Paraphrasing of Synonyms for a Fine-grained Data Representation

Svetla Koeva
Institute for Bulgarian Language
Bulgarian Academy of Sciences
52 Shipchenski prohod Blvd.
Sofia, 1113 Bulgaria
svetla@dcl.bas.bg

ABSTRACT
The paper addressed the question how the paraphrasing of synonyms can be linked with a fine-gained ontology based data representation. Our challenge is to identify for a set of synonyms (including terms and multiword expressions) the best lexical paraphrases suitable for given contexts. Our hypothesis is that: i. the minimal context in which the paraphrasing can be validated is different for different (semantic) word classes; ii. paraphrasing is defined by patterns within the minimal context containing the synonym and its dependent. For each minimal context a different set of rules is defined with respect to the modifiers and complements the words are licensed for. The extracted dependency collocations are linked with the WordNet synonyms. With this we achieve two goals: to define the lexical paraphrases suitable for a given context and to augment available lexical-semantic resources with linguistic information (the dependency collocations in which synonyms are interchangeable).

Categories and Subject Descriptors
I.2. [A**rtificial Intelligence**]: Semantic Networks
J. [Computer applications]: Linguistics

General Terms
Languages

Keywords
semantics, synonymy, paraphrasing, dependency collocations

1. INTRODUCTION
Paraphrasing is used in many areas of Natural Language Processing – ontology linking, question answering, summarization, machine translation, etc. Paraphrasing between synonyms seems a relatively simple task, but in practice an automatic paraphrasing of synonyms might produce ungrammatical or unnatural sentences. The reason is that although there are many synonyms in any natural language, it is unusual for words defined as synonyms to have exactly the same meaning in all contexts in which they are used. In other words, the notion of absolute synonyms remains theoretical. The human knowledge about synonyms – words (and/or multiword expressions) denoting one and the same concept, and semantic relations such as hypernymy, meronymy, antonymy, etc., is encoded in the lexical-semantic network WordNet [16]. The following test for synonymy is applied to WordNet:

Two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value.

The test implies that the WordNet synonyms are cognitive (or propositional) synonyms [2]. Cognitive synonymy is a sense relation that holds between two or more words used with the same meaning in a given context in which they are interchangeable. For example, the pairs {*brain; encephalon*}, {*cry; weep*}, {*big; huge*} are cognitive synonyms. However, cognitive synonyms may differ in their collocational range which means that their interchangeability is restricted. For example the words *educator*, *pedagogue*, and *pedagog* are synonyms linked in the WordNet with the definition '*someone who educates young people*'. In the collocation with the word *certified* most preferred is the word *educator* (*certified educator*), followed by *pedagogue*, while the word *pedagog* is most rarely used. In the collocation *Microsoft certified educator* the word *educator* would not be replaced with either of the words *pedagogue* or *pedagog*. The absolute synonymy is a symmetric relation of equivalence. However, the definition of synonymy as a substitution of words in a given context alternates the meaning of the equivalence relation [16]:

If x is similar to y, then y is similar to x in an a equivalent way.

We focus on WordNet because it is a hand crafted (or hand validated) lexical-semantic network and ontology and offers a large network of concepts and named entities along with an extensive multilingual lexical coverage. In this paper we present a pattern based method for identification of dependency collocations (a pair of grammatically dependent words that co-occur with more frequency than random) in which two words are interchangeable. The difference between grammatical and lexical collocations is pointed out by many researchers. We introduce the notion of ***dependency collocation*** which subsumes grammatical and lexical collocations and adds the condition for a grammatical dependence (such as subject, complement, and modifier) between collocates.

WordNet, together with other semantic resources such as YAGO[1], OpenCyc[2], DBpedia[3], etc., is part of the Linguistic Linked Open Data cloud [1]. Our aim is twofold: to define the lexical paraphrases suitable for a given context and to augment available

[1] http://www.mpi-inf.mpg.de/departments/databases-and-informationsystems/research/yago-naga/yago/

[2] http://www.opencyc.org

[3] http://wiki.dbpedia.org/about

lexical-semantic resources with linguistic information (the dependency collocations in which given words are synonyms).

