Language to Logic Translation with PhraseBank

Adam Pease[1] and Christiane Fellbaum[2]

[1] Articulate Software Inc
278 Monroe Dr. #30, Mountain View, CA 94040
Email: adampease@earthlink.net
[2] Princeton University
Department of Psychology, Green Hall, Princeton, NJ 08544
Email: fellbaum@princeton.edu

Abstract. We discuss a restricted natural language understanding system and a proposed extension to it, which is a corpus of phrases. The Controlled English to Logic Translation (CELT) system allows users to make statements in a domain-independent, restricted English grammar that have a clear formal semantics and that are amenable to machine processing. CELT needs a large amount of linguistic and semantic knowledge. It is currently coupled with the Suggested Upper Merged Ontology, which has been mapped by hand to WordNet 1.6. We propose work on a new corpus of phrases (called PhraseBank) to be added to WordNet and linked to SUMO, which will catalog common English phrase forms, and their deep meaning in terms of the formal ontology. This addition should significantly expand the coverage and usefulness of CELT.

1 Introduction

We first discuss the existing components which make up the Controlled English to Logic Translation system, including its formal ontology and lexicon. We then describe CELT itself. The body of the paper discusses the PhraseBank effort and how it should improve the utility of CELT.

1.1 Upper Ontology

The Suggested Upper Merged Ontology (SUMO) (Niles&Pease, 2001) is a free, formal ontology of about 1000 terms and 4000 definitional statements. It is provided in first order logic, and also translated into the DAML semantic web language. It is now in its 56th version; having undergone three years of development, review by a community of hundreds of people, and application in expert reasoning and linguistics. SUMO has been subjected to formal verification with an automated theorem prover. It has also been mapped to all 100,000 noun, verb, adjective and adverb word senses in WordNet, which not only acts as a check on coverage and completeness, but also provides a basis for application to natural language understanding tasks. SUMO covers areas of knowledge such as temporal and spatial representation, units and measures, processes, events, actions, and obligations. Domain specific ontologies have been created that extend and reuse SUMO in the areas of finance and investment, country almanac information, terrain modeling, distributed computing, endangered languages description, biological viruses, engineering devices, weather and a

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 187–192.

number of military applications including terrorist events, army battlefield planning and air force mission planning. It is important to note that each of these ontologies employs rules. These formal descriptions make explicit the meaning of each of the terms in the ontology, unlike a simple taxonomy, or controlled keyword list.

SUMO has natural language generation templates and a multi-lingual lexicon that allows statements in KIF and SUMO to be expressed in multiple natural languages (Sevcenko, 2002). These include English, German, Czech, Italian, Hindi (Western character set) and Chinese (traditional characters and pinyin). A Tagalog lexicon is under development. Automatic translations can be viewed on line at `http://virtual.cvut.cz/kifb/en/`.

1.2 Restricted Natural Language

The Controlled English to Logic Translation (CELT) (Pease&Murray, 2003) (Murray et al, 2003) system performs syntactic and semantic analysis on restricted natural language input, and transforms it first order logic in Knowledge Interchange Format (KIF) syntax (Genesereth, 1991). The terms in the resulting KIF expressions come from the SUMO. This mapping of WordNet synsets to the ontology provides a deeper semantic analysis of the terms than what can be provided by a lexicon alone. A lexicon provides basic information, much like a dictionary. SUMO provides information about the term's concepts, attributes, and relationships.

CELT can perform active reasoning (via its associate inference engine) to derive answers that are not explicitly stated in the knowledge base. The knowledge is represented in domain knowledge bases (specified domain information), and a mid-level (more general domain information) and upper-level ontology (common sense concepts, world knowledge). The advantage of a tiered, modular knowledge structure is that it is efficient and reusable.

