Semi-automatic compound nouns annotation for data integration systems (Extended abstract) ⋆

Sonia Bergamaschi and Serena Sorrentino ⋆⋆

DBGROUP, DII
University of Modena and Reggio Emilia, Italy
name.surname@unimore.it

Abstract. Lexical annotation is the explicit inclusion of the "meaning" of a data source element according to a lexical resource. Accuracy of semi-automatic lexical annotator tools is poor on real-world schemata due to the abundance of non-dictionary compound nouns. It follows that a large set of relationships among different schemata is discovered, including a great amount of false positive relationships. In this paper we propose a new method for the annotation of non-dictionary compound nouns, which draws its inspiration from works in the natural language disambiguation area. The method extends the lexical annotation module of the MOMIS data integration system.

1 Introduction

The focus of data integration systems is on producing a comprehensive global schema successfully integrating data from heterogeneous structured and semi-structured data sources (heterogeneous in format and in structure) [11, 8, 4]. Therefore, it is important to deal with labels of schemata , i.e. to understand the "meaning" behind the names denoting schemata elements.

Lexical annotation is the explicit inclusion of the "meaning" (synset/sense in WordNet (WN) terminology [15]) of a data source element (i.e. class/attribute name) w.r.t. a thesaurus (WN in our case).

The fundamental peculiarity of a thesaurus, like WN, is the presence of a wide network of semantic relationships between words and meanings. The disadvantage in using a lexical resource is that it does not cover with the same detail different domains of knowledge and many domain dependent terms, say *non-dictionary words*, may not be present in it. Non-dictionary words include compound nouns, acronyms etc. In this work, we will concentrate only on non-dictionary Compound Nouns (CNs).

⋆ **Acknowledgements:** This work was partially supported by MUR FIRB NeP4B (Network Peer for Business project) http://www.dbgroup.unimo.it/nep4b.

⋆⋆ PHD student of the ICT Doctorate School at the University of Modena and Reggio Emilia

In a CN two or more words (in the following called *constituents*) are used to denote a concept. Although CNs are frequently used, both in natural language and in structured and semi-structured data sources, they usually do not have an entry in WN (or other lexical resources). Thus, the result of lexical annotation is strongly affected by the presence of these non-dictionary CNs in the schema. Few works in literature face the problem. In the approach presented in [19] the constituents of a CN are treated as single words: for example the CN "teacher judgment" is split into two tokens ("teacher" and "judgment") and its relatedness to other sources element is calculated as an average over the relatedness between each token and the other element. It follows that a large set of relationships among different schemata is discovered, including a great amount of false positive relationships (as shown in figure 3-a).

Starting from our previous works on lexical annotation of structured and semi-structured data sources [5], we propose a semi-automatic method for the lexical annotation of non-dictionary CNs by creating a new WN meaning.

Our method is implemented in the MOMIS (Mediator envirOnment for Multiple Information Sources) system. However, it may be applied in general in the context of schema mapping discovery, ontology merging and data integration system.

The rest of the paper is organized as follows: in section 2 we present our method for CNs annotation; section 3 describes related works; in section 4 we present experimental results, finally section 5 is devoted to conclusion and future work.

2 Compound Noun annotation

Definition 1.*A CN is a word composed of more than one words called CN constituents. It is used to denote a concept, and can be interpreted by exploiting the meanings of its constituents.*

Definition 2.*Annotation of a CN data source schema element label is the explicit assignment of its meaning w.r.t. a thesaurus.*

In order to perform semi-automatic CNs annotation a method for their interpretation has to be devised.

Definition 3.*The interpretation of a CN is the task of determining the semantic relationships among the constituents of a CN.*

In natural language disambiguation literature different CNs classifications are proposed [13, 12, 1]. In this paper, we choose to use the classification introduced in [13], (where CNs are divided in four categories: *endocentric*, *exocentric*, *copulative* and *appositional*) and to consider only endocentric CNs.

