

Automated Discovery of Telic Relations for WordNet

Marco De Boni

Department of Computer Science,
University of York,
York YO10 5DD
United Kingdom

mdeboni@cs.york.ac.uk

Suresh Manandhar

Department of Computer Science,
University of York,
York YO10 5DD
United Kingdom

suresh@cs.york.ac.uk

Abstract

A method is presented for automatically extending WordNet with the telic relationships proposed in Pustejovsky's lexicon model. The method extracts telic relationships from WordNet glosses by first selecting a telic word through a pattern matcher aided by a part-of-speech tagger and then employing a word disambiguation module to select the specific meaning (synset) of the telic word. The method is shown to be fruitful, inferring a number of useful relationships.

1 Introduction

WordNet (Miller 1995; Fellbaum 1998) is a lexical database which organizes words into synsets, sets of synonymous words, and specifies a number of relationships such as hypernym, synonym, meronym which can exist between the synsets in the lexicon. The explicit relationships in WordNet do not exhaust (nor claim to exhaust) the set of possible relationships between words and there is scope for expansion and improvement. Pustejovsky (1995), for example, presents a model in which each lexical item in a dictionary would be characterized by an argument structure (specifying the number and type of arguments that a lexical item carries), an event structure (characterizing the event type of a lexical item and its internal structure), qualia structure (representing the different modes of predication

possible with a lexical item) and a lexical inheritance structure (identifying how a lexical structure is related to other structures in the dictionary). Each of these structures could then be used to infer a very complex (but also comprehensive) net of relationships between lexical items. Qualia structures, for example, would provide information on "constitutive" relationships (the relation between an object and its constitutive parts, e.g. material, weight, parts and component elements), "formal" relationships (that which distinguishes an object within a larger domain, e.g. orientation, magnitude, shape, dimensionality, colour, position), "telic" relationships (the purpose or function of an object, e.g. the purpose an agent has in performing an act or the built-in function or aim which specifies certain activities) and "agentive" relationships (factors involved in the origin or "bringing about" of an object, e.g. creator, artifact, natural kind, causal chain). Of the relationships identified by Pustejovsky, WordNet partially considers argument structure (only in the case of verbs, as verb groups), inheritance structure (hyponym relationships, but not the "complex type" relationships identified by Pustejovsky) and qualia structure. In the case of qualia structure it only considers "constitutive" relationships, in the form of meronym relationships (member, substance and part) and, in part, "agentive" relationships (in the form of entailment and causal relationships).

There is scope therefore for enhancing WordNet by adding relationships such as argument structure for words that are not verbs, event structure, complex inheritance, and qualia structures such as formal and telic relationships.

This study will focus on telic relationships and in particular will examine a method by which telic relationships can be automatically discovered from the glosses contained in WordNet itself and used to augment WordNet itself.

2 Telic Relationships

The qualia structures identified by Pustejovsky are derived in part from the Aristotelian view of word meaning which identified a set of "modes of explanation" (*aitiai*) that could be applied to words. These modes of explanation identify particular aspects of meaning (*qualia*) which can be used to connect words in a lexicon. One particular aspect of meaning is the "telic" of an object, indicating the purpose or function of an object, for example the purpose an agent has in performing an act or the built-in function or aim which specifies certain activities. Thus the telic of *milk* would be *drink*, as the purpose of milk is to be drunk. Although Pustejovsky never mentions multiple relationships, it is conceivable that a word may have more than one telic relation, as, for example, wood, used both for burning and for making furniture.

The objective was to therefore to extend WordNet by creating a new set of relations telic(A, B), linking two synsets and indicating that there exists a telic relationship between A and B, such that if A is a synset representing a word, B is the telic of A, or, in other words, that A is used to achieve B.

3 Related Work

Machine readable dictionaries and encyclopaedias have been shown to be useful tools in the creation of knowledge-bases. Different approaches have been applied, including pattern-matching (e.g. Chodrow et al., 1985), and specially constructed or broad coverage parsers (see for example Wilks et al. 1996; Richardson et al. 1998; Kang and Lee 2001; Katz et al. 2001). WordNet glosses, brief explanations describing the particular meaning of individual synsets within WordNet, have been successfully used to semi-automatically enhance and create knowledge bases. Moldovan and Rus

(2001), and Harabagiu et al. (1999), for example, parsed the text of the glosses in order to transform them into a logical form to be used respectively as axioms in reasoning about world-knowledge and to enhance WordNet with new derivational morphology relations. Attempts have also been made to automatically build qualia relations from corpora (Pustejovsky et al. (1993); Bouillon et al. (2001)).

