationship between the two nouns, e.g., for *onion tears*, these could include: *tears from onions*, *tears due to cutting onions*, *tears induced when cutting onions*, *tears that onions induce*, *tears that come from chopping onions*, *tears that sometimes flow when onions are chopped*, *tears that raw onions give you*, etc. This richer paraphrasing can handle some compounds that cannot be paraphrased using verbs and prepositions alone, e.g., Downing's [1977] *oil bowl*, which is paraphrasable as *a bowl into which the oil in the engine is drained during an oil change*.[14] Exploring such free paraphrases is the topic of SemEval-2013, task 4.[15] We believe that they have great potential for the semantic interpretation of noun compounds, and we plan to explore them in future work.

We should note that paraphrasing verbs and prepositions are useful not only when used as paraphrases but also as features, and not only for noun-noun compounds. Our results show that Web-derived paraphrasing verbs and prepositions are good features for solving various relational similarity problems including SAT verbal analogy, head-modifier relations, and relations between complex nominals. Thus, in future work, we plan to apply our noun compound interpretation framework to other NLP problems.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their constructive comments and suggestions, which have helped us improve the quality of the manuscript.

REFERENCES

ALSHAWI, H. AND CARTER, D. 1994. Training and scaling preference functions for disambiguation. *Computational Linguistics 20,* 4, 635–648.

BAKER, C. F., FILLMORE, C. J., AND LOWE, J. B. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*. ICCL '98. 86–90.

BAKER, C. F. AND RUPPENHOFER, J. 2002. FrameNet's frames vs. Levin's verb classes. In *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*. 27–38.

BALDWIN, T. AND TANAKA, T. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*. MWE '04. 24–31.

BANNARD, C. AND CALLISON-BURCH, C. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL '05. 597–604.

BARKER, K. AND SZPAKOWICZ, S. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th International Conference on Computational Linguistics*. ICCL '98. 96–102.

BROWN, P. F., PIETRA, V. J. D., PIETRA, S. A. D., AND MERCER, R. L. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics 19,* 2, 263–311.

BUDANITSKY, A. AND HIRST, G. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics 32,* 1, 13–47.

BUTNARIU, C., KIM, S. N., NAKOV, P., Ó SÉAGHDHA, D., SZPAKOWICZ, S., AND VEALE, T. 2009. SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Pro-*

[14]It could be also paraphrased as *bowl for containing oil*, but this is quite an impoverished interpretation.
[15]`http://www.cs.york.ac.uk/semeval-2013/task4/`

ceedings of the NAACL-HLT-09 Workshop on Semantic Evaluations: Recent Achievements and Future Directions. SEW '09. 100–105.

BUTNARIU, C., KIM, S. N., NAKOV, P., Ó SÉAGHDHA, D., SZPAKOWICZ, S., AND VEALE, T. 2010. SemEval-2 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval '10. 39–44.

BUTNARIU, C. AND VEALE, T. 2008. A concept-centered approach to noun-compound interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics*. COLING '08. 81–88.

CAFARELLA, M., BANKO, M., AND ETZIONI, O. 2006. Relational Web search. Technical Report 2006-04-02, University of Washington, Department of Computer Science and Engineering.

CALLISON-BURCH, C., KOEHN, P., AND OSBORNE, M. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. HLT-NAACL '06. 17–24.

COOK, P. AND STEVENSON, S. 2010. Automatically identifying the source words of lexical blends in English. *Computational Linguistics 36,* 1, 129–149.

DEVEREUX, B. AND COSTELLO, F. 2006. Modelling the interpretation and interpretation ease of noun-noun compounds using a relation space approach to compound meaning. Erlbaum, 184–189.

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

FELLBAUM, C., Ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

FILLMORE, C. J. 1968. Universals in linguistic theory. In *The case for case*, Bach and Harms, Eds. Holt, Rinehart, and Winston, New York, 1–88.

FILLMORE, C. J. 1982. Frame semantics. Hanshin Publishing Co., 111–137.

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

GIRJU, R., MOLDOVAN, D., TATU, M., AND ANTOHE, D. 2005. On the semantics of noun compounds. *Journal of Computer Speech and Language - Special Issue on Multiword Expressions 4,* 19, 479–496.

GIRJU, R., NAKOV, P., NASTASE, V., SZPAKOWICZ, S., TURNEY, P., AND YURET, D. 2007. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*. SemEval '07. 13–18.