2. RELATED WORK

There are various attempts to extract automatically candidates for a paraphrase based on the Distributional hypothesis, which states that words occurred in the same contexts tend to have similar meanings [6]. Differences in the approaches can be viewed mainly with respect to the restrictions on the contexts [9]: some approaches (for example, grouping similar terms in document classification) consider all words in a document, others (focused on extracting of semantic relations like synonymy) may take words in a predefined window or extract words in a specific syntactic relation to the target word. Ruiz-Casado et al. [20] label a pair of words as synonyms if the words appear in the same contexts, but this simple approach in many cases might link also hypernyms, hyponyms, antonyms, etc. Semantic relations such as *purpose, agent, location, frequency, material*, etc. are assigned to noun-modifier pairs based on semantic and morphological information about words [17, 18].

Experiments were performed with decision trees, instance-based learning and Support Vector Machines. Turney and Littman [21] and Turney [22] use paraphrases as features to analyze noun-modifier relations. The hypothesis, corroborated by the reported experiments, is that pairs which share the same paraphrases belong to the same semantic relation. Lin and Pantel [14] measure the similarity between paths in dependency trees assuming that if two dependency paths tend to link the same sets of words (for example, *commission, government* versus *crisis, problem*) the meanings of the paths are similar and the words can be paraphrased (for example, *finds a solution to* and *solves*). Padó and Lapata [19] take into account context words that stand in a syntactic dependency relation to the target word and introduce an algorithm for constructing semantic space models. They rely on three parameters which guide model construction: which types of syntactic structures contribute towards the representation of lexical meaning; importance weighs of different syntactic relations; and the representation of the semantic space (as cooccurrences of words with other words, words with parts of speech, or words with argument relations such as subject, object, etc.). Heylen et al. [10] compare the performance of models using a predefined context window and those relying on syntactically related words and show that the syntactic model outperform the other models in finding semantically similar nouns for Dutch. Ganitkevitch et al. [3] extracted a Paraphrase Database using the cosine distance between vectors of distributional features applied on parallel texts.

Hearst [7] introduces lexico-syntactic patterns (for example, X such as Y) in the task for automatic identification of semantic relations (hypernymy and hyponymy). Several techniques aim at providing support for the automatic (or semi-automatic) definition of the patterns to be used for extraction of semantic relations. Hearst [8] proposes to look for co-occurrences of word pairs appearing in a specific relation inside WordNet. Maynard et. al. [15] discuss the use of information extraction techniques involving lexico-syntactic patterns to generate ontological information from unstructured text. Several approaches combine distributional similarity and lexico-syntactic patterns. Hagiwara et al. [5] describe experiments that involve training various synonym classifiers. Giovannetti et al. [4] detect semantically related words combining manually composed patterns with distributional similarity. Turney [23] proposes a supervised machine learning approach for discovering synonyms, antonyms, analogies and associations, in which all of these phenomena are subsumed by analogies. The problem of recognizing analogies is viewed as the classification of semantic relations between words.

The approach proposed here aims at the extraction of collocations in which synonyms occur and interchange and towards the generalization of the shared contexts.

3. PATTERN BASED APPROACH FOR DEPENDENCY COLLOCATIONS

The synonymy in WordNet is limited to a certain set of contexts and cannot be directly applied for automatic paraphrasing. For example the words *car*, *automobile* and *auto* from the synonymous set {*car*; *auto*; *automobile*; *machine*} with a definition '*a motor vehicle with four wheels; usually propelled by an internal combustion engine*' can be interchanged in the collocations with the word *luxury – luxury car*, *luxury automobile*, *luxury auto*, *luxury machine*, with the prepositional phrase *with lights – car with lights*, *auto with lights*, *automobile with lights*, *machine with lights*, and so on. On the other hand, it is hard to find examples in which the word *car* from the collocation *car cash market* is replaced by words *auto*, *automobile* or *machine*.

Our challenge is to identify for a set of synonyms the best lexical paraphrases suitable for given contexts. We accept the view that the meaning of words is expressed through their relations with other words and each word selects the set of semantic word classes with which it can express a specific meaning. For example, the word *director* and the word *professor* are similar in the way they designate the concept for a person, and this determines the fact that both nouns can co-occur with adjectives denoting height, age, etc. The subsets of adjectives that can collocate with the two words differ with respect to their meaning, and not all adjectives that are compatible with one noun are compatible with other as well (*chief executive officer*, ?*chief executive professor*). The meaning of the word *professor* also implies that it may be specified with expressions for disciplines as complements (*professor of physics*), while, in comparison, the word *director* may not. Both words can be specified for institutions through selecting the respective complements. Therefore, the closer the similarity between two words is the bigger is the number of the contexts which they share. Our hypothesis is that:

i. the minimal context in which the paraphrasing has to be identified is different for different word classes;

ii. paraphrasing is defined by patterns within the minimal context containing the synonym and its dependent (dependency collocations).

