Combining Techniques for Event Extraction in Summary Reports

Nancy McCracken, Necati Ercan Ozgencil and Svetlana Symonenko

Center for Natural Language Processing
School of Information Studies
335 Hinds Hall, Syracuse University
njm@ecs.syr.edu, neozgenc@ecs.syr.edu, ssymonen@syr.edu

Abstract

The semantic role labels of verb predicates can be used to define an event model for understanding text. In the system described in this paper, the events are extracted from documents that are summary reports about individual people. The system constructed for the event extraction integrates a statistical approach using machine learning over Propbank semantic role labels with a rule-based approach using a sublanguage grammar of the summary reports. The event model is also utilized in identifying patterns of event/role usage that can be mapped to entity relations in the domain ontology of the application.

Introduction

Event extraction from text is vital for understanding the meaning of text and representing it in ways that can be useful to reasoning systems that reason about information from text. This paper describes the use of an event model derived from the Propbank definitions of semantic role labels for events. It focuses on two aspects: the first is the use of combined statistical and knowledge-based techniques for event extraction from text in a summary report genre, and the second is to map the resulting event model to an entity-based ontology for reasoning about individuals.

Applications that make use of extractions from text for reasoning may need to apply extraction to many types of text genre. For example, intelligence analysts read news summaries and reports about particular areas or topics, as well as newswire stories. While event extraction has been most closely associated with newswire text, event modeling is a natural way to organize and represent information from other types of (formal) text as well, because it can be based on the verb and phrase structure of the sentence. However, for many of these applications, it is not feasible to annotate enough text in the genre to use statistical techniques for event extraction. One possible solution is to use a statistically trained extractor to extract as much as possible and to use rule- or knowledge-based techniques to close the gap for the different text genre. In the application discussed here, the documents were summary reports on individual people.

While event modeling may be a natural way to represent information that is extracted from raw text, the resulting information structure may not be the most natural for the application or reasoning system that will be using the information. In the system described here, the reasoning system is based on a domain ontology of relations about individuals. In text that contains events about the individual, another task is then to map the semantic roles of the events to the ontology relations.

Background

One resource for statistical event extraction is Propbank, which contains the annotation of text for its predicate/argument structure with respect to target verbs. (Palmer, Gildea, and Kingsbury 2005) A similar system for annotating the semantic roles of verbs has been developed in FrameNet (Baker, Fillmore, and Lowe 1998). These resources have been used as the training data for the Semantic Role Labeling (SRL) problem and have been the subject of significant research over the last several years, including the shared task of CoNLL in 2004 and 2005 (Carreras and Màrquez 2005). Event extraction has also been a part of the Automatic Content Extraction evaluation, most recently in 2005 (ACE 2005) All of these projects have primarily annotated newswire text.

Statistical systems trained on these datasets have made significant progress since their inception, from the original work in (Gildea and Jurafsky 2002) to more recent systems such as (Pradhan et al. 2005). However, these systems may not perform as well on different text genre. For example, in the CoNLL shared task of 2005, the systems were trained and tested on the Wall Street Journal (WSJ) test set, but were also then tested on a portion of the Brown corpus, which contains text from genre different from the training set. Overall the systems did not perform as well on the Brown corpus: the top score on the WSJ test set had recall of 82% and precision of 77%, resulting in an F-measure of 79, while the top score on the Brown test set had recall of 74% and precision of 63%, resulting in an F-measure of 68.

The other major issue in this project is the role that the event representation plays in the mapping between the text and the domain ontology. The automatic acquisition of semantic domain knowledge from the text has long been a goal in Information Extraction, including work by (Riloff 1996). Later work has focused on acquiring domain concepts for automatic ontology engineering, and even concept relations, as in (Maedche and Staab 2000). Other work has focused on mapping lexical items from text to an existing ontology. For example, in (Basili, Pannacchiotti, and Zanzotto 2005), the model of learning ontology concepts from text includes the ideas of gathering or clustering instances of concepts from text, including entities and events, using lexical resources to group and name these instances, and semi-automatically mapping them to concepts in the domain ontology.

