Efficient Semantic Inference over Language Expressions

Roy Bar-Haim and Ido Dagan
Computer Science Department, Bar-Ilan University
Ramat-Gan 52900, Israel
{barhair, dagan}@cs.biu.ac.il

1 The Need for an Inference Formalism over Language Expressions

According to the traditional formal semantics approach, inference is conducted at the logical level. Texts are first translated into some logical form and then new propositions are inferred from interpreted texts by a logical theorem prover. However, text understanding systems usually employ shallower lexical and lexical-syntactic representations, sometimes augmented with partial semantic annotations like word senses, named-entity classes and semantic roles. While practical semantic inference is mostly performed over linguistic rather than logical representations, such practices are typically partial and quite ad-hoc, and lack a clear formalism that specifies how inference knowledge should be represented and applied.

2 Our Approach - a Proof System over Parse Trees

In (Bar-Haim et al., 2007) we proposed a step towards filling this gap, by defining a formalism for semantic inference over parse-based representations. All semantic knowledge required for inference is represented as *entailment rules*, which encode parse tree transformations, and each rule application generates a new consequent sentence (represented as a parse tree) from a source tree. Figure 1 shows a sample entailment rule, representing a passive-to-active transformation.

From a knowledge representation and usage perspective, entailment rules provide a simple unifying formalism for representing and applying a very broad range of inference knowledge. Some examples of this breadth are illustrated in Table 1. From a knowledge acquisition perspective, representing entailment rules at the lexical-syntactic level allows easy incorporation of rules learned by unsupervised methods, which seems essential for scaling inference systems. Interpretation into stipulated semantic representations, which is often difficult and is inherently a supervised semantic task for learning, is circumvented altogether. Our overall research goal is to explore how far we can get with such an inference approach, and identify the scope in which semantic interpretation may not be needed.

3 Inference Efficiency

Formally, application of an entailment rule corresponds to generation of a new sentence, a consequent, semantically entailed by the source sentence. The inferred consequent itself may be the source of further rule applications and so on. However, explicitly generating a new sentence (or parse tree) for each rule application may quickly lead to exponential explosion. Consider, for example, a sentence containing five content words, each one having two synonyms. The number of derivable sentences would be 3^5 (including the source sentence). Thus, representing each entailed sentence explicitly leads to severe efficiency problems. Intuitively, we would like to add for each rule application just the entailed part (e.g. the synonymous words) to the source sentence representation. However, we still want the inference process to be formulated over individual sentences, rather than over some superposition of sentences whose semantics is unclear.

In (Bar-Haim et al., 2008) we proposed a solution to this inherent problem. We present a novel data structure, termed *compact forest*, which allows effi-

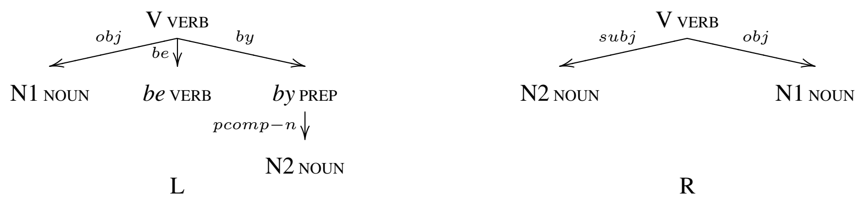

Figure 1: An entailment rule: passive to active transformation

Rule Type	Sources	Examples
Syntactic	Manually-composed	Passive/active, apposition, relative clause, conjunctions
Lexical -Syntactic	Learned with unsupervised algorithms (DIRT, TEASE), and derived automatically by integrating information from WordNet, Nomlex, and VerbNet	*X's wife, Y* $\rightarrow$ *X is married to Y* *X bought Y* $\rightarrow$ *Y was sold to X* *X is a maker of Y* $\rightarrow$ *X produces Y*
Lexical	WordNet, Wikipedia	*steal*→*take*, *Albanian*→*Albania* *Janis Joplin*→*singer*, *Amazon*→*South America*
Polarity	Manually-composed, utilzing VerbNet and PARC's polarity lexicon	Verbal negation, modal verbs, conditionals, verb polarity

Table 1: Representing diverse knowledge types as entailment rules.

cient generation and representation of entailed consequents, where each consequent is represented by a dependency tree. We show how all inference operations defined in our framework, including an extension to handle co-reference, can be implemented over compact forests. Figure 2 shows a sample compact forest, containing both the source sentence and the sentence resulting from the application of the rule in Figure 1 to it.

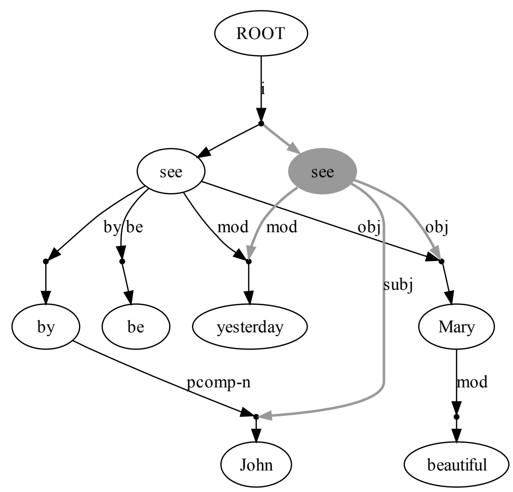

Figure 2: A compact forest containing both the source sentence "beautiful Mary was seen by John yesterday", and the inferred sentence "John saw beautiful Mary yesterday" resulting from the application of the rule in Figure 1. Parts added to the compact forest by this rule application are shown in gray.

4 Evaluation

In (Bar-Haim et al., 2007) we present an empirical evaluation of our proof system, based on relation extraction from a large corpus. In (Bar-Haim et al., 2008), our proof system, complemented with an entailment classifier measuring the remaining gap between the source and target texts, has been evaluated on RTE datasets, achieving quite competitive results on RTE-3. In both cases, the contribution of the various knowledge sources for entailment rules has been shown.

References

Roy Bar-Haim, Ido Dagan, Iddo Greental, and Eyal Shnarch. 2007. Semantic inference at the lexical-syntactic level. In *Proceedings of AAAI*.

Roy Bar-Haim, Jonathan Berant, Ido Dagan, Iddo Greental, Shachar Mirkin, Eyal Shnarch, and Idan Szpektor. 2008. Efficient semantic deduction and approximate matching over compact parse forests. In *Proceedings of TAC*. To appear.