4 Method

In order to use the WordNet glosses to add telic relations to WordNet itself, it was necessary to:

- Extract the telic relation (or, possibly, relations) from the gloss using some parsing method. The result of this process was expected to be a word or a group of words representing the telic relation(s) for the synset which provided the gloss.
- Transform the telic word into a synset by disambiguating its meaning, thus avoiding the creation of misleading relationships

In the present experiments, pattern-matching, enhanced by a very simple part-of-speech parser was used to find the relevant information, i.e. a word representing the telic of the given synset. The extracted information was then passed to a word disambiguation module which returned a synset for the given telic word.

4.1 Identification and extraction of telic words

Initially the WordNet glosses were cleaned, removing example sentences. The glosses were then analyzed for patterns indicating that the gloss contained information about the telic relations for a particular synset. It was noted that within the given glosses telic relations could be found in the presence of the following patterns:

"... to TELIC_VERB by the use of...", as in *mammography* (synset id 100649306), which has as gloss "a diagnostic procedure to detect breast tumors by the use of X rays", indicating that the telic relation for *mammography* is to detect breast tumors (i.e. it is used to detect breast tumors).

"... used for TELIC" , as in *tracing_paper* (synset id 110816432), with gloss "a semitransparent paper that is used for tracing drawings", indicating that the telic relation for tracing paper is to trace drawings.

"... used to TELIC", as in *cardiac_glycoside* (synset id 110805579), defined as "obtained from a number of plants and used to stimulate the heart in cases of heart failure", indicating that the telic relation for cardiac glycoside is the stimulation of the heart.

"... use of ... to TELIC" as in *trickery* (synset id 100485559), with gloss "the use of tricks to deceive someone (usually to extract money from them)", indicating that the telic of trickery is to deceive someone.

"... used as ... in TELIC_ING-VERB" as in *Plasticine* (synset id 110453999), defined as "a synthetic material resembling clay but remaining soft; used as a substitute for clay or wax in modeling (especially in schools)", indicating that the telic relation for Plasticine is modeling.

"... used in TELIC_ING-VERB" as in *seal_oil* (synset id 110781016), with gloss "a pale yellow to red-brown fatty oil obtained from seal blubber; used in making soap and dressing leather and as a lubricant", indicating that the telic relations for seal_oil should be making soap, dressing leather and lubrication.

"... used in ... as a TELIC" as in *giant_taro* (synset id 108093257) with gloss "large evergreen with extremely large erect or spreading leaves; cultivated widely in tropics for its edible rhizome and shoots; used in wet warm regions as a stately ornamental", indicating that the telic relation of a giant taro is its use as a stately ornamental.

"... for use as TELIC" as in *houseboat* (synset id 102838388) defined its gloss as "a barge that is designed and equipped for use as a dwelling" indicating that the telic relation for a houseboat is a dwelling.

"... for use in ... TELIC_ING-VERB" as in *wherry* (synset id 103611080), with gloss "light rowboat for use in racing or for transporting goods and passengers in inland waters and harbors", indicating that the telic relation for a wherry is racing and the transportation of goods and passengers.

A subset of the WordNet glosses possibly containing information regarding telic relations was therefore taken and each gloss was split where more than one telic relation was indicated, as in the presence of conjunctions or disjunctions (as in synset 103357011, "a sailing ship [...] used in fishing and sailing along the coast", where two telic relations, a) fishing, and b) sailing, are present) and semicolons (as in synset 110842812, "any of a group of synthetic

steroid hormones used to stimulate muscle and bone growth; sometimes used illicitly by athletes to increase their strength" where two telic relations are present, a) the stimulation of muscle and bone growth, and b) to increase strength).

