Using and Extending WordNet to Support Question-Answering

Peter Clark[1], Christiane Fellbaum[2], Jerry Hobbs[3]

[1]Boeing Phantom Works, Seattle (USA)
[2]Princeton University, Princeton (USA)
[3]USC/ISI, Marina del Rey (USA)
peter.e.clark@boeing.com, fellbaum@clarity.princeton.edu, hobbs@isi.edu

Abstract. Over the last few years there has been increased research in automated question-answering from text, including questions whose answer is implied, rather than explicitly stated, in the text. WordNet has played a central role in many such systems (e.g., 21 of the 26 teams in the recent PASCAL RTE3 challenge used WordNet), and thus WordNet is being increasingly stretched to play more semantic tasks in applications. As part of our current research, we are exploring some of the new demands which question-answering places on WordNet, and how it might be further extended to meet them. In this paper, we present some of these new requirements, and some of the extensions that we are currently making to WordNet in response.

Keywords: WordNet, question answering, textual entailment, world knowledge

1 Introduction

Advanced question-answering is more than simply fact retrieval; typically, much of the knowledge that an author wishes to convey is never explicitly stated in text (by one estimate the ratio of explicit:implicit knowledge is 1:8, [1]). Rather, the reader fills in the missing pieces using his/her background knowledge, creating a "mental model" of the scenario the text is describing, allowing him/her to go beyond facts explicitly stated. For example, given:

"A soldier was killed in the gun battle"

a reader would infer that, plausibly, the solder was shot, even though this fact is never explicitly stated.

A key requirement for this task is access to a large body of world knowledge. However, machines are currently poorly equipped in this regard, and developing such resources is challenging. Typically, manual acquisition of knowledge is too slow, while automatic acquisition is too messy. However, WordNet [2,3] presents one avenue for making inroads into this problem: It already has broad coverage, multiple lexico-semantic connections, and significant knowledge encoded (albeit informally)

in its glosses; it can thus be viewed as on the path to becoming an extensively leveragable resource for reasoning. Our goal is to explore this perspective, and to accelerate WordNet along this path. The result we are aiming for is a significantly enhanced WordNet better able to support applications needing extensive semantic knowledge.

2 Semantic Requirements on WordNet

To assess WordNet's strengths and limitations for supporting textual question-answering, have been working with the task of "recognizing textual entailment" (RTE) [4,5], namely deciding whether a hypothesis sentence, H, follows from an initial text T. For example, from:

(1.T) Satomi Mitarai bled to death.

the following hypotheses plausibly follow:

(1.H1) Satomi Mitarai died.
(1.H2) Mitari lost blood.

Similarly, from:

(2.T) Hanssen, who sold FBI secrets to the Russians, could face the death penalty.

it plausibly follows that:

(2.H1) The FBI had secrets.
(2.H2) Hanssen received money from the Russians.
(2.H3) Hanssen might be executed.
(2.H4) The Russians bought secrets from Hanssen.

Our methodology has been to define a test suite of such sentences, analyze the types of knowledge required to determine if the entailment holds or not, and then determine the extent to which WordNet can provide this knowledge already and where the gaps are. For these gaps, we are exploring ways in which they can be partially filled in.

The test suite we developed contains 244 T-H entailment pairs (122 of which are positive entailments) such as those shown above. The pairs are grammatically fairly simple, and were deliberately authored to focus on the need for lexico-semantic knowledge rather than advanced linguistic processing. Determining entailment is very challenging in many cases. Each positive entailment pair was analyzed to identify the knowledge required to answer them. For example, for the pair:

(3.T) Iran purchased plans for a nuclear reactor from A.Q.Khan.
(3.H) The Iranians bought plans for building a nuclear reactor.

the computer needs to know:

"Iranian" is a person from Iran (derivational link)

"buy" and "purchase" are approximately equivalent (synonyms)
"plans for X" can mean "plans for building X" (world knowledge)

This process was repeated for all 122 positive entailments. From this, we found the knowledge requirements could be grouped into approximately 15 major categories, namely knowledge of:

1.Synonyms
2.Hypernyms
3.Irregular word forms
4.Proper nouns
5.Adverb-adjective relations
6.Noun-adjective relations
7.Noun-verb relations and their semantics (e.g., a consumer is the AGENT of consume event)
8.Purpose of artifacts
9.Polysemy vs. homonymy (related vs. unrelated senses of a word form)
10.Typical/plausible behavior (planes fly, bombs explode, etc.)
11.Core world knowledge (e.g., time, space, events)
12.Specific world knowledge (e.g., bleeding involves loss of blood)
13.Knowledge about actions and events (preconditions, effects)
14.Paraphrases (linguistically equivalent ways of saying the same thing)
15.Other

