

The Construction of an Object-Oriented Corpus Using Prepositional Paraphrases*

Joakim Nivre

School of Mathematics and Systems Engineering
Växjö University
SE-35195, Växjö, Sweden
nivre@msi.vxu.se

Noah A. Smith

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
nasmith@cs.cmu.edu

Abstract

The present work examines corpus construction comprising nouns and noun compounds. This work presents arguments claiming that the unpredictable nature of words may be circumvented by alternatively addressing the predictable nature of concepts themselves. Thus, an object-oriented approach is used to build a corpus using prepositional paraphrases by means of semantic relations comprising the set of all prepositions. The resulting corpus also provides an intuitive naming convention and hence corpus, having the ability to greatly simplify technical terminology.

1 Introduction

Establishing the semantics of noun compounds remains a core problem in linguistics often requiring the use of empirical knowledge (pragmatics) to aid in the interpretation of concepts. The present work examines corpus construction comprising nouns and noun compounds. Although current corpora like WordNet (Fellbaum, 1998) employ an *is-a* relation to form relations between words, grouped as synsets, this work presents arguments claiming that the unpredictable nature of synsets may be circumvented by alternatively addressing the predictable nature of

concepts themselves. To justify the construction of this corpus, a proof is offered arguing unsupervised learning requires both the corpus and the method for examining the corpus to be logical, predictable.

In an attempt to uncover the core set of semantic relations linking the modifiers and the head of all noun compounds, previous works (Levi 1978, Warren 1978, Finin 1980, Isabelle 1984) have introduced various semantic relations. These relations have varied from prepositional combinations (Warren 1978, Lauer 1995) to more precisely-defined abstract predicates (HAVE, USE) (Nastase and Szpakowicz 2003, Girju 2007). In general, these semantic relations stand as the foundation of corpus construction.

The present work commences with a proof justifying the construction of a new corpus. The construction of this corpus largely departs from the prior-art by assuming an object-oriented approach. Therefore the major focus rests on concepts as opposed to words or synsets; one can also think of this as a relationship between definitions. Within the object-oriented framework, objects(words) are grouped by classes(the most generalized form of a word) and modified by members of their respective classes; members which are generated through prepositions.

Redefinition defines the remaining work. The lexical hierarchical concepts of branching and non-branching are dismissed by way of proof and replaced by *independent* and *dependent*, respectively.

In lieu of experimental results, the work concludes by presenting a promising new corpus by solving a series of outstanding linguistic problems.

*This document has been adapted from the instructions for earlier ACL proceedings, including those for ACL-05 by Hwee Tou Ng and Kemal Oflazer, those for ACL-02 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats. Those versions were written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence*.

These results are accompanied by examples of simplified technical terminology.

2 Corpus Construction Justification

When dealing with prepositional paraphrases (N_1PN_2) in the literature of (Lauer 1995, Nastase and Szpakowicz 2003, Girju 2006), it is assumed prepositions dictate the relation between two nouns (the head and the noun modifier). In mathematics this relation is termed a mapping, otherwise known as a function.

All functions are defined as single-valued, meaning that given an input a function will yield a unique output. Previous semantic analyses have been spawned in response to the ambiguity of semantic relation sets, for example (e.g. the eight prepositions of Lauer(1995) adapted from Warren(1978)). This vagueness finds root in the multi-valued nature (polysemy) of prepositions. Without denying the existence of the semantic relations of prepositions, Girju (2006, 2007) has responded with a set of precisely-defined semantic relations to offer unambiguous, single-valued relations of deletable predicates as a better alternative.

Despite the efforts to obtain well-defined, single-valued semantic relations, researchers have frequently overlooked the fact that given a noun compound and a semantic relation such as a deletable predicate, which can be represented as three separate functions, the noun compound will never be single-valued (unambiguous) unless all three functions (the two nouns and the deletable predicate) are single-valued.

Let f be single valued function such $f(x) \neq f(y) \forall x \neq y$ $f(x)g(x) = (f \circ g)(x) = f(g(x))$, by composition Now suppose g is multi-valued function, then $\exists y : g(x) = a$ and $g(x) = b \forall a \neq b$ $f(g(x)) = f(a)$ and $f(b)$ therefore $(f \circ g)(x)$ is multivalued

In summary, the justification for corpus construction is the need for unambiguity.

3 Redefinition of Branching and Non-Branching

In previous work (Cruse, 1986), branching and non-branching classifications are used in determining the relationship between a series of words. Being resourceful, we can take full advantage of pre-existing work in mathematical independence and dependence.

When dealing with objects, a class (c_k) represents a property shared between two words (sets) W_i and W_j , where i, j , and $k \in \mathbb{N}$.