The minimal context for English involves different combinations of the following: adjectival modifier in pre-position, one or several, prepositional complement in post-position; and noun modifier in pre-position.

For adjectives the minimal context starts with the adjective (the target synonym) and ends with a noun modified by the adjective (for example ***new idea***, ***new*** *brilliant* ***idea***, ***fresh idea***, ***fresh*** *brilliant* ***idea,*** but not *New Idea Magazine*).

For nouns the minimal context is one of the following: an adjective modifier in the leftmost position and the head noun (the target synonym) at the right position; a noun modifier in the

leftmost position and the head noun (the target synonym) at the right position (for example ***gold light***, ***amber light***, but not *Gold Light Gallery*); the noun (the target synonym) in the leftmost position and a prepositional complement – a preposition and a noun at the right position (for example ***flood of requests***, ***torrent of abuse***).

For verbs the minimal context is one of the following: the verb (the target synonym) in the leftmost position and an object noun at the right position (for example ***compose music***, ***write music***, ***compose*** *nice* ***music***, *but not compose music online*); the verb (the target synonym) in the leftmost position, a preposition and an object noun at the right position (for example ***lies*** *in the* ***hands***, ***rests*** *in the* ***hands***, but not *rests in the hands of the United States Congress*).

The dependency collocations in our approach always contain the two constituents occupying the leftmost and rightmost position in the minimal context (in some cases linked with a preposition). The minimal context is defined by linguistic rules, which describe eligible constituents between the leftmost and rightmost position. The minimal contexts and the syntactic structures of dependency collocations are different for different languages. We have developed rules for Bulgarian and English but only rules for English are illustrated in this paper. More minimal contexts relevant for synonymy validation can be defined further, for example comprising coordinative constructions, subject verb dependencies, and so on.

4. IMPLEMENTATION

The rules are formulated within the linguistic formalism called Est and applied through the parser ParseEst [12]. The Est formal grammar is a regular grammar. The rules are abstractions for strings of words and do not define a hierarchical (linguistic) structure. An element in the rule can be a word, a lemma, a grammatical tag, and a lexicon. The boolean operators, the Kleene star and Kleene plus can be applied on the elements and on groups of elements. The formalism maintains unification and supports cascading application of rules by preset priority. Right and/or left context can be defined in a similar way, as a sequence of elements.

The rules have to exhaust all lexical and grammatical combinations and permutations. A given word can be specified by the class to which it belongs: lemma, part-of-speech and grammatical categories. For example, the part-of speech tag 'NC' defines common noun, the tag 'NCs' – singular common noun, the regular expression 'NC.' – singular and plural common noun, etc. The word permutations are expressed as different paths in the rules. For each minimal context, a different rule is defined with respect to the modifiers and complements the target word classes are licensed for. The rule (1) below matches a minimal context for a noun (only part of the rule is presented here).

(1)

```
<group>
        <e l="NOUN LEMMA"/>
        <e p="R"/>
        <star><e p="DT"/></star>
        <star><e p="A"/></star>
        <e p="NCs"/>
</group>
```

The rule says that the head noun can be modified by a prepositional phrase in post-position. The structure of the prepositional phrase is constrained to a preposition, zero or more determiners, zero or more adjectives, and a noun. This general rule is multiplied by replacing its element "NOUN LEMMA" with the WordNet synonyms, for example l="teacher" and l="instructor". Our approach makes use of handcrafted rules running on preliminary annotated texts with part-of-speech tags, tags for grammatical categories, and lemmas. Apache OpenNLP[4] with pre-trained models and Stanford Core-NLP5[5] are used for the annotation of the English texts – sentence segmentation, tokenisation, and POS tagging [13].

