2010 IEEE Fourth International Conference on Semantic Computing

Reference Resolution Supporting Lexical Disambiguation

Marjorie McShane, Stephen Beale and Sergei Nirenburg
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, Maryland, USA
{marge,sbeale,sergei}@umbc.edu

Abstract— **This paper describes ongoing work in carrying out the semantic analysis of texts and reference resolution in a control structure that permits each process to inform the other, rather than in a more traditional, unidirectional fashion (semantics followed by reference resolution). We concentrate on situations in which a polysemous predicate cannot be lexically disambiguated until the meaning of one of its arguments has been specified, and that can only be accomplished with the help of reference resolution procedures. As a sidebar, we briefly introduce our "feature value bundling" approach to configuring reference resolution engines without the need for large annotated corpora.**

Keywords-NLP; semantics; reference resolution

I. INTRODUCTION

This paper presents our ongoing work on interleaving the processing of semantic analysis and reference resolution such that each can inform the other in the most beneficial way to support the reasoning of human-like intelligent agents. The paper begins with some definitions, background notes about the OntoSem environment, and our basic approach to lexical disambiguation and reference resolution. We then present the algorithm for resolving our selected class of phenomena, followed by examples of its usage in the system. We conclude with future directions of work.

We define semantic analysis as the interpretation of text meaning as rendered using an unambiguous metalanguage – in our case, the text meaning representation language of the OntoSem environment [1]. We define reference resolution as anchoring the *meaning* of each referring expression (RE) in the mental model of the intelligent agent processing the text, such that new information about entities can supplement or amend old information (if any), leading to memory population not unlike what a human would carry out [2]. Our treatment of reference, therefore, goes beyond the typical *co*reference task of NLP in several ways: we treat all referring expressions, not just the subset of pronouns typically treated in the well-known pronominal coreference task [3]; we go beyond establishing textual coreference to anchoring meaning in an agent's memory; we seek to achieve full understanding of coreference relations across texts, as realized by anchoring coreferential instances of entities and events to the same anchors in memory; and we are developing methods of treating residual referential underspecification and ambiguity that parallel our methods of treating residual lexical ambiguity.

In the OntoSem environment, the text analyzer takes as input raw text and carries out its tokenization and morphological, syntactic, semantic and pragmatic analysis to yield text meaning representations. Text analysis in OntoSem relies on: the OntoSem language-independent ontology of over 8,000 concepts, each of which is described by a large number of properties whose values can be locally defined or inherited; the OntoSem lexicon of English of about 35,000 senses that contains linked syntactic and semantic zones, the latter of which uses ontological concepts to describe word meaning; agent memory, also called the fact repository; the OntoSem text analyzers; and the text meaning representation language itself, which is the unambiguous metalanguage for representing text meaning in all resources and in the automatically generated text meaning representations.

Our basic approach to lexical disambiguation is to use mutual constraints of predicates and their arguments in a bidirectional way. For example, the unambiguous meanings of the subjects and direct objects in (1) and (2) permit the analyzer to automatically understand that the highly ambiguous verb *have* in (1) means the event INGEST, whereas in (2) it is used as a light verb that, in conjunction with the direct object *migraine*, means the event MIGRAINE.

1. The woman had a burger.
2. The woman had a migraine.

What makes this disambiguation possible is the fact that the lexicon includes different senses of the word *have* that include mutually exclusive semantic constraints: one expects the THEME to be an INGESTIBLE whereas another expects the THEME to be a DISEASE or SYMPTOM (there are many more sense of *have* as well, many of them constructions and idioms).

Similarly, an unambiguous predicate head can imply the meaning of its arguments: e.g. in (3), even though the system cannot know what the invented word *trala* means, it knows that it must be some SURFACE-OF-OBJECT because the ontological concept TILE-EVENT (the only meaning of 'tile' in our lexicon) includes the specification that the THEME of TILE-EVENT is SURFACE-OF-OBJECT.

3. The workmen tiled the trala.

978-0-7695-4154-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICSC.2010.103

For further details on semantic analysis in OntoSem, see [4],

Our basic approach to reference resolution is what we call "feature value bundling" [5]. We have compiled a large inventory of features of REs that are relevant for predicting which of the candidate sponsors for a referring expression is the actual sponsor. These features range from "surfacy" (e.g., the respective gender and number of the RE and the candidate), to syntactic (e.g., the respective syntactic functions of the RE and the candidate), to semantic (e.g., the meaning of the events selecting the candidate and the RE as their case-roles) to pragmatic (e.g., the location of speaker changes in a dialog). We then manually create combinations of coreference-promoting feature values – what we call "feature value bundles" – that we believe will have significant power to predict the correct sponsor. We then vet those hypotheses using a corpus and assign each bundle a score indicating the confidence of prediction attested by the corpus. This feature value bundling methodology was originally created to support research and development on the more difficult and less studied referring expressions, like *it, this* and *that,* for which no adequate annotated corpora exist; however, we are finding it useful for a wide range of referring expressions. As in many approaches to reference resolution, our system keeps a running list of REs that might function as sponsors for later REs; however, unlike most systems, ours stores the candidates both as strings and as instances of ontological concepts. As such, our heuristics can refer to both the form and the meaning of referring expressions.