It was then necessary to identify one word (or compound word) that could summarize the telic relationship found in the gloss. In particular, it was necessary to identify one verb or noun that would represent the telic for a chosen synset. In order to avoid over-generalization, words such as "be", "do", "make" and "thing" were avoided and where these were found, a more specific word was sought. So, for example, in seeking the telic relationship for conditioner (synset id 102485262), defined as "a substance used in washing (clothing or hair) to make things softer", the relationship that was sought was to "make soft", i.e. to soften, not simply to "make", which is far too general to be of any use. In these cases a more specific noun or verb was sought and, in the absence of a more specific noun or verb the adjective attached to the noun was modified to find a more specific verb (in the case of conditioner, the adjective "soft" was used to derive the verb "soften").

4.2 Telic Word Sense Disambiguation

Having identified one word that summarized the telic relationship, it was necessary to identify the specific sense of the telic word, i.e. to find the synset which represented the telic relationship. It was therefore necessary to consider the set of possible synsets to which the telic word could belong and choose the synset that best represented the meaning of the telic word. One approach to disambiguation of word sense is to use some measure of relatedness between the word and its context, or, in other words, to calculate the semantic similarity or conceptual distance between the word and its context (Miller and Teibel 1991, Rada et al. 1989). WordNet has been shown to be fruitful in the calculation of semantic distance, determining similarity by calculating the length of the path or relations connecting two concepts; different approaches either using all WordNet relations (Hirst-St-Onge (1998) or only is-a relations (Resnik (1995); Jiang-Conrath (1997); Lin

(1998); Leacock-Chodorow (1998)) have been proposed (for an evaluation see Budanitsky and Hirst (2001)). In determining conceptual distance (also referred to as conceptual density), Mihalcea and Moldovan (1999) and Harabagiu et al. (1999) found WordNet glosses, considered as micro-contexts, to be useful. A similarly inspired approach was taken in this study, with the meaning of a telic word being constrained by a) the gloss from which it was extracted and b) the set of glosses of the synsets to which it could belong.

Therefore, given a word w , whose meaning was represented by the gloss (definition) GW_w , and its telic word t , it was necessary to find the synset ts representing the correct meaning of t from the set of all synsets T to which ts could belong. The correct synset ts (in other words, the correct sense) for the telic word t was taken to be the synset that maximized the semantic distance between GW_w and ts , as follows:

$$ts = \text{Argmax}_{ts \in T} sd(GW_w, ts)$$

where $sd(a, s)$ is a function calculating the semantic distance sd between a sentence a and a particular meaning of word w , represented by its synset s , which returns a number between 0 and 1 indicating the relatedness between the sentences, where 1 indicates they have the same meaning and 0 indicates they have no meaning in common.

In order to calculate the semantic distance sd the set TS_{ts} was constructed by taking all the words in the gloss GT of ts and all the words in the glosses of all the hyponyms and hypernyms of ts (to a depth of 3 hypernyms and 3 hyponyms) as follows:

$$TS_{ts} = \{w : \begin{array}{l} w \in GT \vee \\ w \in \text{hyperg}(ts, 3) \vee \\ w \in \text{hypog}(ts, 3) \end{array} \}$$

Where w represents a word; $\text{hyperg}(s, d)$ is a function which returns a set made up of all the words in the glosses of the hypernyms of a synset s , to a depth d ; and $\text{hypog}(s, d)$ is a function which returns a set made up of all the words in the glosses of the hyponyms of a synset s , to a depth d .

TS was then compared with GW_w (which was considered the set of words making up a gloss) by using a form of term overlap measure to measure their semantic relatedness. Initially, a set of stop-words SW was used to ignore words that were too common to be of any use (e.g. "the", "do") thus producing the two reduced sets RGW_w and RTS :

$$\begin{aligned} RGW_w &= GW_w - SW \\ RTS &= TS - SW \end{aligned}$$

The remaining words were then analyzed to find their stems thus producing the two sets SGW_w and STS made of the stems of the words belonging to RGW_w and RTS .