Event Extraction Using Propbank

In the application described here, each text document is a summary report about an individual person. The reports contain two parts, where the first is a factual accounting of incidents in the person's life, with some similarities to a newswire style of reporting, and the second part is a summary evaluation of that individual. In this paper, we will report on how well the Propbank system of semantic role labels was able to represent the factual events in the first part of the report and contribute to the representation of the summary information.

The extraction system described here uses two main components and the first of these is a statistical system trained on the WSJ sections of Propbank. The annotations of Propbank label each of the arguments of a predicate verb with argument numbers, where the core arguments are labeled Arg0, Arg1, … Arg5. In addition, adjunctive arguments are labeled ArgM-… and of these, this system used ArgM-TMP (temporal), ArgM-LOC (locative), and ArgM-NEG (negation). For example, in the following phrase from the Propbank WSJ corpus

> *When Disney offered to pay Mr. Steinberg a premium for his shares, the New York investor didn't also pay a premium to other shareholders.*

The arguments for the first occurrence of the predicate "pay" can be annotated as

> When [Arg0 Disney] offered to [V pay] [Arg2 Mr. Steinberg] [Arg1 a premium] for [Arg3his shares], the New York investor …

These Argument labels can also be given meaning from the accompanying framesets in Propbank, where, for example, in the frameset for "pay",

Arg0 is "payer",
Arg1 is "money",
Arg2 is "recipient",
Arg3 is "commodity"

In our system these labels are used as semantic role labels in a (quantified) frame representation of the event. Currently, the only quantification used is negation.

The semantic role labeling system is based on one that was developed for the CoNLL Shared Task in 2005 (Ozgencil and McCracken 2005). In common with many statistical systems for the SRL problem, it uses a parse tree representation of the text, one classifier to identify candidate argument phrases, another classifier to suggest sets of labelings of those candidates with semantic roles for each predicate, and a final constraint resolver that picks an argument labeling that satisfies the constraints among the labels. The two classifiers in this system use the libSVM package (Chang and Lin 2001) and use parse trees from the Charniak parser. (Charniak and Johnson 2005) This SRL system performs at an F-measure of 77 (using only one type of parse tree) on the WSJ test set of the Propbank corpus.

But the application of this system to the summary report genre of this project results in greatly lower recall than on the newswire text. Analysis of example text shows that the SRL program labels semantic roles with 83% precision, which is comparable to the newswire text of the WSJ. However, it only achieves 57% recall, which is considerably lower than both the WSJ and Brown tests. But the fact that the system's shortcoming is in missing semantic role labels gives the opportunity to supply additional labels through another technique.

Knowledge-based Approach

The summary reports used in this project have many similarities to newswire articles, but there are also differences and these differences have predictable aspects. The reason is that the reports are written by a small group of people trained to write the reports in the language of a set of content guidelines and using a particular style that focuses on the individual subject of the reports. This makes it possible to use a sublanguage grammar approach to processing the text.

A sublanguage is defined as the particular language usage patterns which develop within the written or spoken communications of a community that uses this sublanguage to accomplish some common goal or to discuss topics of common interest. Early research in Sublanguage Theory, for example see (Grishman and Kittredge 1986) and (Liddy et al. 1991) has shown that there are linguistic differences amongst various types of discourse (e.g. news reports, email, manuals, requests, arguments, interviews) and that discourses of a particular type that are used for a common purpose within a group of individuals exhibit characteristic linguistic (lexical, syntactic, semantic, discourse, and pragmatic) features. In particular in these documents, several types of special purpose lexical features were normalized before text processing. Additionally, it can be observed that the syntax of sentences falls into a smaller number of grammar patterns than standard newswire text.

Analysis of the errors in the SRL shows that some errors are caused by incorrect parses and some are caused by the failure of the SRL labeling to correctly identify the labels. In either case, the analysis shows that a significant number of the failures occur as a result of a small number of sentence patterns. These patterns consist of compound verb phrases and clauses in which the syntactic subject of the sentence is distant from the predicate verb. It is also the case that the report genre contains many more third person pronouns "he" and "she" as the subject of the sentence and that the SRL system misses these as the agentive argument in some sentences.