Branching: $W_i \cap W_j = \emptyset$, where $i \neq j$

Non-Branching: $W_i \cap W_j = c_k$, where $i \neq j$

By replacement of variables:

Independent: $A_i \cap A_j = \emptyset$, where $i \neq j$

Dependent: $A_i \cap A_j \neq \emptyset$, where $i \neq j$

Therefore branching equates to independence and non-branching equates to dependence.

Example 1: Dependency

(W_1)capacitor - reservoir for electrons

(W_2)reservoir - reservoir for water

(c)reservoir

4 Prepositional Paraphrase Interpretation

4.1 Post ad Pre-modification (Right and Left-bracketing)

Contrary to left and right-bracketing schemes (Pustejovsky, 1993), the current work assumes right and left bracketing to be preposition and object-dependent. Maintaining inherent in every preposition is a mapping (linking) orientation. In other words, given the framework head(modifier), and the noun compound $S=(Adj_1)N_1P(Adj_2)N_2$, S is usually pre-modified ($N_1(N_2)$) or post-modified ($N_2(N_1)$).^{1 2 3} In the case of *of*, the mapping orientation may be altered based on whether the modifier is a member of the object.

Condition 1: For $P=of$, from

¹ N_1 represents the first noun and all its synonyms in a prepositional paraphrase

² N_2 represents the second noun and all its synonyms in a prepositional paraphrase

³*because* is a special case as demonstrated in condition 3

S is mapped through either $N_2(N_1)$

Note: For *of* $N_2(N_1)$ is used only when N_1 is an element of N_2

Condition 2: For **P**=*at, in, of, to, with, without, during, before, after, outside of*

S is mapped through $N_1(N_2)$ or in general $N_{i-1}(N_1)$

Condition 3: For **P**=*between*

When mapped, *between* is decomposed into *versus* and the objects of *between* or two instances of *betweens* object.

a. between N_2 and $N_3 \rightarrow N_2$ versus N_3 .⁴

b. between $N_2 \rightarrow N_2$ versus N_2

c. between N_2 and $N_3 \rightarrow N_1(N_2, N_3)$, where N_2 and N_3 are on the same ordered set

The employment of the three conditions are demonstrated below.

(1.a) $[N_1P_1N_2]$ conflict with violence \rightarrow fight

(1.b) $[AdjN_1N_2]$ armed conflict with violence \rightarrow war

(2) $[N_1P_1N_2]$ trial without verdict \rightarrow mistrial

In a more complicated example, synonyms may be employed.

(3) *development of Persian lexical units \rightarrow development of Persian words which ultimately becomes Persian etymology*

Ambiguous Subclass reference Phrases such as *sort of, kind of, variety of, form of, semblance of, type of*, do not map unless they are modifiers of a classified, head noun. The aforementioned phrases are examples of ambiguous subclass references. In essence, they refer to a random choice of subclasses. The word structure CAR and MUSIC would be considered classified word structures since their subclasses (i.e. *type of car, type of music*) are also termed *model* and *genre*, respectively.

Number of Number of N_2 : does not normally result in a mapping, since *number of* refers to an unspecified number of N_2 . However *number of* can signify an enumeration by the presence of a periodic as opposed to quantifiable N_2 . *Time* and *chance* cannot be quantified, hence they are both examples of periodic nouns. In combination with number

of, the prefix *re-* can often be used when there is a periodic number of mapping.

(4.a) *The number of times Ive told you is too many! I wont tell you again!*

(4.b) *My reiterations are too many! I wont tell you again.*

In the proceeding analysis, it is important to note that each preposition is multi-valued due to its numerous connotations. These connotations are sometimes identified through definite or indefinite articles (*the, a, an*, etc) that may accompany prepositions (e.g. *of the, by the, beyond a*, etc.), but the case of articles will not be addressed in the present work. Also neglected is the case of hyphenated noun compounds (*blue-green, player-coach, member state, member-state*) and compound nouns (*hometown, nightclub, night club*).

4.2 Cascading Prepositional Paraphrases and Noun Compounds

A cascading noun compound ($n_1n_2 \dots n_i$) comprises a series of three or more nouns. A pairing of any two nouns (e.g. $[n_1n_2], [n_3n_1], [n_2n_3]$, etc.) is termed a constituent noun compound. In the analysis of cascading noun compounds, the rightmost constituent ($n_{i-1}n_i$) acts as the head. As previously mentioned, the mapping of each constituent is based upon the inherent mapping orientation(s) of the preposition. In summary, this technique mimics the dependency compounding algorithm of Lauer(1994), but differs in that the acceptance is based on membership; i.e. determining whether an object contains a modifier as a member.

(5.a) $[Adj_1N_1P_1N_2P_2N_3]$ armed conflict with violence between countries

(5.b) country versus country armed conflict with violence

(5.c) civil armed conflict with violence

(5.d) civil war

⁴ N_3 represents the third noun and all its synonyms in a prepositional paraphrase

Item	noun	prepositional paraphrase
1	engine	converter
2	transducer	converter of energy
3	electric motor	electric to mechanical converter of energy

Table 1: Converter.