The user asks queries and makes assertions to CELT in a specified grammatical format. This subset of English grammar is still quite extensive and expressive. The advantage of the controlled English is that when the grammar and interpretation rules are restricted, then every sentence in the grammar has a unique parse. This eliminates the problems of ambiguity with other parsing approaches that would result in retrieving non-appropriate answers. For further discussion of controlled English grammars and applications, see Sowa (1999).

To overcome some of the limitations of CELT syntax, such as only handling indicative verbs and singular nouns, we developed other methods to extend its coverage. We use morphological processing rules, derived from the "Morphy" code of WordNet, to transform other verb tenses and plural verbs into the various tenses and numbers required. Discourse Representation Structures (DRSs) (Kamp & Reyle, 1993) handle context to resolve anaphoric references, implications, and conjunctions.

CELT does not limit the parts of speech or the number of word senses a word can have. Nor is the number of words limited. More importantly, CELT is not a domain specific system. It is a completely general language which can be specialized and extended for particular domains along with domain specific vocabulary. WordNet is being leveraged to provide core coverage of common English words. Currently we have about 100,000 words senses in our system. Individual words are identified based on the parse and lexicon.

2 Phrases in Language Understanding

Much of current NLP work, including part of speech and semantic tagging, focuses on language at the word level. But statistics show that speakers do not compose messages by freely combining words according to the rules of syntax and morphology. Much of language is composed of chunks or phrases, where specific lexical items co-occur in set patterns (Mul'cuk, 1998) The most frequent verbs in English (based on the Brown Corpus statistics) include “have,” “do,” “make,” “take,” and “give.” These verbs also are among the most polysemous and their meanings are represented by dozens of distinct senses in lexical resources, including WordNet. Clearly, they represent a challenge for any natural language processing application. One type of phrase are verb-noun chunks involves so-called “light” or “support” verbs (Church&Hanks, 1990), such as “have a shock,” “do the laundry,” “make a face,” and “give birth (to).” Thus, “take” occurs most frequently not in what might be called its primary sense, roughly paraphrasable as “get hold of with one's hands,” but in a collocations like “take walk” or “take a hit.” Other examples are “have a shock,” “do the laundry,” “do lunch,” “make a face,” “make progress,” “give birth (to),” “give grief (to).” These phrases are characterizable by two properties. First, the noun carries most of the semantic weight, with the verb providing relatively little information. Second, the verb phrase is often roughly synonymous with a simple verb that is morphologically related to the noun: “do/have lunch-lunch,” “take a walk-walk,” “make progress-progress,” etc.

Other examples are verb phrases like “pay attention/heed/homage,” which require the particular choice of a verb in a sense that is specific to these phrases. English has hundreds or perhaps thousands of such phrases. The author of a large-scale study of the uses of “take” (Church&Hanks, 1989) estimates that there are at least 10 000 phrases that follow the pattern “support verb plus noun”. The focus of our proposed work is on such phrases and phrase patterns. We believe that the automatic processing of natural language queries and answers will be greatly enhanced in an approach that considers chunks and phrases.

CELT first classifies phrases and identifies the patterns according to which they are composed and which define their meanings. In the current system, the corpus of phrases is quite limited, numbering only a few dozen. After having been parsed, the words in the frame-slot representation can be disambiguated against WordNet. Currently, the disambiguation of a polysemous word is performed by selecting the first sense of that word in WordNet, which displays the senses in the order determined by the frequency with which they were annotated to tokens in the Brown Corpus (Francis and Kucera, 1964) Miller et al. found that selecting the most frequent sense yields an accuracy rate of 65% (Miller et al, 1993). This method is clearly not good enough for reliable disambiguation. Moreover, the tagging effort was limited to a small number of words, covering a thematically unbalanced subset of the Brown Corpus. A reliable system must include more accurate lexical disambiguation.