For example, in the following sentence, the syntactic subject of the sentence "he" is neither identified as an argument of the predicate "arrest", nor of the predicate "attempt to cash".

> *He confessed to being arrested while in college, 4 years ago, for stealing a professor's payroll check and then attempting to cash this check using a false ID in the professor's name.*

And in this example, "he" is not labeled as an argument to the predicate "leaving".

> *When the police arrived, he drove off and later was involved in an accident, leaving before the police's arrival.*

Examination of the documents shows that this type of sentence, involving coordination structures and clauses, occur in significant quantities and follow a particular sentence grammar pattern. Therefore, the system also includes a rule-based extraction component that uses sublanguage grammar rules and a general semantic role labeling rule to label the semantic roles of the syntactic subject in these sentences.

The most significant part of this grammar pattern is that the subject of the sentence is a simple noun phrase and, if there is no other obvious local subject of other verbs in the remaining part of the sentence, this main subject is also the subject of those verbs. The next observation is that the main subject is almost always either an Arg0 label of other verbs, the Prototypical Agent, or an Arg1, the Prototypical Patient or Theme, depending on whether the verb is in active or passive voice. Therefore, the main role of the grammar rules is to analyze the voice of the verb according to a very local analysis of the words surrounding the verb; typically, some form of the word "be" or "have" precedes the verb within a few words. Here is an example passive voice rule pattern to match text words with Penn Treebank style POS tags

```
be|$VERB ($anyword|$RB)* ($anyword|VBN)
```

where $VERB matches any of the verb POS tags, $anyword matches any word, and $RB matches any of the adverbial POS tags. These voice patterns are used by a Semantic Role Labeling rule that adds either and Arg0 or and Arg1 label to verbs that have not already been given labels by the statistical SRL role labeler.

This augmentation of the labeling grammar is very simple, containing on the order of 10 rules, but works very well on the text examples in this genre. Errors are very few and are primarily omissions due to past participle verbs not correctly tagged as VBN. It is possible that this technique of copying role labels from one verb to another within the same sentence could also benefit SRL on other genres, even the original news text genre, and this is a topic of future investigation.

In addition to the two forms of Semantic Role Labeling described here, the event extraction system was integrated with a previously existing knowledge-based system for entity extraction, relation extraction and categorization of entities. This integrated system was thus able to assign semantic categories to the arguments of the events as well as to add relations between the entities. For example, one of the categories is "person" and the relations on entities of type person include "citizen-of", "residence" and "relatives". Existing rule sets were also used for dates and frequencies of events.

Mapping to the Domain Ontology

In addition to the event extraction component, the system must produce its output in the form of concepts from an ontology representing the domain. This ontology was constructed by knowledge engineers and represents a set of guidelines about the content of the reports. It was constructed in Protégé, which is a Java-based system for developing certain types of ontologies and integrating them with other systems, and is used to support a first-order logic reasoning system to draw conclusions about the individual. The ontology and reasoning system are being developed by Syracuse Research Corp.

The goal in producing this mapping was to make it as automatic as possible and to assist the human to map concepts from the text and to the domain ontology. The first and simplest part was to align the entity categories from the text to the domain ontology entitity concepts, which was a simple set of six concepts: Person, Organization, Country, GeographicRegion (other than country), Drug and Crime. In addition, some simple mappings were possible between entity relations from the text, such as Citizen-Of, to the same concept in the domain ontology.

The more challenging part of the mapping was to map to the domain ontology relations by providing patterns based on the event model. The most common patterns were to derive a relation about an individual based on the individual participating in the event by being the value of one of the semantic roles.

For example, one relation in the domain ontology is "Arrested-For", which can hold between a Person and a Crime. This relation can be derived from events such as might occur in the text with the predicate "arrested":

She was arrested in July 1998 for welfare fraud.