Of these, WordNet already has rich coverage of synonyms, hypernyms, adverb-adjective relations, and noun-adjective relations. It also has knowledge of noun-verb relations, although it does not distinguish between the different semantic type of this relation (e.g., AGENT, INSTRUMENT, EVENT); and some knowledge about semantic similarity highly polysemous verbs. In addition, WordNet has some knowledge of irregular word forms and proper nouns, and additional information is easily obtainable from other existing resources. The remaining knowledge types are still lacking; our goal is to extend WordNet to help provide more of this kind of knowledge. Note that we do not view WordNet as the sole supplier of knowledge, rather we wish to increase its utility as a contributing knowledge resource of systems performing advanced question-answering.

3 Recent WordNet Extensions

Based on this analysis, we are making several extensions to WordNet, which we describe in the following sections.

3.1 Morphosemantic links

WordNet contains mostly paradigmatic relations, i.e., relations among synsets with words belonging to the same part of speech (POS). Version 2 introduced cross-POS links, so-called "morphosemantic links" among synsets that were not only

semantically but also morphologically related [6]. There are currently tens of thousands of manually encoded noun-verb (sense) connections, linking derivationally related nouns and verbs, e.g.,:

abandon#v1 - abandonment#n3
rule#v6 - ruler#n1
catch#v4 - catcher#n1

Importantly, the appropriate senses of the nouns and verbs are paired, e.g., "ruler" and "rule" refer to the measuring stick and the marking or drawing with a ruler, respectively, rather than to a governor and governing, which makes for a different pair. What WordNet does not currently inform about, however, is the nature of the relation. For example:

abandonment#n3 is the EVENT of abandon#v1
ruler#n1 is the INSTRUMENT of rule#v6
catcher#n1 is the AGENT of catch#v4

Knowledge of the nature of such relations is essential for many question-answering tasks. For example, given

(4.T) "Dodge produces ProHeart devices",

it is needed to realize that "producer" refers to the AGENT ("Dodge"), "production" refers to the EVENT ("produces"), and "product" the RESULT ("ProHeart devices"), a prerequisite for correctly answering questions asking about the producer/production/product.

The scale of adding this information manually is somewhat daunting; there are approximately 21,500 noun-verb (sense) links needing to be typed in WordNet. (We have not yet considered morphosemantic links among synsets from other parts of speech, which could also contribute to WordNet's usefulness as a tool for automated question answering.)

We have devised the following semi-automated approach:

1. We extract the noun-verb pairs with a particular morphological relation, (e.g., "-er" nouns such as "builder"-"build")
2. We determine the default relation for these pairs (e.g., The noun is the AGENT of the action expressed by verb)
3. We manually go through the list of pairs, marking pairs not conforming to default relation.
4. We inspect and group the marked pairs, assigning the correct relations to them.

This methodology is substantially faster than simply labelling each pair one by one, as only exceptions to the default relation need to be manually classified. In addition, this method has revealed the surprisngly high degree to which generally accepted one-to-

one mappings of morphemes with meanings is violated;. Furthermore, it is interesting to see that across the morphological classes, a limited inventory of semantic relations applies (for details see [7]).

3.2 Purpose links

A second type of knowledge often needed in question-answering is the function or purpose of artifacts (natural entities like stones and trees do not have an inherent function). For example, given:

(5.T) "The soldier was killed in a gun fight"
(5.H) "The soldier was shot"

we need to know that a gun is for shooting in order to infer that 5.H plausibly follows from 5.T. Knowledge what an artifact is intended for and how it is typically used enables a computer to make a plausible guess about implicit events that are not overtly expressed in a text. So our goal is to add links among noun and verb synsets in WordNet such that the verbs denote the intended and typical function or purpose of the nouns.

The number of such links is potentially huge, as almost any object can be used for almost any function. Thus, one can kill someone with a stiletto shoe, using is as a weapon. Similarly, a tree stump could be sat on when no chair is available. Worse, just about any solid object of a certain size can be used for hitting. We try to limit our links to those expressing the intended function, similar to the Role qualia of Pustejovsky [8]. Corpus data, e.g., [9], can be used to identify the most frequent noun-verb cooccurrences and usually confirm one's intuition about which noun-verb synset pairs should be linked.

Manually adding the links is a daunting task. However, a semi-automated approach is possible, using existing morphosemantic links in WordNet. As noted by Clark and Clark [10], English has a productive rule and fairly regular rule whereby many nouns can be used as verbs, and in many cases, the verb denotes the noun's intended function (or, put differently, the noun is the Instrument for carrying out the action expressed by the verb). Examples are "gun"(n)-"gun"(v): A gun is for gunning; "pencil"(n)-"pencil"(v): A pencil is for penciling, a hyponym of writing. In cases where there is no corresponding verb, e.g., for "car"(n), we can search up the hypernym tree until a more general noun is found which does have a corresponding verb, e.g., "car"(n) is a "transport"(n), linked to "transport"(v), thus a "car" is for "transporting".