By classifying phrases and establishing phrase patterns according to their semantics, we can match the component words of the phrases to WordNet entries with a very high degree of accuracy. For example, our classification will permit us to state with high degree of confidence that the sense of the verb “make” in a context where the parser has identified the word “trouble” as its direct object must be assigned sense 3 in WordNet: verb.creation: make, create (make or cause to be or become; “make a mess in one's office”; “create a furor”). The phrases will also be matched to template logical forms, allowing CELT to output a range of logic statements that more precisely capture the semantics of the sentence than would be otherwise possible by looking only at word senses and the syntactic parse.

One possible straightforward solution for the automatic processing of such phrases would be to ignore the light verb and treat the noun as the related verb. Thus, “take a walk” would be interpreted as “walk,” and “give birth” as “birthe.” But this turns out not to be an acceptable approach. First of all, the verbs are often polysemous, and the system would have to decide which sense to associate with the noun in such phrases. Second, to understand a text, a system needs to analyze the syntactic relations among sentence constituents, to, to put it simply, to understand “who does what to whom.” While the subject in both the phrases “take a walk,” “have lunch,” and “give birth” and in the corresponding verbs “walk,” “lunch,” and “birthe” is the Agent of the event, this is not the case in superficially similar phrases like “take a hit” and “have a shock” where the subject is the Undergoer, or Patient, in the event, and does not play the same semantic role (Agent, Stimulus) as the subject of “hit” and “shock.” A system that ignores the light verb and equates the noun with the related verb would seriously misinterpret the text in such cases.

Moreover, some phrases include the the same noun, but different verbs: “do lunch/have/take lunch,” “take/give a break (to).” In the first case, the meaning difference is subtle (“do” implying a social event), whereas in the second, the meaning of the two phrases is entirely unrelated.

A second solution would be to treat the entire phrase as a lexical unit. In fact, the lexical status of phrases like “take a walk” is unclear. On the one hand, they are partly compositional; one might argue that “take” in “take a vacation,” “take a walk,” and “take lunch” has an independent meaning and denotes the participation in an event. On the other hand, the phrases are idiosyncratic collocations: why do we say “make a decision” and not “take a decision” and why is it “take a photo” and not “make a photo” (as in French)? The restrictions on such phrases have to learned and stored in speaker’s mental lexicons.

But treating these phrases as a unit is not unproblematic for language processing. First, the lexicon would have to be augmented with a very large number of phrases; some of the patterns are in fact productive. More seriously, the parser would need to recognize the verb and the noun as a unit in all and only all the relevant cases so as to match it against the lexicon entry. This can be difficult in cases where the verb and the noun are not adjacent and do not conform to the lexicon entry, as in “take a long walk” or “inappropriate remarks were made.”

Instead, we propose an approach that avoids these problems. We classify light verb phrases and light verb phrase patterns semantically. For example, we collect phrases like “have a shock” and “have a surprise,” distinguishing them from superficially similar phrases like “have dinner” and “have a nap.” In the first case, the verb means “experience” (currently WordNet sense 11) and selects for a mental or emotional state. The subject is an Experiencer, and the event is a punctual achievement (Vendler 1967, Dowty 1991) In the second case, the phrases denote activities or processes and “have” here means roughly “partake of” or “engage in” (there’s currently no corresponding WordNet sense).

Actually, there is some kind of mutual selection of specific senses, (or co-composition, in Pustejovsky’s sense). Not only the verb, but the noun, too, is polysemous. For example, nouns like “dinner” and “nap” exhibit systematic polysemy between a process/activity and a result/product reading. (Cf: dinner lasted 3 hours=activity; dinner was on the table=product.)

So the question is, for each of the phrases, which noun reading do we get with which verb? In other words, the goal is not only to disambiguate the verb but also the noun.

WordNet generally does not include collocations or phrasemes like “make a remark” and “take a walk,” because the lexemes in WordNet’s synsets should be treatable as units by NLP systems. But a system that considers “make a remark” as internally unmodifiable will have problems dealing with tokens like “make a nasty remark” or “remarks were made.”