Each word in RGW_w was then compared to all the words in RTS , using all the available WordNet relationships (is_a, satellite, similar, pertains, meronym, entails, etc.), with the additional relationship, "same_as", which indicated that two words were identical. Each relationship was given a weighting indicating how related two words were, with a "same as" relationship indicating the closest relationship, followed by synonym relationships, hypernym, hyponym, then satellite, meronym, pertains, entails. Each word w_i in RGW_w was therefore assigned a weighting r_i indicating its relatedness to RTS , and the total semantic distance tsd between RGW_w and RTS was calculated as the normalized sum of all the weightings r of RGW_w . The normalization was carried out by dividing by the number of words in the gloss by sum of the result + 1, in order for short glosses not to be disadvantaged.

$$tsd = |RTS| / (\sum_{w \in RGW} r(w, RTS) + 1)$$

A number of experiments were conducted to see to what depth the hyper- and hyponyms of a candidate synset should be considered, i.e. to decide, given a synset S , and its hyponyms HS , if the hypernyms HS' of the set HS also be considered, if the hypernyms HS'' of HS' should be considered, and so forth to an arbitrary depth n . It was found that a depth of 3 hypernyms and 3 hyponyms gave satisfactory results in an acceptable time; however further research would be necessary to optimize this parameter.

5 Results

2449 telic relationships were derived, relating to 1841 different synsets (i.e. a synset could have more than one telic relationship). A sample (about 10% of the total) of the derived relationships was examined manually and it was estimated that 78% of the relationships were actually telic relationships, while the rest either denoted other types of relationships (e.g. the context in which something is used), or denoted telic relationships that could not be summarized in one word (as in synset 110556533, "activating agent", whose telic should be "increase the attraction to a specific mineral"). 9% of the relationships were in effect telic relationships, but were counterintuitive, in part because of the limitations posed by the adopted method, which considered telics in isolation from their possible objects (e.g. a cancer drug having as telic "kill", because its function is to kill cancer cells). Of the correct relationships, the correct synset (i.e. the correct meaning) was chosen 77% of the time. The disambiguation algorithm failed mainly where there were very subtle differences in meaning in WordNet, as in the difference in meaning for the word "represent" between synset 201841374 (take the place of), 200566766 (express indirectly; be a symbol of) and 200265192 (to establish a mapping (of mathematical elements or sets)), or the difference in meaning for the word "stain", between synset 200196870 (produce or leave stains; "Red wine stains table cloths") and 201053918 (make a spot or mark onto; "The wine spotted the tablecloth").

Total relationships found	2449
Number of different synsets	1841
Actual telic relationships	78% (1910)
Of which with correct synset	77% (1470)

A number of useful telic relationships were derived:

Example 1: from synset 102853717, indicating "incubator, brooder", defined as "a box designed to maintain a constant temperature by the use of a thermostat; used for chicks or premature infants" it was derived that incubators are used for (or: the telic of an incubator is) maintaining a constant temperature: the telic word was therefore correctly identified as

"maintain", with the particular meaning given by synset 201829600, i.e. "keep in a certain state, position, or activity", which was correctly chosen by the algorithm in preference to, for example, synset 200723279 (maintain by writing regular records), synset 200607420 (support against an opponent) and 200496801 (observe correctly).

Example 2: from synset 100450328, indicating "desensitization technique, desensitization procedure, systematic desensitization", defined as "a technique used in behavior therapy to treat phobias and other behavior problems involving anxiety; client is exposed to the threatening situation under relaxed conditions until the anxiety reaction is extinguished", the algorithm correctly inferred that the telic of desensitization technique was "treat" in the particular meaning given by synset 200054862, i.e. "provide treatment for" as in "The doctor treated my broken leg"; this meaning was chosen in favour of incorrect alternatives such as 201547305 (provide with a treat) and 200699711 (deal with verbally or in some form of artistic expression).

Example 3: from synset 102399372, indicating "cash_register, register", "a cashbox with an adding machine to register transactions; used in shops to add up the bill" it was correctly inferred that the telic was 201805970, "add up", with the meaning "add up in number or quantity", and not "add up" as in synset 201786912, "be reasonable or logical or comprehensible" or in synset 201792159, "develop into".

6 Conclusions and further work

WordNet contains a significant amount of implicit information in the form of synset glosses. This study has shown how this implicit information can be made explicit by automatically extracting new relationships. In particular, the algorithm presented usefully extended WordNet by automatically inducing a number of telic relationships from the glosses.