We are currently inspecting the list of so-called zero-derived (homographic) noun-verb pairs in WordNet and classifying them as described in 3.1. Those pairs where the noun is an Instrument will be encoded with purpose links. Similarly, all noun-verb pairs from the different morphological classes (-er, -al, -ment, -ion, etc.) that were classified as expressing an Instrument relation can be labeled as "Purpose."

The automatic extraction of pairs related via a specific affix (Step 1 in 3.1 above) generates a list of candidate pairs that is validated and corrected by the same lexicographer who manually inspects the pairs for their semantic relation. Most pairs that are generated are valid, but a few false hits must be discarded. For example, the noun synset {coax, ethernet cable} was paired with the verb "coax", which would lead to the statement "An ethernet cable is for coaxing". In the majority of cases the computer's guess is sensible, and hence construction of the database is much faster than working from scratch.

3.3 World Knowledge - WordNet Glosses

WordNet contains a substantial amount of knowledge within its glosses. In particular, note that knowledge about a word (sense) is not just contained in that sense's gloss and example sentences, but also in its use in other glosses and example sentences. For example, for the word "lawn", WordNet includes mention that a lawn:

- needs watering;
- can have games played on it;
- can be flattened, mowed;
- can have chairs on it and other furniture;
- can be cut/mowed;
- can have things growing on it;
- has grass;
- can have leaves on it; and
- can be seeded.

Despite this promise, this knowledge is largely locked up in informal English text, and difficult to extract in a machine-usable form (although there has been some work on translating the glosses to logic, e.g., [11,12]. The glosses were not originally written with machine interpretation in mind, and as a result the output of machine interpretation is often syntactically valid but semantically meaningless logic. To address this challenge, we are proceeding along two fronts: first, we are developing an improved language processor specifically designed for interpreting the WordNet glosses; second, we are manually rephasing some of the glosses to create more regularity in their structure, so that the resulting machine interpretation is improved.

To scope this work, we are focusing on "Core WordNet" Because WordNet contains tens of thousands of synsets referring to highly specific animals, plants, chemical compounds, etc. that are less relevant to NLP, the Princeton WordNet group has compiled a CoreWordNet, consisting of 5,000 synsets that express frequent and salient concepts. These were selected as follows. First, a list with the most frequent strings from the BNC was automatically compiled and all WordNet synsets for these strings were pulled out. Second, two raters determined which of the senses of these strings expressed "salient" concepts [13]. The resulting top 5000 concepts comprises the core that we are focusing on, and as a result of this method of data collection

contains a mixture of general and (common) domains-specific terms. (CoreWordNet is downloadable from http://wordnet.cs.princeton.edu /downloads.html)

3.4 World Knowledge - Core Theories

In addition to the specific world knowledge that might be obtained from the glosses, question-answering sometimes requires more fundamental, "core" knowledge of the world, e.g., about space, time, events, cognition, people and activities. Because of its more general nature, such knowledge is less likely to come from the WordNet glosses, and instead we are encoding some of this knowledge by hand as a set of "core theories". Although these theories contain only a small number of concepts (synsets), these concepts are also often general, meaning that information about them can be applied to a large number of other WordNet concepts. For example, WordNet has 517 "vehicle" nouns, and so any general knowledge about vehicles in general is potentially applicable to all these subtypes; similarly WordNet has 185 "cover" verbs, so general knowledge about the nature of covering can potentially apply to all these subtypes. In general, the broad coverage of WordNet can be funneled into a much smaller defined core, which can then be richly axiomatized, and the resulting axioms applied to much of the wider vocabulary in WordNet.

To identify these theories, we sorted words in Core WordNet into groups based on (a somewhat intuitive notion of) coherence, resulting in 15 core theories (listed with a selection of the words in them):

- Composite Entities: perfect, empty, relative, secondary, similar, odd, ...
- Scales: step, degree, level, intensify, high, major, considerable, ...
- Events: constraint, secure, generate, fix, power, development, ...
- Space: grade, inside, lot, top, list, direction, turn, enlarge, long, ...
- Time: year, day, summer, recent, old, early, present, then, often, ...
- Cognition: imagination, horror, rely, remind, matter, estimate, idea, ...
- Communication: journal, poetry, announcement, gesture, charter, ...
- Persons and their Activities: leisure, childhood, glance, cousin, jump, ...
- Microsocial: virtue, separate, friendly, married, company, name, ...
- Material World: smoke, shell, stick, carbon, blue, burn, dry, tough, ...
- Geo: storm, moon, pole, world, peak, site, village, sea, island, ...
- Artifacts: bell, button, van, shelf, machine, film, floor, glass, chair, ...
- Food: cheese, potato, milk, break, cake, meat, beer, bake, spoil, ...
- Macrosocial: architecture, airport, headquarters, prosecution, ...
- Economic: import, money, policy, poverty, profit, venture, owe, ...