GIRJU, R., NAKOV, P., NASTASE, V., SZPAKOWICZ, S., TURNEY, P. D., AND YURET, D. 2009. Classification of semantic relations between nominals. *Language Resources and Evaluation 43,* 2, 105–121.

GOLDBERG, A. E. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago.

GREFENSTETTE, G. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Translating and the Computer 21: Proceedings of the 21st International Conference on Translating and the Computer*.

GRISHMAN, R. AND STERLING, J. 1994. Generalizing automatically generated selectional patterns. In *Proceedings of the 15th conference on Computational linguistics*. COLING '94. 742–747.

HALEVY, A., NORVIG, P., AND PEREIRA, F. 2009. The unreasonable effectiveness of data. *Intelligent Systems, IEEE 24,* 2, 8–12.

HEARST, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*. 539–545.

HENDRICKX, I., KIM, S. N., KOZAREVA, Z., NAKOV, P., Ó SÉAGHDHA, D., PADÓ, S., PENNACCHIOTTI, M., ROMANO, L., AND SZPAKOWICZ, S. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. SEW '09. 94–99.

HENDRICKX, I., KIM, S. N., KOZAREVA, Z., NAKOV, P., Ó SÉAGHDHA, D., PADÓ, S., PENNACCHIOTTI, M., ROMANO, L., AND SZPAKOWICZ, S. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval '10. 33–38.

HENDRICKX, I., KOZAREVA, Z., NAKOV, P., SÉAGHDHA, D. O., SZPAKOWICZ, S., AND VEALE, T. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '13.

JACKENDOFF, R. 1983. *Semantics and Cognition*. MIT Press, Cambridge, MA.

KATZ, J. AND FODOR, J. 1963. The structure of a semantic theory. *Language* 39, 170–210.

KAUCHAK, D. AND BARZILAY, R. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. HLT-NAACL '06. 455–462.

KIM, S. N. AND BALDWIN, T. 2005. Automatic interpretation of compound nouns using WordNet similarity. In *Proceedings of 2nd International Joint Conference on Natural Language Processing*. IJCNLP '05. 945–956.

KIM, S. N. AND BALDWIN, T. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*. ACL-COLING '06. 491–498.

KIM, S. N. AND NAKOV, P. 2011. Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. 648–658.

KLEIN, D. AND MANNING, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. ACL '03. 423–430.

KOEHN, P. 2005. Europarl: A parallel corpus for evaluation of machine translation. In *Proceedings of the X MT Summit*. 79–86.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A., AND HERBST, E. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. ACL '07. 177–180.

LAPATA, M. 2002. The disambiguation of nominalizations. *Computational Linguistics 28,* 3, 357–388.

LAPATA, M. AND KELLER, F. 2005. Web-based models for natural language processing. *ACM Trans. Speech Lang. Process. 2,* 1, 1–31.

LAUER, M. 1995. Designing statistical language learners: Experiments on noun compounds. Ph.D. thesis, Department of Computing Macquarie University.

LAVIE, A. AND DENKOWSKI, M. J. 2009. The Meteor metric for automatic evaluation of machine translation. *Machine Translation 23*, 105–115.

LEVI, J. 1978. *The Syntax and Semantics of Complex Nominals.* Academic Press, New York.

LIN, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML '98. 296–304.

LIN, D. AND PANTEL, P. 2001. Discovery of inference rules for question-answering. *Natural Language Engineering 7,* 4, 343–360.

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 Workshop on Computational Lexical Semantics*. 60–67.

NAKOV, P. 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.

NAKOV, P. 2008a. Improved statistical machine translation using monolingual paraphrases. In *Proceedings of the European Conference on Artificial Intelligence*. ECAI '08. 338–342.

NAKOV, P. 2008b. Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*. WMT '08. 147–150.

NAKOV, P. 2008c. Noun compound interpretation using paraphrasing verbs: Feasibility study. In *Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications*. AIMSA '08. 103–117.

NAKOV, P. 2008d. Paraphrasing verbs for noun compound interpretation. In *Proceedings of the LREC'08 Workshop: Towards a Shared Task for Multiword Expressions*. MWE '08. 46–49.

NAKOV, P. 2013. On the interpretation of noun compounds: Syntax, semantics, entailment. *Journal of Natural Language Engineering*.

NAKOV, P. AND HEARST, M. 2007a. UCB: System description for SemEval Task #4. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*. SemEval '07. 366–369.