Definition 4.*An Endocentric CN consists of a head (i.e. the categorical part that contains the basic meaning of the whole CN) and modifiers, which restrict this meaning. A CN exhibits a modifier-head structure with a sequence of nouns composed of a* head *noun and one or more modifiers where the head noun occurs always after the modifiers.*

The constituents of endocentric compounds are noun-noun or adjective-noun, where the adjective derives from a noun (e.g. "dark room" where the adjective "dark" derives from the noun "darkness").

Our restriction is motivated by different elements: (1) the vast majority of CNs of schemata fall in endocentric category; (2) endocentric CNs are the most common type of CNs in English; (3)exocentric and copulative CNs, which are represented by a unique word, are often present in a dictionary; (4) appositional compound are not very common in English and less likely used as element of a schema.

We consider endocentric CNs composed of only two constituents, because CNs consisting of more than two words can be constructed recursively by bracketing them into pairs of words and then interpreting each pair. CNs which have an entry in WN (e.g. "travel agent" and "company name") will be treated as single words, while for CNs which do not have an entry in WN (non-dictionary CNs) we apply our annotation method.

Our method can be summed up into four main steps: (1) CN constituents disambiguation; (2) redundant constituents identification; (3) CN interpretation via semantic relationships; (4) creation of a new WN meaning for a CN.

2.1 CN constituents disambiguation

In this phase the correct WN synsets of each constituent are chosen in two steps:

1. *Compound Noun syntactic analysis*: this phase performs the syntactic analysis of CN constituents, in order to identify the syntactic category of its head and modifier. If the CN does not fall under the endocentric syntactic structure (noun-noun or adjective-noun where the adjective derives from a noun), it is ignored;
2. *Disambiguating head and modifier*: this phase is part of the general lexical disambiguation problem. By applying our CWSD (Combined Word Sense Disambiguation) algorithm [5], each word is semi-automatically mapped into its corresponding WordNet 2.0 synsets.

As shown in figure 1-a, for example, for the CN "teacher judgment" we obtain the two constituents annotated with the correspondent WN meanings (i.e. "$teacher_{\#}1$" and "$judgment_{\#}2$").

2.2 Redundant constituents identification and pruning

During this phase we control if a CN constituent is a *redundant word*. Redundant words are words that do not contribute new information as their semantics contribution can be derived from the schema or from the lexical resource. For example, usual in database integration, is when the name of a class (e.g. "company") is reported on the name of one of its attribute (e.g. "company name"): the constituent class name is not considered because the relationship holding among a class and its attributes can be derived from the schema.