We are first focusing on Time and Event words. We have developed underlying ontologies of time and event concepts, explicating the key notions in these domains [14,15]. For example, the temporal ontology axiomatizes topological temporal concepts like before, duration concepts, and concepts involving the clock and calendar. The event ontology axiomatizes notions like subevent, and the internal

structure of events and processes. We are then defining, or at least characterizing, the meanings of the various word senses in terms of these underlying theories. For example, to fix something is to bring about a state in which all the components of the thing are functional. This effort is of course a very labor intensive project, but since we are concentrating on the synsets in the core WordNet, we believe we will achieve the maximum impact for the labor we put into it.

Because of the richness of WordNet's hypernym links, in principle these axioms can be heavily reused for reasoning about WordNet word senses. A number of the textual entailment problems in our test suite appeal directly to this knowledge, for example to judge the validity of this entailment:

(6.T) Baghdad has seen a spike in violence since the summer.
(6.H) There was greater violence in Baghdad since the summer.

requires reasoning about the core notion of change in a quantity ("spike", "rise"), rather than anything specific about Baghdad, violence, or summer. This kind of knowledge - namely the meaning of these core words and their relationships - is being encoded in these core theories.

4. Status and Summary

The work that we have described here is still a work in progress: To date, we have corrected/validated about half of the machine-generated database of morphosemantic links; made an initial start on the purpose links; have completed a first pass on logical forms for WordNet glosses and are focussing on improving both the phrasing and interpretation of Core WordNet; and have completed some of the core theories and are in the process of linking their core notions to WordNet word senses. Our goal is that these extensions will substantially improve WordNet's utility for language-based problems that require reasoning as well as basic lexical information, and we are optimistic that these will improve WordNet's ability to meet the increasingly strong requirements demanded by modern day language-based applications.

Acknowledgements. This work was supported by the AQUAINT Program of the Disruptive Technology Office under contract number N61339-06-C-0160.

References

1. Graesser, A. C. (1981). "Prose Comprehension Beyond the Word". NY:Springer.
2. Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM. 38.11:39-41
3. Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
4. Giampiccolo, D., Magnini, B., Dagan, I., Dolan, B. (2007). The Third PASCAL Recognizing Textual Entailment Challenge. In Proc. 2007 Workshop on Textual Entailment and Paraphrasing, pp 1-9. PA: ACL.

5. Clark, P., Harrison, P., Thompson, J., Murray, W., Hobbs, J., and Fellbaum, C. (2007). On the Role of Lexical and World Knowledge in RTE3. ACL-PASCAL Workshop on Textual Entailment and Paraphrases, June 2007, Prague, CZ.
6. Miller, G. A. and Fellbaum, C. (2003). Morphosemantic links in WordNet. Traitement automatique de langue, 44.2:69-80.
7. Fellbaum, C., Osherson, A., and Clark, P.E. (2007). Putting Semantics into WordNet's "Morphosemantic" Links. In: Proceedings of the Third Language and Technology Conference, Poznan, Poland, October 5-7.
8. Pustejovsky, J. (1995). The Generative Lexicon. Cambridge, MA: MIT Press.
9. Clark, P., Harrison, P. (2003). The Reuters Tuple Database. (Available on request from peter.e.clark@boeing.com).
10. Clark, E. and Clark, H. (1979). When nouns surface as verbs. Language 55, 767-811.
11. Harabagiu, S.M., Miller, G.A., Moldovan, D.I. (1999). WordNet 2 - A Morphologically and Semantically Enhanced Resource, Proc. SIGLEX 1999, pp1-8.
12. Fellbaum, C., Hobbs, J., (2004). WordNet for Question Answering (AQUAINT II Project Proposal). Technical Report, Princeton University.
13. Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. (2006). Adding dense, weighted, connections to WordNet. In: Proceedings of the Third Global WordNet Meeting, Jeju Island, Korea, January 2006.
14. Hobbs, Jerry R., and Feng Pan, 2004. An Ontology of Time for the Semantic Web, ACM Transactions on Asian Language Information Processing, Vol. 3, No. 1, March 2004.
15. Hobbs, Jerry R., 2007. Encoding Commonsense Knowledge. Technical Report, ISI, http://www.isi.edu/~hobbs/csk.html