Item	noun	prepositional paraphrase	
1	antenna	receiver of electromagnetic energy	electromagnetic receiver
2	antenna	radiator of electromagnetic energy	electromagnetic transmitter
3	antenna	transceiver of electromagnetic energy	electromagnetic transceiver

Table 2: Converter.

5 Prepositional Paraphrase Construction

The basic tenets which govern the construction of the present corpus are simple and comprise two major steps.

1. nominalization of the function of a concept; when possible
2. use of any preposition and a noun to modify the nominalization

As noted by (Downing, 1977) the *Animal* class, which corresponds to the hypernymy relation of Girju (2006, 2007), is distinct in that function is rarely the basis for modification.⁵ Instead, appearance and habitat dictate modifier choice. Therefore this procedure does not pertain to the *Animal* class.

Without hesitation, we observe a sample construction.

Misnomers: What are misnomers? Misnomers are by default multi-valued.

Famously, Downing's example *Please sit in the apple juice seat* derives its confusion from the fact that *apple juice* is an accepted compression of *container of apple juice*. This in turn prevents *apple juice* from being a true function since it is multi-valued. Employing with and of from our prepositional set, *apple juice seat* would become seat with the container of apple juice. The apple-juice seat problem provides a practical example presented by multi-valued naming

⁵The plant, Virginia creeper *Parthenocissus quinquefolia* is a rare exception

conventions. Other examples are plentiful in technical fields.

Consider *antenna*, *RADAR*, and *frequency modulation*, *NMR* (Nuclear Magnetic Resonance or Noun Modifier Relationship). These examples range from being either ambiguous to semantically empty, making it impossible for an engine or a random observer to reliably and non-empirically interpret the concepts they represent. Even worse, this may result in redundancy or contradiction.

Example 2:

First consider *antenna*, which is semantically empty, since the physical description of an antenna as a long cylindrical device has long since been antiquated. WordNet gives the following definition. *antenna: an electrical device that sends or receives radio or television signals*⁶

The problems with this definition are numerous: Anything that sends radio or television signals is by default an electrical device.⁷ Can an antenna receive and send signals? Can an antenna send or receive radio and television signals?

Using , the previously mentioned naming convention an antenna can be described as a transmitter or receiver. Television and radio signals are both electromagnetic waves, therefore an antenna is a transmitter of electromagnetic energy, receiver of electromagnetic energy, or in the case where an antenna transmits and receives signals a transceiver of electromagnetic energy - transceiver is used for lack of a better word. Thus, antenna is actually a culmina-

⁶<http://wordnet.princeton.edu>

⁷{n: electrical device} a device that produces or is powered by electricity. <http://wordnet.princeton.edu>

Item	Noun	Prepositional Paraphrase
1	General	Commander of group of regiments (division)
2	Brigadier	Commander of brigade (group of regiments)
3	Colonel	Commander of regiment (group of battalions/squadrons)
4	Lieutenant Colonel	Commander of battalion/squadron (group of companies)
5	Major	Commander of company (group of platoons)
6	Captain	(Substitute) commander of company (group of platoons)
7	Lieutenant	Commander of platoon (group of squads)

Table 3: Military Rank.

Item	Noun	Prepositional Paraphrase	Noun Compound
1	Piston	reciprocator	reciprocator
2	Spark plug	ignition by intermittence	intermittent igniter
3	Fuel injector	injector of fuel	fuel injector
4	Cylinder	cavity of compressor	compressor cavity
5	Combustion chamber	chamber for combustion	combustion chamber
6	Camshaft	coordinator for valves (intake and exhaust)	valve coordinator
7	Crankshaft	coordinator of reciprocator	reciprocator coordinator
8	Cylinder and piston	compressor	compressor
9	Timing belt	coupler for coordinators	coordinator coupler
10	Alternator	generator of AC Electricity	generator

Table 4: The Automobile.

tion of the three different concepts mentioned above, hence it is multi-valued due to its reliance on resemblance as a semantic relation.

89

5.1 Problems

Problem 1:

In *Lexical Semantics*, Cruse (1986) considers the possibility of forming a branching representation of military rank.¹⁰ The following proof shows that a representation must either be non-branching (dependent) or branching (independent), and can not be both.

⁸There will be slight variations of the definitions of some fields. For example variations of *field marshal* are primarily used in European countries

⁹www.easternct.edu/personal/faculty/pocock/ranks.htm

¹⁰*It is interesting to speculate on how a set of terms could be devised for military personnel which formed a branching hierarchy*

If $A_1 \cap A_2 \neq \emptyset, \exists$ some element a Assume $A_1 \cap A_2$ are independent then, $A_1 \cap A_2 = \emptyset$ Since $a \neq \emptyset$, the independence of $A_1 \cap A_2$ implies $A_1 \cap A_2$ is not dependent and vice versa.

