Customisable Semantic Analysis of Texts

Vivi Nastase and Stan Szpakowicz

School of Information Technology and Engineering
University of Ottawa, Ottawa, Ontario, Canada
{vnastase,szpak}@site.uottawa.ca

Abstract. Our customisable semantic analysis system implements a form of knowledge acquisition. It automatically extracts syntactic units from a text and semi-automatically assigns semantic information to pairs of units. The user can select the type of units of interest and the list of semantic relations to be assigned. The system examines parse trees to decide if there is interaction between concepts that underlie syntactic units. Memory-based learning proposes the most likely semantic relation for each new pair of syntactic units that may be semantically linked. We experiment with several configurations, varying the syntactic analyzer and the list of semantic relations.

1 Introduction

Deep processing of natural language data often requires suitably annotated data. Recognition of semantic relations is such a task that benefits from the availability of annotated texts from which we can learn to analyze new data. Manual semantic annotation is a time-consuming activity, and it is seldom possible to capitalize on the annotation effort of other researchers. This is because they work with a different set of semantic phenomena, for example a different list of relations, or because they consider different types of texts or different domains. We present a customizable, domain-independent tool for certain style of semantic analysis. It relies on syntactic information usually supplied by parsing. When the tool achieves its full functionality, its user will be able to impose her own list of semantic relations, select the type of relations she is interested in (between eventsm between an event and an entity, and so on), and plug in her own parser.

Knowledge acquisition from texts spans the range between fully automatic and fully user-driven systems. Automation relies on manually built resources and on statistical or machine-learning methods that extract classifiers from annotated data. The shortcomings of such methods include high cost of annotation and low accuracy of such classifiers on new data. User-driven systems, with friendly interfaces that domain experts use to identify knowledge in texts, allow much higher accuracy (insofar as humans agree on semantic relations). On the other hand, they require time to train people with minimal AI or NLP background, and to encode knowledge.

Our approach falls between these extremes. We rely on parsers for the grammatical structure of sentences, in which we identify concepts and pair up those