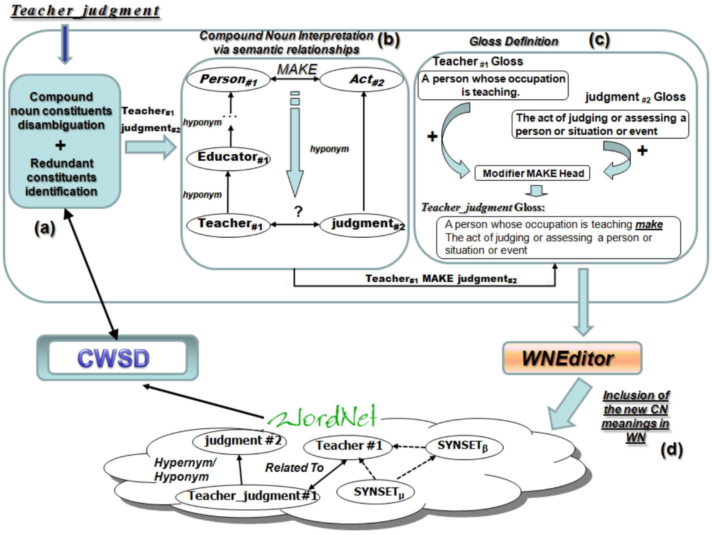

Teacher_judgment
Compound Noun Interpretation via semantic relationships
(b)
Gloss Definition
(c)
Compound noun constituents disambiguation
+
Redundant constituents identification
(a)
Teacher#1
judgment#2
Person#1
MAKE
Act#2
hyponym
Educator#1
hyponym
hyponym
?
Teacher#1
judgment#2
Teacher #1 Gloss
A person whose occupation is teaching.
judgment #2 Gloss
The act of judging or assessing a person or situation or event
+
+
Modifier MAKE Head
Teacher_judgment Gloss:
A person whose occupation is teaching make The act of judging or assessing a person or situation or event
Teacher#1 MAKE judgment#2
CWSD
WNEditor
Inclusion of the new CN meanings in WN
(d)
WordNet
judgment #2
Teacher #1
SYNSETβ
Hypernym/ Hyponym
Related To
Teacher_judgment#1
SYNSETμ

Fig. 1. The CNs annotation process.

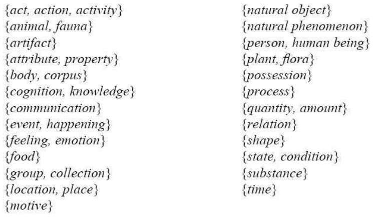
2.3 CN interpretation via semantic relationships

This phase concerns selecting among a set of predefined relationships the ones that best capture how the meanings of head and modifier are related. The set of semantic relationships to be considered for the interpretation of CNs, is a problem widely discussed in the natural language disambiguation literature [9, 16, 17]. In [12], Levi defines a set of nine possible semantic relationships to interpret CNs: CAUSE (“flu virus”), HAVE (“college town”), MAKE (“honey bee”), USE (“water wheel”), BE (“chocolate bar”), IN (“mountain lodge”), FOR (“headache pills”), FROM (“bacon grease”) and ABOUT (“adventure story”). On the contrary, Finin in [7] claimed an unlimited number of semantic relationships. In [13] the problems of relationships set is sidestep: the semantics of a CN is then simply the assertion of an unspecified relation between its constituents.

We choose the Levi semantic relationships set, as it is the best choice in the simplified context, w.r.t. natural language, of data integration. According to [6], our method is based on the following assumption:

Definition 1. *The semantic relationship between a head and its modifier of a CN is derived from the one holding between their top level WN nouns in the WN hierarchy*

The WN noun hierarchy has been proven to be very useful in the CNs interpretation task [17, 2]. The top level concepts of the WN hierarchy are the 25 *unique beginners* (shown in figure 2) for WN English nouns defined by Miller in [15]. These unique beginners were selected after considering all the possible adjective-noun or noun-noun combinations that could be expected to occur and are suitable to interpret noun-noun or adjective-noun CNs as in our case.



{act, action, activity}
{animal, fauna}
{artifact}
{attribute, property}
{body, corpus}
{cognition, knowledge}
{communication}
{event, happening}
{feeling, emotion}
{food}
{group, collection}
{location, place}
{motive}
{natural object}
{natural phenomenon}
{person, human being}
{plant, flora}
{possession}
{process}
{quantity, amount}
{relation}
{shape}
{state, condition}
{substance}
{time}

Fig. 2. The 25 unique beginners for WN nouns.

For each possible couple of unique beginners we associate the relationship from the Levi's set that best describes the meaning of the couple. For example, for the unique beginner couple "person and act" we choose the Levi's relationship MAKE (e.g. "person MAKE act"), that expresses that a person performs an act. Thus, as shown in figure 1-b, we interpret the CN "teacher judgment" by the MAKE relationship because "teacher" is an hyponym of "person" and "judgment" is an hyponym of "act".

Our method requires an initial human intervention to associate to each couple of unique beginners the right relationship. It may be considered acceptable, when compared with the much greater effort required for other approaches based on the pre-tagged corpus where the number of CNs to be annotated is much higher [16–18, 10].

The method is independent by the domain under consideration and can be applied to lexical resources providing as WN a wide network of hyponym/hypernym relationships between meanings.