NAKOV, P. AND HEARST, M. 2007b. UCB system description for the WMT 2007 shared task. In *Proceedings of the Second Workshop on Statistical Machine Translation*. WMT '07. 212–215.

NAKOV, P. AND HEARST, M. 2008. Solving relational similarity problems using the web as a corpus. In *Proceedings of the 46th Annual Meeting on Association for Computational Linguistics*. ACL '08. 452–460.

NAKOV, P. AND HEARST, M. A. 2006. Using verbs to characterize noun-noun relations. In *AIMSA*, J. Euzenat and J. Domingue, Eds. Lecture Notes in Computer Science Series, vol. 4183. Springer, 233–244.

NAKOV, P. AND KOZAREVA, Z. 2011. Combining relational and attributional similarity for semantic relation classification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. RANLP '11. 323–330.

NAKOV, P., SCHWARTZ, A., AND HEARST, M. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of SIGIR'04 Workshop on Search and Discovery in Bioinformatics*. 81–88.

NASTASE, V. AND SZPAKOWICZ, S. 2003. Exploring noun-modifier semantic relations. In *Fifth International Workshop on Computational Semantics*. IWCS '03. 285–301.

OCH, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. ACL '03. 160–167.

OCH, F. J. AND NEY, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics 29,* 1, 19–51.

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

Ó SÉAGHDHA, D. 2009. Semantic classification with WordNet kernels. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. NAACL '09. 237–240.

PAPINENI, K., ROUKOS, S., WARD, T., AND ZHU, W.-J. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. 311–318.

PETRUCK, M. 1996. Frame semantics. In *Handbook of Pragmatics*, J. Verschueren, J.-O. stman, J. Blommaert, and C. Bulcaen, Eds. John Benjamins, 1–13.

PORTER, M. 1980. An algorithm for suffix stripping. *Program 14,* 3, 130–137.

ROSARIO, B., HEARST, M. A., AND FILLMORE, C. 2002. The descent of hierarchy, and selection in relational semantics. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. 247–254.

RUGE, G. 1992. Experiment on linguistically-based term associations. *Information Processing and Management 28,* 3, 317–332.

RUSSO-LASSNER, G., LIN, J., AND RESNIK, P. 2005. A paraphrase-based approach to machine translation evaluation. Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland. August.

SAEED, J. 2003. *Semantics* 2 Ed. Blackwell.

SÉAGHDHA, D. O. AND COPESTAKE, A. 2007. Co-occurrence contexts for noun compound interpretation. In *Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions*. MWE '07. 57–64.

SÉAGHDHA, D. O. AND COPESTAKE, A. 2009. Using lexical and relational similarity to classify semantic relations. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '09. 621–629.

SHINYAMA, Y., SEKINE, S., AND SUDO, K. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of the second international conference on Human Language Technology Research*. 313–318.

SNOVER, M., DORR, B., SCHWARTZ, R., MICCIULLA, L., AND MAKHOUL, J. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*. AMTA '06. 223–231.

TANAKA, T. AND BALDWIN, T. 2003. Noun-noun compound machine translation: a feasibility study on shallow processing. In *Proceedings of the ACL 2003 workshop on Multiword expressions*. MWE '03. 17–24.

TATU, M. AND MOLDOVAN, D. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. HLT-EMNLP '05. 371–378.

TOUTANOVA, K. AND MANNING, C. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. EMNLP/VLC '00. 63–71.

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*. ACL '10. 678–687.

TURNEY, P. AND LITTMAN, M. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning Journal 60,* 1-3, 251–278.

TURNEY, P. D. 2006. Similarity of semantic relations. *Computational Linguistics 32,* 3, 379–416.

VANDERWENDE, L. 1994. In *Proceedings of the 15th conference on Computational linguistics - Volume 2*. COLING '94. 782–788.

WARREN, B. 1978. Semantic patterns of noun-noun compounds. In *Gothenburg Studies in English 41, Goteburg, Acta Universtatis Gothoburgensis*.

ZHAI, C. 1997. Fast statistical parsing of noun phrases for document indexing. In *Proceedings of the fifth conference on Applied natural language processing*. ANLP '97. 312–319.

ZHANG, Y. AND VOGEL, S. 2004. Measuring confidence intervals for the machine translation evaluation metrics. In *10th International Conference on Theoretical and Methodological Issues in Machine Translation*. TMI '04. 4–6.

ZHOU, L., LIN, C.-Y., AND HOVY, E. 2006. Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP '06. 77–84.

Received February 2007; revised March 2009; accepted June 2009