Enhancing Recall in Information Extraction through Ontological Semantics

Sergei Nirenburg, Marjorie McShane and Stephen Beale
Institute for Language and Information Technologies
University of Maryland, Baltimore County
Baltimore, MD, USA

1. Introduction.

We proceed from the assumption that extracting and representing the meanings of texts that serve as sources for information extraction will enhance the latter's quality. In particular, we believe that resolving reference in these texts will lead to higher levels of recall in IE because additional information will become available for extraction once it can be captured not simply by matching character strings in the IE template but by knowing that *George W. Bush, President Bush, the current president of the US, the leader of the free world*, and *the winner of the 2000 National election* all refer to the same entity and, therefore, whatever information in the text is introduced by any of the above (and other reference means, notably, pronominalization and ellipsis) is relevant.

2. The Environment

At the core of our environment are general-purpose syntactic and semantic analyzers developed over the past 10 years at the Computing Research Lab of New Mexico State University and the University of Maryland Baltimore County. We will very briefly describe the semantic analysis process (a detailed description can be found in Nirenburg and Raskin 2003), including the treatment of reference, and then relate it to the task of enhancing recall in information extraction.

Ontological-semantic processing for text analysis relies on the results of a battery of pre-semantic text processing modules (see Figure 1). The output of these modules provides input to and background knowledge for semantic analysis.

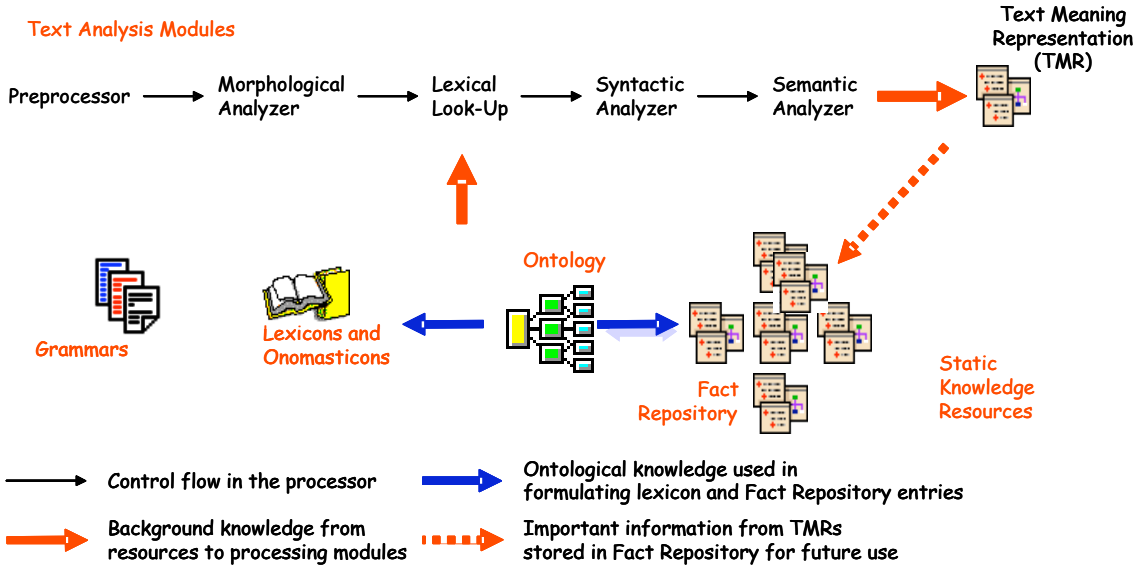

Figure 1. Ontological-semantic processing for text analysis.

Semantic analysis takes as input results from the earlier stages of processing and produces a **text meaning representation** (**TMR**). The central task for semantic analysis is to construct an unambiguous propositional meaning by processing selectional restrictions, which are listed in the ontology and the semantic zones of lexicon entries. Other issues include treating such phenomena as aspect, modality and non-literal language (which, incidentally, is important for the treatment of reference as well), and building a discourse structure associated with the basic propositional structure of the text.

The major "static knowledge sources" for text analysis are: the **TMR language**, the **ontology**, the **fact repository** and a **lexicon** that includes an **onomasticon**.