We first plan to collect a large number of phrasemes like “make a remark,” “take a walk,” and “have a surprise.” Next, we classify the expressions in terms of their semantics. For example, in “make a remark/comment/point/joke,” the object nouns denote a linguistic expression, whereas in “make a mistake/blunder/error/faux pas” the noun denotes a kind behavior. The verbs in these phraseme classes have a different semantics, too. In “make a comment/joke” etc. the verb means “create mentally,” whereas in “make a mistake/blunder” etc. “make” means “commit” or “perform.” In phrases like “have a surprise/shock/...,” the verb means “suffer” or “undergo,” and the noun denotes a mental state or feeling. The full semantics of the phrase will be expressed in a template logical expression in KIF and using SUMO terms. Spaces in the template will be left to fill in with the contents of slots in the parse frame. As a simplified example, “John takes a walk.” would be parsed into a frame like [John, subject][takes a walk, VP template 547] which would be keyed to a logical template below left, which would be filled in with the results of the parse and combined with the logical output of word-level interpretation to yield the logic expression at below right

```
(exists (?walk <subject>)           (exists (?walk ?john)
  (and                                (and
    (instance ?walk Walking)            (instance ?walk Walking)
    (agent ?walk <subject>)))           (instance ?john Human)
                                        (names ‘‘John’’ ?john)
                                        (agent ?walk ?john)))
```

References

1. Church, K. W. and Hanks, P., (1990). Word association norms, mutual information and lexicography. Computational Linguistics, 16(1):22–29.
2. Dowty, D., 1991: Thematic Proto-Roles and Argument Selection, Language 67, 547–619.
3. Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database (language, speech, and communication). Cambridge, MA: MIT Press.
4. Francis, W., and Kucera, H., (1964). Brown Corpus Manual. Revised 1979. Available at `http://www.hit.uib.no/icame/brown/bcm.html`.
5. Genesereth, M., (1991). “Knowledge Interchange Format”, In Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning, Allen, J., Fikes, R., Sandewall, E. (eds), Morgan Kaufman Publishers, pp. 238–249.
6. Kamp, H. & Reyle, U. (1993). From discourse to logic. New York: Kluwer Academic Publishers.
7. Melc’uk, I., (1998). Collocations and Lexical Functions. In: Cowie, Ed. 23–53.
8. Miller, G., (1995). WordNet: A Lexical Database for English. Communications of the ACM, Vol. 38 No. 11, 39–41.
9. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1993). “Introduction to WordNet: An On-line Lexical Database.”.

10. Murray, W. R., Pease, A., and Sams, M. (2003). Applying Formal Methods and Representations in a Natural Language Tutor to Teach Tactical Reasoning. 11th International Conference on Artificial Intelligence in Education (AIED) conference in Sydney. pp 349–356. IOS Publications.
11. Niles, I. & Pease A., (2001). "Towards A Standard Upper Ontology." In Proceedings of Formal Ontology in Information Systems (FOIS 2001), October 17–19, Ogunquit, Maine, USA, pp. 2–9. See also `http://ontology.teknowledge.com`.
12. Niles, I., & Pease, A., (2003). Mapping WordNet to the SUMO Ontology. Proceedings of the IEEE International Knowledge Engineering conference, Las Vegas, NV, June 23–26.
13. Pease, A., and Murray, W., (2003). An English to Logic Translator for Ontology-based Knowledge Representation Languages. In Proceedings of the 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China.
14. Sevcenko, M.: Online Presentation of an Upper Ontology, In: Proceedings of Znalosti 2003, 19–21 February 2003, Ostrava, Czech Republic.
15. Sowa, J. F.. (1999). Controlled English.
Available at `http://users.bestweb.net/~sowa/misc/ace.htm`.
16. Vendler, Z., 1967: Verbs and Times, in Linguistics in Philosophy, Cornell University Press, Ithaca, NY.