The nature of the work reported here is linguistic analysis aimed at near-term implementation for the natural language support of cognitively human-like intelligent agents functioning in working applications. Although there are many phenomena related to the interaction between lexical disambiguation and reference resolution, here we will concentrate on just one: situations in which a polysemous predicate cannot be lexically disambiguated until the meaning of one of its arguments has been specified, and that can only be accomplished with the help of reference resolution procedures. We will use an example-based methodology of description and minimize the use of formalism in order to emphasize that the approach is not system-specific but, rather, is likely to be required by any system seeking to carry out both semantic analysis and reference resolution.

II. The Algorithm and Examples

The processing algorithm we will be discussing is as follows:

If a predicate (verb) cannot be confidently disambiguated using bidirectional constraints with its arguments
  And if one or more of its arguments is an underspecified referring expression (e.g., *it*)
    And if exactly one resolution of the pronoun (based on coreference with candidate sponsors) leads to a confident semantic unification with exactly one of the meanings of its selecting predicate
    Then establish the given coreference and use the unifying predicate analysis
    Else
      If more than one resolution of the pronoun leads to a confident semantic unification with one or more meanings of the selecting predicate
        And if the underspecified referring expression can be resolved confidently using non-semantic methods (i.e., high-confidence, "surfacy" feature bundles)
        Then resolve the referring expression and use its meaning to help disambiguate the predicate
… ; we do not discuss additional conditions in this paper

As promised above, we will use an example-driven method of illustrating the algorithm since we believe that examples can be the best way to succinctly and informally convey how a system operates. We will first walk through one example, then present a number of other examples with only minimal commentary.

At a first glance, one might not even detect the lexical or referential ambiguity in an example like (4) since we as people resolve ambiguity so effortlessly. However, for a system attempting to disambiguate every aspect of an input, the challenges of resolving polysemous *save* and underspecified *it* are formidable.

(4) A dialog box will open and ask if you want to open *the file*, save **it** or cancel.

Lexically, *save* has at least 3 meanings: 'rescue from harm', 'store in a digital file' and 'store for the future'. *It* can refer to 'a dialog box', 'the file', or even the propositions 'a dialog box will open', 'ask if you want to open the file', or 'you want to open the file'. (Reference to propositions, realized as spans of text, has received relatively little attention in NLP but is a prominent phenomenon in language use; see, e.g., [2] and [5].)

Table 1 shows the three verbal senses of *save* that are recorded in our lexicon. The middle column presents a very abbreviated version of their semantic descriptions that includes only those aspects of meaning central to an understanding of this disambiguation task. For example, the description of *save-v1* indicates that this word sense has the meaning of the concept RESCUE and that its THEME should be some kind of ANIMAL – more specifically, any lexical item mapped to the concept ANIMAL or any of its descendants. The third column provides an example of usage, since the meaning of ontological concepts cannot properly be understood without consulting the concept's description in the ontology. (Although concept names look like and are typically similar to the meaning of English words, they are not English words.)

Table 1: Senses of *save*

v1	RESCUE (THEME ANIMAL)	save a bear cub
v2	SAVE-COMPUTER-DATA (THEME COMPUTER-DATA, COMPUTER-FILE)	save a file
v3	STORE-FOR-FUTURE (THEME OBJECT)	save a seashell

When the analyzer encounters (4), all three meanings of *save* are available. Since one of the arguments of *save* is an underspecified referring expression (*it*), the analyzer will attempt every available resolution of *it* – using the meanings of the candidate sponsors in the candidate list – and see if any of them makes a strong suggestion about what *save* means.