The ontology provides a metalanguage for describing the meaning of lexical units of a language as well as for the specification of meaning encoded in TMRs. The ontology contains specifications of concepts corresponding to classes of things and events in the world. Formatwise, the ontology is a collection of frames, or named collections of property-value pairs. The ontology contains about 5,500 concepts, each of which has, on average, 16 properties defined for it. Figure 2 shows a portion of the description of the concept ROOM (not all inheritance is shown). Small caps are used to distinguish ontological concepts from English words.

Concept: ROOM		
DEFINITION	VALUE	area set off by walls within a building
IS-A	VALUE	INTERIOR-BUILDING-PART
SUBCLASSES	VALUE	APARTMENT
		BATHROOM
		DINING-ROOM
		HOTEL-ROOM
		KITCHEN
		PUBLIC-ROOM
		STUDY-ROOM
HAS-OBJECT-AS-PART	SEM	CEILING
		CLOSET
		FLOOR
		WALL
		WINDOW
PART-OF-OBJECT	SEM	BUILDING
		LODGING-CORPORATION
		SERVICE-BUILDING
Inherited from: PLACE		
SOURCE-OF	DEFAULT	CHANGE-LOCATION
		MOTION-EVENT
		TRANSFER-OBJECT
	SEM	EVENT
DESTINATION-OF	DEFAULT	MOTION-EVENT
		TRANSFER-OBJECT
	SEM	EVENT
PATH-OF	SEM	CHANGE-LOCATION
		EVENT

Figure 2. Part of the description of the ontological concept ROOM (not all inheritance is shown).

This ontology has been shown to be able to represent the meanings of over 40,000 entries in a Spanish lexicon. We also have an English lexicon of about 45,000 entries and have developed an efficient methodology for the acquisition of the ontology and the lexicon (Nirenburg and Raskin 2003, Chapter 9).

The fact repository contains a list of remembered instances of ontological concepts. For example, whereas the ontology contains the concept CITY, the fact repository contains entries for London, Paris and Rome; and whereas the ontology contains the concept SPORTS-EVENT, the fact repository contains an entry for the Salt Lake City Olympics. A sample fact repository entry is shown in Figure 3.

HUMAN-33599		
NAME	George W. Bush	
ALIAS	George Bush, President Bush, the president of the United States, the US president, ...	
SOCIAL-ROLE	PRESIDENT	
GENDER	male	
NATIONALITY	NATION-1	(i.e., The United States of America)

DATE-OF-BIRTH	July 6, 1946	
SPOUSE	HUMAN-33966	(i.e., Laura Bush)

Figure 3. An excerpt from a sample entry in the fact repository.

The ontological semantic lexicon contains not only semantic information, it also supports morphological and syntactic analysis. Semantically, it specifies what concept, concepts, property or properties of concepts defined in the ontology must be instantiated in the TMR to account for the meaning of a given lexical unit of input.

The entries in the onomasticon directly point to elements of the fact repository. Onomasticon entries are indexed by name (e.g., *New York*), while their corresponding entries in the fact repository are named by appending a unique number to the name of the ontological concept of which they are instances (e.g., Detroit might be listed as CITY-213).

3. Resolving Reference

Most NLP work in reference resolution focuses on finding *textual* antecedents (or postcedents) for *pronouns* using *knowledge-lean* methods. For us, by contrast, resolving reference involves linking *every referring entity* to its *real-world anchor in the FR* using *a broad range of semantic knowledge and heuristic clues.* We present just a sampling of reference issues with their required processing and expected output.

Pronouns. Resolving a reference to a pronoun like *he* requires not only linking this pronoun to a coreferential element in the text (e.g., *The President*) but further linking it to its real-world entity stored in the FR (e.g., George W. Bush). We supplement the same types of heuristics (e.g., text distance, syntactic structure) as most researchers but supplement them with ontological-semantic analysis of candidate coreferential entities.

Approximations. Resolving approximations requires positing a concrete range whose calculation depends upon semantic heuristics: e.g., *around 8:00* might be 7:45-8:15, whereas *around 8:06* will be *8:05-8:07.* Whereas we have found that a 7% rule works quite well in most cases (i.e., expanding the range to 7% of the given number in each direction), exceptions – like *around 8:06* -- must be detected and treated separately..