The rules are run on a corpus[6] [13] and match for a given pair of synonyms their minimal contexts, i.e. ***months*** *of investigation ENG2014348156n*, ***breaking*** *the longstanding political stalemate ENG2000351165v*, ***acute*** *pain ENG2000769157a*. For adjectives and verbs the target synonym is at the first position in the collocation. For nouns – either at the first or at the last position of the collocation. The collocations for different word classes are extracted from the minimal contexts as follows. For nouns: the first adjective and the last noun or the first noun, a preposition if any, a determiner, if any, and the last noun, i.e. *months of investigation*. For adjectives: the first adjective and the last noun, i.e. *acute pain*. For verbs: the first verb, a preposition, if any, a determiner, if any, and the last noun, i.e. *breaking the stalemate*. The results for the Princeton WordNet2.0 base concepts (PWN 2.0 BCS) are presented in Table 1.

Table 1. Number of Rules snd Collocations for PWN 2.0 BCS

	Nouns	**Verbs**	**Adj**	**Total**
Rules	4624	2997	70	7691
Collocations	223347	396434	5108	624889
Unique collocations	59877	73201	4528	137606

The lemmas of the dependent collocates and the information for the number of occurrences in a corpus are linked with the respective WordNet literals in the field LNote (a note related to a literal), as it is shown in (2)[7].

(2)

```
<SYNONYM><LITERAL>present<SENSE>2</SENSE>
      <LNOTE>proposal,2</LNOTE>
      <LNOTE>budget,1</LNOTE>
      <LNOTE>plan,2</LNOTE>
</SYNONYM>
```

Since the task is not a classification one a validation against an annotated corpus is not applicable. A validation is performed by an expert during the process of the developing of rules: every change within a rule has been checked against a certain number of matches.

[4] http://incubator.apache.org/opennlp/

[5] http://nlp.stanford.edu/software/corenlp.shtml

[6] The experiments are made on the monolingual parts of the Bulgarian-English parallel corpus: 280.8 and 283.1 million tokens respectively.

[7] The PWN2.0 enriched with collocations of synonyms is published at: http://dcl.bas.bg/wordnet_collocatons.xml

The pattern matching approach allows a focused extraction of dependency collocations – not all collocations are extracted but only those in which a particular dependency is expected. The rules are applied without prior word sense disambiguation. However, we consider that the focused use of different minimal contexts for different semantic word classes may lead to correct identification of collocations. Sometimes even humans cannot distinguish between hypernyms and hyponyms if their lemmas coincide. The approach allows the accumulation of information – in case some new rules are formulated or the existing rules are applied on different corpora.

5. CONCLUSION AND FUTURE WORK

To conclude, it is difficult to define synonymy taking into account all different ways in which synonyms may differ; to provide a reliable tests for identification of synonyms, and to calculate all possible contexts in which two words are synonyms. On the other hand, dependency collocations provide suitable contexts for paraphrasing with synonyms. This is a step towards an improving of intuitive definitions of synonyms and for a precise linking of the synonymous words and expressions with the contexts in which two or more words are interchangeable.

The dependency collocations consist of the head word lemma – a noun or a verb, and the dependent word lemma – an adjective or a noun, and provide information about the combinatory properties between particular semantic word classes. Each lemma, which is present in the WordNet structure, is classified into semantic primitives such as person, animal, plant, cognition, communication, etc. [16]. On the bases of the dependency collocations and the classification of semantic primitives different inferences can be calculated. For example, nouns for professions participate in the following collocational patterns, generalized for parts of speech and semantic primitives:

– (dependent adjective – (head noun denoting a profession, semantic primitive: noun.person)) (for example, *young engineer*, *blond professor*);

– ((head noun denoting a profession, semantic primitive: noun.person) – (dependent noun specifying a domain, semantic primitive: noun.cognition)) (for example, *director of theater*, *rector of university*);

– ((dependent noun specifying a domain, semantic primitive: noun.cognition) – (head noun denoting a profession, semantic primitive: noun.person)) (for example, *theater director*, *university rector*);

– ((head noun denoting a profession, semantic primitive: noun.person) – (dependent noun specifying an affiliation, semantic primitive: noun.group)) (for example, *teacher at university*, *instructor at school*).

Some WordNets, for example GermaNet, distinguish between semantic classes of adjectives, thus different semantic classifications might be further applied.

One of the main goals of our future work will be to apply WordNet based semantic classifications in order to obtain generalizations about combinatory preferences of words, in particular, to generate collocational patterns for WordNet synonyms. Further, the collocations can be extended by means of relatedness between two concepts in WordNet [11], possibly restricted to the direct hyponyms of the head collocate. Since the Princeton WordNet is converted to RDF/OWL[8], our future plans also include the conversion of the dependency collocations of the WordNet synonyms to RDF/OWL representation.