2.4 Creation of a new WN meaning for a CN

During this phase, we create a new WN meaning for a CN; it can be divided in two prevalent steps:

1. *Gloss definition*: during this step we create the gloss to be associated to a CN, starting from the relationship associated to a CN and exploiting the glosses of the CN constituents. Figure 1-c shows an example of this phase. The glosses of the constituents "teacher" and "judgment", are joined according to the relationship MAKE.
2. *Inclusion of the new CN meaning in WN*: the insertion of a new CN meaning in the WN hierarchy implies the definition of its relationships with the other WN meanings. As the concepts denoted by a CN are a subset of the concepts denoted by the head we assume that a CN inherits most of its semantic from its head [13]. Starting from this consideration, we can infer that the CN is related, in the WN hierarchy, with its head by an hyponym relationship. Moreover, we represent the CN semantics related to its modifier by inserting a generic relationship RT (*Related term*), corresponding to WN relationships as *member meronym*, *part meronym* etc. However, the insertion of this two

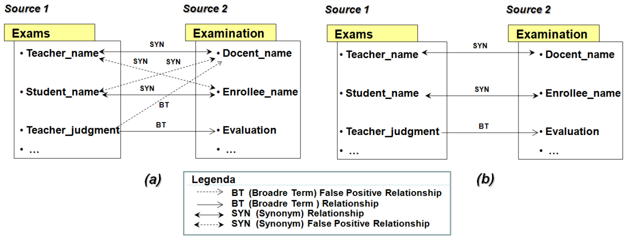


Fig. 3. Relationships discovered considering the CNs constituents as single words (a) and with our method (b).

relationships is not sufficient; it is necessary to discover also the relationships of the new inserted with the other WN meanings. For this purpose, we use the WNEditor tool to create/manage the new meaning and to set relationships between it and WN ones [3]. WNEditor automatically retrieves a list of candidate WN meanings sharing similarities with the new meaning. Then, the user is asked to explicitly declare the type of relationship (hyponymy, meronymy and so on) to relate the new meaning to another, if any. Figure 1-d shows an example of this step.

A CNs unified semantic interpretation is fundamental in data integration/ontology mapping discovery: by considering the CN constituents as single words we assign an independent meaning to each constituent (e.g. "teacher judgment" is annotated with two meanings, one for "teacher" and one for "judgment"). In this way the CN will be related to several "semantically distant" elements in the schema. It follows that the discovered relationships among the elements of different schemata is a large set including most of false positive relationships. Figure 3 shows, the semantic relationship between two schemata with/without our method.

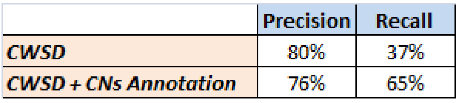
3 Related works

Interpreting CNs has received much attention in different areas, such as, machine translation, information extraction and applications as question answering.

Many works in literature involve costly pre-tagged corpus and heavy manual intervention [16–18, 10]. These approaches are based on a statistic co-occurrence of a relationship r between two words on corpus that contain different CNs manually labeled with the right semantic relationship. Moreover, there are other two prevalent problems with corpus-based methods: (1) in these approaches, there has been some underlying assumption in terms of domain or range of interpretations; this leads to problems in scalability and portability to novel domains; (2) there is a trade-off between how much training data (pre-tagged corpus) are used and the performance of the method. According with [6], we claim that the cost of acquiring knowledge from manually tagged corpus for different domains may overshadow the benefit of interpreting the CNs.

On the contrary, in the context of data integration and schema mapping only a few paper address the problem of CNs interpretation. In [19] a preliminary CNs comparison for ontology mapping is proposed. This approach suffers of two main problems: first, they start from the assumption that the ontology entities are accompanied with comments that contain words that express the relationship between the constituents of a CN; second, it is based on a set of rules manually created, that it is not general but is created on a set of specific keywords that are not necessarily present in comments.

The well know CUPID algorithm [14], during the schema elements normalization phase, considers abbreviations, acronyms, punctuation, etc. but not the problem of CNs interpretation.



	Precision	Recall
CWSD	80%	37%
CWSD + CNs Annotation	76%	65%

Fig. 4. Comparing the result of lexical annotation performed by CWSD with/without our CNs annotation.