The analyzer will begin by coreferring *it* with *dialog box*, which is semantically analyzed as COMPUTER-DIALOG-BOX. This meaning does not meet the narrow constraints on the THEME of *save-v1* or *save-v2* because COMPUTER-DIALOG-BOX is not a descendant of ANIMAL, COMPUTER-DATA or COMPUTER-FILE. This meaning *does* meet the broad constraint on the THEME of *save-v3*, making this sense selection a viable option; however, it is not a confident option because there is a great ontological distance between the very specific concept COMPUTER-DIALOG-BOX and the very general constraint OBJECT. Next the analyzer will corefer *it* with *the file*, which was disambiguated in its own clause as COMPUTER-FILE. COMPUTER-FILE is a direct match of a selectional constraint for the THEME of sense 2 and offers a high-confidence coreference link that will be selected over the low-confidence link offered by sense 3. The analyzer will not, in the case, consider the text span propositions to be viable candidates because propositions typically refer to EVENTs and none of our senses of *save* expects an EVENT as its THEME. In sum, for example (4) semantic analysis is sufficient to both resolve the meaning of the referring expression and choose a meaning of the selecting verb.

However, imagine that the bidirectional semantic correlation of the head and its argument could not confidently suggest exactly one resolution for example (4). In that case, the system would first attempt to resolve the pronoun using "surfacy" heuristics combined in the feature bundling strategy; if successful, it would use the pronoun's meaning to unidirectionally impose a meaning on its selecting predicate. As it turns out, our example matches a feature value bundle that that has been attested to have very high predictive power of coreference:

- C (the candidate) is the most recent candidate that matches the RE in gender/number/animacy
- C and RE have matching syntactic functions
- C and RE are in a VP conjunction structure
- C and RE have matching case-roles (both are THEMEs under any semantic interpretation of the predicate)

This feature value bundle would confidently select *the file* (COMPUTER-FILE) as the resolution of the pronoun *it*. Once the meaning of *it* was established, it would be matched to the constraint on the theme of SAVE-COMPUTER-DATA in sense 2, and that meaning would be unidirectionally imposed on the predicate *save*.

At this point, one might ask, Why not always use the computationally less expensive feature-bundling strategy first, before resorting to more expensive semantic analysis? One could, but we choose not to because (a) in our environment all texts are processed semantically anyway, so unless we do some selective processing of sentences in a corpus, we will always have an antecedent list containing both strings and concepts; (b) we believe that semantic evidence, when available, is stronger and more certain than any other kind of evidence; and (c) in some contexts there will be no available feature bundles that can confidently predict the resolution of the pronoun in isolation.

The remaining examples are presented using the same formalism as above.

(5) *Combat stress* is a natural result of the heavy mental and emotional work required when facing danger in tough conditions. Like physical fatigue and stress, handling *combat stress* depends on the level of your fitness/training. *It* can come on quickly or slowly, and **it** gets better with rest and replenishment.

Table 2. Senses of *get better*

v1	represented as an increase in the value of evaluative modality; its THEME is any EVENT	His sax playing got better.
v2	HEAL (THEME ANIMAL, DISEASE)	His cough got better.

Analysis: If *it* is coreferential with *combat stress* (COMBAT-STRESS, which is a type of DISEASE), then HEAL (*get better-v2*) has a perfect filler for its THEME case-role.

A corroborating strongly predictive feature bundle:

- C is a matching pronoun
- C is the most recent candidate that matches in gender/number/animacy
- C and RE have matching syntactic functions (both are subjects)
- C and RE are in a VP conjunction structure with 'and'
- long chain of coreference (C is part of a 3-member chain even before coreferring with RE)

(6) When it comes to the causes of *autism*, here are the facts: we know *it* runs strongly in families, although **it** is not strictly inherited like muscular dystrophy or hemophilia.

Table 3. Senses of *inherit*

v1	INHERIT-GOODS (THEME FINANCIAL-OBJECT)	inherit a fortune
v2	INHERIT-GENETICALLY (THEME GENE, DISEASE, CHARACTERISTIC, etc.)	inherit cystic fibrosis

Analysis: If *it* is coreferential with *autism* (AUTISM, which is a type of DISEASE), then INHERIT-GENETICALLY (*inherit-v2*) has a perfect filler for its THEME case-role.

A corroborating strongly predictive feature bundle:

- C is a matching pronoun
- C is the most recent one that matches in gender/number/animacy
- C and RE have matching syntactic functions (both are subjects)
- C and RE are in a main/subordinate relationship
- chain of coreference (C is part of a 2-member chain even before coreferring with RE)

(7) *Primitive Medicine* is timeless. *It* is as old as the Paleolithic cave-dwellers. *It* is as new as today. Early

evidences of *its* practice can be traced back 10,000 years. Yet **it** is being practiced in some part of the world at this very hour…

Table 4. Senses of *practice*

v1	PRACTICE (THEME EVENT)	practice soccer
v2	PLAY-MUSICAL-INSTRUMENT (THEME MUSICAL-INSTRUMENT)	practice the trumpet
v3	HAS-RELIGION (RANGE RELIGION)	practice Catholicism
v4	WORK-ACTIVITY (THEME FIELD-OF-STUDY)	practice allopathic medicine

Analysis: If *it* is coreferential with *primitive medicine* (FIELD-OF-MEDICINE, which is a type of FIELD-OF-STUDY), then WORK-ACTIVITY (*practice-v4*) has a perfect filler for its THEME case-role.