6. REFERENCES

[1] Chiarcos, C., Hellmann, S., Nordho, S. 2011.Towards a Linguistic Linked Open Data Cloud: The Open Linguistics Working Group. In TAL 52(3), 245–275.

[2] Cruse 1986: Cruse D. A. 1986. Lexical Semantics. Cambridge: Cambridge University Press.

[3] Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. 2013. PPDB: The paraphrase database. In Proceedings of NAACL-HLT. Atlanta, Georgia: Association for Computational Linguistics, 758–764.

[4] Giovannetti, E., Marchi, S. and Montemagni, S. 2008. Combining Statistical Techniques and Lexico-Syntactic Patterns for Semantic Relations Extraction from Text. In Proceedings of the 5th Workshop on Semantic Web Applications and Perspectives.

[5] Hagiwara, M. O. Y. and Katsuhiko, T. 2009. Supervised Synonym Acquisition using Distributional Features and Syntactic Patterns. In Information and Media Technologies 4(2), 558–582.

[6] Harris, Z. 1985. Distributional Structure. In Katz, J. J. (ed.) The Philosophy of Linguistics. New York: Oxford University Press. 26–47.

[7] Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France, 539-545.

[8] Hearst, M. A. 1998. Automated Discovery of WordNet Relations. Cambridge MA: MIT Press.

[9] Heylen, Kris; Peirsman, Yves; Geeraerts, Dirk. 2008. Automatic Synonymy Extraction: A Comparison of Syntactic Context Models. In LOT Occasional Series, 11:101–116.

[10] Heylen, K., Peirsman ,Y., Geeraerts, D., Speelman, D. 2008. Modelling word similarity: an evaluation of automatic synonymy extraction algorithms. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 3243v3249.

[11] Hirst, G., and St-Onge, D. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum, C. (ed.) WordNet: An electronic lexical database. Cambridge MA: MIT Press. 305–332.

[12] Karagiozov, Diman, Anelia Belogay, Dan Cristea, Svetla Koeva, Maciej Ogrodniczuk, Polivios Raxis, Emil Stoyanov and Cristina Vertan. 2012. i-Librarian – Free Online Library for European Citizens. In INFOtheca, no. 1, vol. XIII, May. BS Print: Belgrade. 27-43.

[13] Koeva, S., Stoyanova, I., Leseva, S., Dimitrova, T., Dekova, R., Tarpomanova, E. The Bulgarian National Corpus: theory and practice in corpus design. In Journal of Language Modeling, 1, 65–110.

[8] http://www.w3.org/TR/wordnet-rdf/

[14] Lin, D. and P. Pantel. 2001. Discovery of Inference Rules for Question Answering. Natural Language Engineering 7(4):343–360.

[15] Maynard D., A. Funk and W. Peters. 2009. Using Lexico-Syntactic Ontology Design Patterns for ontology creation and population. In Proceedings of ISWC Workshop on Ontology Patterns (WOP 2009), Washington, 36–52.

[16] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. 1990. Introduction to WordNet: An On-line Lexical Database. In International Journal of Lexicography, 3(4):235–244.

[17] Nastase, V., and Szpakowicz, S. 2003. Exploring noun modifier semantic relations. In Proceedings of IWCS 2003, 281–301.

[18] Nastase, V., J. S. Shirabad, M. Sokolova and S. Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In Proceedings of the 21st National Conference on Artificial Intelligence, Boston, Mass., 781–787.

[19] Padó, S. and M. Lapata. 2007. Dependency-based construction of semantic space models. In Computational Linguistics , 33(2):161–199.

[20] Ruiz-Casado, m., E. Alfonseca and P. Castells. 2005. Using context-window overlapping in Synonym Discovery and Ontology Extension. Proceedings of the International Conference. In Recent Advances in Natural Language Processing, RANLP-2005, Borovets, Bulgaria, 2005.

[21] Turney, P., and Littman, M. 2003. Learning analogies and semantic relations. Technical Report Technical Report ERB-1103. (NRC #46488), National Research Council, Institute for Information Technology.

[22] Turney, P. 2005. Measuring semantic similarity by latent relational analysis. In Proceedings of IJCAI 2005, 1136–1141.

[23] Turney, P. D. 2008. A Uniform Approach to Analogies, Synonyms, Antonyms and Associations. In Proceedings of the 22nd International Conference on Computational Linguistics, 905–912.