4 Experimental Results & Conclusion

We implemented our method for CNs annotation in the MOMIS system. CNs annotation is performed during the lexical knowledge extraction phase: during this phase each schema element of a local source is semi-automatically annotated by the CWSD algorithm. We experimented our method over a real data sources environment which includes three sources of an application scenario (ICT-A partner search [1]) of the NeP4B project, with a global amount of 491 schema elements. These sources are particularly suitable to test our method, because they contain a lot of CNs. The annotation results have been evaluated in terms of recall (the number of correct annotations divided by the total number of schema elements) and precision (the number of correct annotations divided by the total number of annotations). The table in figure 4, shows the result of lexical annotation performed by CWSD with/without our CNs annotation method. Without CNs annotation, CWSD obtains a very low recall value, because a lot of CNs are present in the sources. The application of our method, together with CWSD permits to increase the recall without significantly worsening precision. However, the recall value is not very high; this is caused by the presence among sources of a lot of acronym terms. During the evaluation process, a CN has been considered correctly annotated if the Levi's relationship selected by the user is the same returned by our method.

[1] The schema of the scenario sources (in XML format) can be found at *www.dbgroup.unimo.it/nep4b/NeP4BScenarioICTA.xml*

The experimental results showed the effectiveness of our method, which significantly improves the result of the lexical annotation process. Future work will be devoted to investigate on the role of the set of semantic relationships chosen for the CNs interpretation process. Moreover, we will investigate on the problem of acronyms and abbreviations expansion.

References

1. K. Barker and S. Szpakowicz. Semi-automatic recognition of noun modifier relationships. In *COLING-ACL*, pages 96–102, 1998.
2. L. Barrett, A. R. Davis, and B. J. Dorr. Interpretation of compound nominals using wordnet. In *CICLing*, pages 169–181, 2001.
3. D. Beneventano, S. Bergamaschi, F. Guerra, and M. Vincini. Synthesizing an integrated ontology. *IEEE Internet Computing*, 7(5):42–51, 2003.
4. S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Record*, 28(1):54–59, 1999.
5. S. Bergamaschi, L. Po, and S. Sorrentino. Automatic annotation for mapping discovery in data integration systems. In S. Gaglio, I. Infantino, and D. Saccà, editors, *SEBD*, pages 334–341, 2008.
6. J. Fan, K. Barker, and B. W. Porter. The knowledge required to interpret noun compounds. In *IJCAI*, pages 1483–1485, 2003.
7. T. W. Finin. The semantic interpretation of nominal compounds. In *AAAI*, pages 310–312, 1980.
8. A. Halevy. Data integration: a status report. In *Proceedings of the German Database Conference, BTW-03*, Leipzig, Februar 2003.
9. S. N. Kim and T. Baldwin. Automatic interpretation of noun compounds using wordnet similarity. In *IJCNLP*, pages 945–956, 2005.
10. M. Lapata. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388, 2002.
11. M. Lenzerini. Data integration: A theoretical perspective. In L. Popa, editor, *PODS*, pages 233–246. ACM, 2002.
12. J. N. Levi. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York, 1978.
13. R. Lieber. *Morphology and Lexical Semantics*. Cambridge University Press, Cambridge, 2004.
14. J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with cupid. In *VLDB*, pages 49–58, 2001.
15. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.
16. D. Moldovan, A. Badulescu, M. Tatu, D. Antohe, and R. Girju. Models for the semantic classification of noun phrases. In *In HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 60–67, 2004.
17. V. Nastase, J. Sayyad-Shirabad, M. Sokolova, and S. Szpakowicz. Learning noun-modifier semantic relations with corpus-based and wordnet-based features. In *AAAI*, 2006.
18. B. Rosario. The descent of hierarchy, and selection in relational semantics. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, 2002.
19. X. Su and J. A. Gulla. Semantic enrichment for ontology mapping. In *NLDB*, pages 217–228, 2004.