A non-branching (dependent) representation of military rank is provided in Table 3 to prove independence using the *commander* class.

Problem 2: The automobile. As shown in Table 4, the prepositional paraphrase of common nouns and compound nouns can be exploited to offer intuitive naming conventions.

Problem 3: Pills. An example used to show the limitations of the eight prepositions of Lauer(1995) was the headache pill, fertility pill example (O Seaghdha, 2008). By using *against* and *for*, *headache pill* can be interpreted as *pill against headache* while *fertility pill* can be interpreted as *pill for fertility*. Although not practiced in this case, ideally the head (*pill*) will be named after its function.

Noun	Conceptual PP	Noun Compound
Cavity magnetron	generator of frequency by resonance	resonant frequency generator
Diode	conductor of only positive current	positive-bias conductor
Voltage	motion of electricity by force	electromotive force, potential (difference)
RADAR	location by radio waves	radiometric location
RF Amplifier	amplifier of potential	potential amplifier
Mixer	-	frequency adder or subtracter
AGC	regulator of potential	potential regulator
STALO	regulator of frequency	frequency regulator
IF amplifier	amplifier of potential	potential amplifier
IF filter	extractor of intermediate frequency	intermediate frequency extractor
IF limiter	limiter of potential	potential limiter

Table 5: RADAR.

Noun	Conceptual PP	Noun Compound
Inverter	converter of DC to AC electricity	DC-AC converter
Rectifier	converter of AC to DC electricity	AC-DC converter
Stranded cost recovery	reimbursement for commoditization	commoditization reimbursement
Rain garden	garden with diverter of rain	-
Step-up Transformer	amplifier of potential	potential amplifier (potentiator)
Step-down Transformer	attenuator of potential	potential attenuator
Panelboard	-	electricity allocator

Table 6: Miscellaneous Terminology.

Problem 4: RADAR. The inclusions of tables 5 and 6 deem it necessary to divide prepositional paraphrases (PP) into interpreted (based on the noun) and conceptual (based on the function) prepositional paraphrases.

6 Conclusion

Given a minimally supervised environment, the present paper has shown the necessity of not only defining the semantic relations unambiguously, but also the necessity of defining nouns comprising the noun compound unambiguously. Following this result, efforts were made to show how to create the critical head and modifier nouns.

While these efforts are preliminary, examples were provided to show existence of a naming convention applicable to technical terminology. The limitations present relate to demonstrating the uniqueness of this scheme and its general

applicability to a concept with more than one function (e.g. mitochondrion) and answering whether there are exceptions to the *Animal* class naming convention noted by (Downing, 1977).

Acknowledgments

Do not number the acknowledgment section.

References

- David A. Cruse. 1986. *Lexical Semantics*. Cambridge Press.
- Tim Finin. 1981. The semantic interpretation of nominal compounds. *In Proceedings of the 1st National Conference on Artificial Intelligence Journal (AAAI-80)*. Stanford, CA
- Roxana Girju. 2006. Out-of-context noun phrase semantic interpretation with cross-linguistic evidence. *In Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM-06)*, Arlington, VA.

- Roxana Girju. 2007. Experiments with an annotation scheme for a knowledge-rich noun phrase interpretation system. In *Proceedings of the ACL-07 Linguistic Annotation Workshop*, Prague, Czech Republic.
- Pierre Isabelle. 1984. Another look at nominal compounds. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING-84)*, Stanford, CA.
- Christiane Fellbaum.editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Vivi Nastase and Stan Szpakowicz. 2003. *Exploring noun-modifier semantic relations*. In Proceedings of the 5th International Workshop on Computational Semantics(IWCS-03).Tilburg, The Netherlands.
- Mark Lauer. 1994. *Conceptual Association for Compound Noun Analysis*. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Student Session, Las Cruces, N.M.
- Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds.*, Ph.D. thesis. Department of Computing Macquarie University NSW 2109 Australia.
- Judith N. Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press.
- Vivi Nastase and Stan Szpakowicz. 2003. *Exploring noun-modifier semantic relations*. In Proceedings of the 5th International Workshop on Computational Semantics(IWCS-03).Tilburg, The Netherlands.
- Diarmuid O Seaghdha. 2008. *Learning compound noun semantics. Technical Report Number 735, UCAM-CL-TR-735, ISSN 1476-2986. Cambridge, United Kingdom*
- James Pustejovsky and Anick Bergler. 2003. *Lexical semantic techniques for corpus analysis*. *Computational Linguistics*, 19(2):331358.
2009. <http://wordnet.princeton.edu/perl/webwn?s=antenna>.