Future work will involve a manual review of the relationships found to ensure that the derived relationships are of sufficient high-quality to be used in practice. It will be also necessary to address the limitations of the method of

representation chosen, which constrained telic relationships to be between two words as opposed to relationships between a word and a sentence: in a number of instances a single word was not enough to represent a telic relationship. Another problem that needs to be tackled is the fine granularity of meaning in WordNet, which in some cases made it very difficult to choose a particular synset (meaning) for a telic

relationship. Another direction for future work will be the application of the proposed methodology to derive telic relationships from other machine readable dictionaries. A further interesting direction for research would be exploring the possibility of moving on from machine readable dictionaries in order to derive telic relationships from generic corpora.

References

- Bouillon, P., Claveau, V., Fabre, C., Sebillot, P., "Using part-of-speech and semantic tagging for the corpus-based learning of qualia", in Bouillon and Kanzaki (eds.), Proceedings of the First International Workshop on Generative Approaches to the Lexicon, Geneva, 2001.
- Budanitsky, A., and Hirst, G., "Semantic distance in WordNet: and experimental, application-oriented evaluation of five measures", in Proceedings of the NAACL 2001 Workshop on WordNet and other lexical resources, Pittsburgh, 2001.
- Chodrow, M., Byrd, R., Heidorn, G, "Extracting semantic hierarchies from a large on-line dictionary", In Proceedings of the 23rd Annual Meeting of ACL, 1985.
- Fellbaum, C., "WordNet, An electronic Lexical Database", MIT Press, 1998.
- Harabagiu, S. A., Miller, A. G., Moldovan, D. I., "WordNet2 - a morphologically and semantically enhanced resource", In Proceedings of SIGLEX-99, University of Maryland, 1999.
- Hirst, G., and St-Onge, D., "Lexical chains as representations of context for the detection and correction of malapropisms", in Fellbaum (ed.), "WordNet: and electronic lexical database", MIT Press, 1998.
- Jiang, J. J., and Conrath, D. W., "Semantic similarity based on corpus statistics and lexical taxonomy", in Proceedings of ICRCL, Taiwan, 1997.
- Kang, S-J, Lee, J-H, "Semi-Automatic Practical Ontology Construction by Using a Thesaurus, Computational Dictionaries and Large Corpora", Proceedings of the Workshop on Human Language Technology, ACL-2001, Toulouse, 2001.
- Katz, B., Lin, J., Felshin, S., "Gathering Knowledge for a Question Answering System from Heterogeneous Information Sources", Proceedings of the Workshop on Human Language Technology, ACL-2001, Toulouse, 2001.
- Lin, D., "An information-theoretic definition of similarity", in Proceedings of the 15th International Conference on Machine Learning, Madison, 1998.
- Miller, G, and Teibel, D., "A proposal for lexical disambiguation", in Proceedings of DARPA Speech and natural Language Workshop, California, 1991.
- Miller, G. A., "WordNet: A Lexical Database", Communications of the ACM, 38 (11), 1995.
- Moldovan, D, and Rus, V., "Logic Form Transformation of WordNet and its Applicability to Question Answering", in Proceedings of ACL-2001, Toulouse, 2001.
- Pustejovsky, J., Bergler, S., Anick, P., "Lexical Semantic techniques for corpus analysis", Computational Linguistics, 19 (2), 1993.
- Pustejovsky, J., "The Generative Lexicon", MIT Press, 1995.
- Rada, R., Mili, H., Bicknell, E. and Blettner, M., "Development and application of a metric on semantic nets", in IEEE Transactions on Systems, Man and Cybernetics, vol.19, n.1, 1989.
- Rada Mihalcea and Dan Moldovan, A Method for Word Sense Disambiguation of Unrestricted Text, in Proceedings of ACL '99, Maryland, NY, 1999.
- Resnik, P., "Using information content to evaluate semantic similarity", in Proceedings of the 14th IJCAI, Montreal, 1995.
- Richardson, S. D., Dolan, W. D., Vanderwende, L., "Mindnet: acquiring and structuring semantic information from text", in Proceedings of COLING-98, 1998.
- Wilks, Y. A., Slator, B. M., Guthrie, L. M., "Electric Words: dictionaries, computers and meaning", MIT Press, 1996.