A corroborating strongly predictive feature bundle:

- long chain of coreference (C is part of a 4-member chain even before coreferring with RE)
- most members of the chain have matching syntactic function (subject)

(8) [Lord Illingworth]: Then why does he write to me? [Mrs. Arbuthnot]: What do you mean? [Lord Illingworth]: What *letter* is this? [Mrs. Arbuthnot]: *That*—is nothing. Give *it* to me. [Lord Illingworth]: *It* is addressed to me. [Mrs. Arbuthnot]: You are not to open **it**.

Table 5. Senses of *open*

v1	OPEN (THEME BAG, WINDOW, ENVELOPE, LETTER, etc.)	open a letter
v2	cause to BE-AVAILABLE	open a road
v3	begin + EVENT	open a conference (with a speech)

Analysis: If *it* is coreferential with the chain of coreferred elements meaning LETTER, then OPEN (*open-v1*) has a perfect filler for its THEME case-role.

A corroborating strongly predictive feature bundle:

- C is a matching pronoun
- C is the most recent one that matches in gender/number/animacy
- long chain of coreference (C is part of a 4-member chain even before coreferring with RE)

III. DISCUSSION

One might say that a main thread in the overall program of research and development in the OntoSem environment is that all aspects of natural language processing are connected and are aimed at populating an agent's memory so that it can carry out sophisticated reasoning with the results mimicking those of people. As such, we do not draw hard lines between traditionally divided realms like syntactic analysis, word sense disambiguation and reference resolution. Due to feasibility constraints, we cannot, it is true, allow heuristics from all modules of text processing to fire at once: e.g., some part-of-speech decisions are made before syntactic analysis is launched, less probable syntactic parses are removed before semantic analysis occurs, etc. However, we attempt to postpone difficult cases of upstream decision-making until semantics can act as an arbiter. The same is true of reference resolution. It would be infeasible to attempt to resolve reference without the benefit of any semantic analysis decisions having been made; however, this does not mean that semantic analysis must be completed, with no outstanding options, before reference resolution is attempted. A real key to achieving outstanding text analysis, we believe, is to be able to automatically evaluate confidence in each stage of analysis, and leave low-confidence decisions open until later stages of processing can register a vote. At the time of writing, we are working on developing such confidence-assigning engines for each stage of processing.

We have been using our text processing capabilities in real-world applications – most recently, in dialog systems in the medical domain [6]-[7]. The applications in question – a medical education system called Maryland Virtual Patient and a CLinician's ADvisor called CLAD – are prime examples of applications for which high-quality, deep text understanding are needed, the processing of difficult phenomena cannot be postponed, and the returns of developing methods for effectively treating difficult phenomena should be great. In short, the work reported here is not being carried out in a conceptual bubble – it is being incorporated into working, forward-looking systems.

REFERENCES

[1] Nirenburg, S., Raskin, V., 2004. *Ontological Semantics*, The MIT Press, Cambridge, Mass.

[2] McShane, Marjorie, 2009. Reference resolution challenges for an intelligent agent: The need for knowledge. IEEE Intelligent Systems, vol. 24, no. 4, pp. 47-58.

[3] Hirshman, L., Chinchor, N., 1998. MUC-7 coreference task definition. Version 3.0. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Applications International Corporation.

[4] Beale, S., 1997. Using branch-and-bound with constraint satisfaction in optimization problems. Proc. AAAI-97. Providence, Rhode Island.

[5] McShane, Marjorie. 2009, Advances in difficult aspects of reference resolution: Working Notes. ILIT Working Paper #01-09, Nov. 18, 2009 (62 pp.) Available at http://ilit.umbc.edu/MargePub/SGER-As-Paper.pdf.

[6] Nirenburg, S., McShane, M., Beale, S., 2008. A simulated physiological/cognitive "double agent". In: Workshop on Naturally Inspired Cognitive Architectures at AAAI 2008 Fall Symposium.

[7] McShane, M., Nirenburg, S., Beale, S., 2008. Two kinds of paraphrase in modeling embodied cognitive agents. In: Workshop on Biologically Inspired Cognitive Architectures, AAAI 2008 Fall Symposium.