Relative Scalars. Resolving relative scalars (e.g., *expensive*) requires selecting the relevant range on the scale defined for modified entity. For example, an *expensive bomber* costs far more than an *expensive pistol*, which can be reasoned based on the fact that the property COST (which indicates the range of typical cost) in the ontological frame for the concept MILITARY-JET has a numerical filler that is orders of magnitude higher than the same property for GUN.

Definite Descriptions. Resolving reference to definite descriptions (i.e., noun phrases with *the*) requires first determining if the signals coreference. Non-coreferential definite descriptions include always-definite NPs (*the winter; on the other*) and NPs used in certain constructions, like appositives (*Bill Gates, the chairman of Microsoft*) and restrictive modification (*the hope of ensuing peace*). All other definite descriptions require coreference resolution, be they identical to their coreferent (*the conflict... the conflict*), synonymous (*the treaty...the pact*) in a hypernym/hyponym relationship (*the bank... the financial institution*), in a meronym relation (*I walked in the room and found the window open*), etc. We have the conceptual infrastructure to carry out such analysis, as well as automatically corefer, e.g., *the move* in (2) with the meaning of the entire preceding sentence (1); our current work focuses on improving our algorithms to best exploit and extend these resources.

(1) The Standard & Poor's Corporation, a leading credit rating agency, cut its ratings on the debt of United to "default," its lowest ranking.

(2) **The move** by S.& P. helped fuel speculation that United, the world's second-biggest airline, was on the verge of seeking bankruptcy court protection from its creditors.

Syntactic and Semantic Ellipsis. Syntactic ellipsis is the non-representation of semantic information that is signaled by a syntactic gap: e.g., *Italy voted against the proposal and France did [vote against the proposal] too.* Semantic ellipsis is similar but without the syntactic gap to act as a trigger: *The subcommittee started with [a discussion of, debate about] the gun issue.* Ontological semantic analysis permits us to resolve ellipsis – sometimes quite specifically and other times more generally – based on the lexically stipulated selectional restrictions of text entities. For example, since we know that *start* regularly triggers semantic ellipsis (just like *finish [the pizza], prefer [Hemingway],* etc.), we created a lexical sense of this word that expects a PHYSICAL-OBJECT as a complement and explicitly calls a procedure that seeks to resolve the missing EVENT based on the semantic collocation between the overt text elements (*subcommittee / gun*). In other words, the given lexicon sense posits an EVENT whose agent is COMMITTEE (the mapping for *subcommittee*) and whose theme is GUN (the mapping for *gun*), then the semantic analyzer searches the ontology for the EVENT that best meets these selectional restrictions. Positing a lexical sense that expects a PHYSICAEL-OBJECT as a complement is not strictly necessary: the semantic analyzer has recovery procedures that would be triggered when the selectional restrictions for the first sense of *start* (start + EVENT 'start reading') were violated. However, encoding expectations about ellipsis in the lexicon, to the extent reasonable, helps the analysis process by reducing the search space for error recovery.

Resolving reference is arguably one of the most difficult aspects of text processing, alongside the metaphor and metonymy. We have spread our net wide in attempting to treat reference issues not only because we believe we have the infrastructure to achieve some success but also because we consider this aspect of text processing an opportunity to improve the results of applications like extraction, summarization and question-answering, where reference relations cannot simply be "carried over" – as is sometimes the case in machine translation – but must be explicitly resolved for each referring entity so that sentences containing those entities can be fully exploited.

4. IE in Ontological Semantics

Unlike the rest of IE systems, information extraction that uses the mechanisms and knowledge sources of Ontological Semantics operates against the results of ontological-semantic text analysis, the TMRs, not against open text. In the TMRs, ambiguity and reference are resolved, to the best of the analyzer's ability; ontological and extra-ontological semantic information is encoded, and referring expressions are linked to their corresponding entities (typically, instances of ontological concepts). We are currently conducting experiments comparing the IE against TMRs before and after reference resolution. We are using texts from the domain of business (specifically, bankruptcy reports) and our hypothesis is that results of reference resolution should lead to enhancement in the levels of recall in IE. We hope to present the initial results of our experimentation at the conference.

References.

Nirenburg, S. and V. Raskin. 2003. Ontological Semantics. MIT Press. Forthcoming.