From the semantic role labeling, an event frame is constructed:

Event = arrested
Arg1 = she
Arg2 = welfare fraud
Date = July 1998

The system must then provide patterns for such an event to be mapped to the relation

Arrested-For (she, welfare fraud)

In this case, the name of the concept "Arrested-For" is semantically similar to the text of the event "arrested", but there are many occurrences of event verbs that should be mapped to concepts with a different name. Furthermore, there is a many-to-one mapping between the lexical text of events and the concept in the ontology.

Mappings are generalized as much as possible by using verb classes. When available in the Propbank frames, the verb class is from VerbNet (Kipper, Dang, and Palmer 2000). Otherwise, it is constructed by using WordNet (Miller 1995) to identify synonymous event verbs, with the same semantic role usage, and to map them to the same concept. Unfortunately, automatically using the VerbNet classes proved to map too many lexical items to one concept and manual intervention was required to preserve the meaning of the concept. Constraints on the arguments of the relation are generally provided by the type structure of the domain ontology.

Some relations in the ontology are unary, where they represent that a property or attribute of the individual is true. These relations are also obtained by mapping patterns. For example,

She lied on the application because she thought she would not get the job if the arrest information was listed on the application.

This sentence produces several event frames, including

Event = lied
Arg0 = she

From this event, the unary relation is produced:

Falsification (she)

Building the ontology mapping patterns is still ongoing; the mapping pattern software has been used to map to ontology concepts for four of the thirteen guideline content areas of the domain ontology, including approximately 10 relations per guideline.

Once the patterns for the mapping between event representations and the ontology are in place, the system can perform the mapping to produce instances of relations from the ontology in the output. As a final step, automatic entity coreference is used to merge lexical text instances of the arguments to relations, in particular, text instances of referring phrases such as the pronouns "he" and "she" are combined into the name of the individual as one instance of the concept Person in the ontology.

Conclusions and Ongoing Work

In the system described here, it is shown that a combination of statistical event extraction and sublanguage grammar rules can be used to extract events in a report genre that is similar to newswire text, but different in significant ways. This combination of techniques allows a system to be built that utilizes statistical techniques without new training data. Furthermore, the event representations have been useful in mapping to an entity relation ontology. While initial results look promising, evaluation of the ontology relations is still ongoing. Gold standards have been produced on a small test collection for the four initial guideline areas and the system performance will be tested on this set in the coming months.

The significance of this system is that it integrates the statistical system into a knowledge-based system and leverages the significant research in semantic role labeling for a different text genre with small amounts of training data. It also uses the event model to map to the domain ontology of the application project.

Acknowledgments. This work is partially supported by the Disruptive Technology Office (DTO).

References

Automatic Content Extraction (ACE). 2005. See www.nist.gov/speech/tests/ace.

Baker, C.F.; Fillmore, C.J.; and Lowe, J.B 1998. The Berkeley FrameNet Project. In Proceedings of the COLING-ACL, Montreal, Canada.

Basili, R.; Pannacchiotti, M.; and Zanzotto, F.M. 2005. Language Learning and Ontology Engineering: an integrated model for the Semantic Web. In Proceedings of the MEANING2005, 2nd Workshop, Trento, Italy.

Carreras, X. and Màrquez, L. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In Proceedings of CoNLL-2005.

Chang, C.-C. and Lin, C.-J. 2001. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Charniak, E. and Johnson, M. 2005. Coarse-to-fine N-best Parsing and MaxEnt Discriminative Reranking. In Proceedings of the Association for Computational Linguistics.

Gildea, D. and Jurafsky, D. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics* 28(3):245-288.

Grishman, R. and Kittredge, R. I. (Eds.). 1986. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*: Lawrence Erlbaum Associates.

Kipper, K.; Dang, H.T.; and Palmer, M. 2000. Class-based construction of a verb lexicon. In *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX.

Liddy, E.D., Jorgensen, C.L., Sibert, E.E. and Yu, E.S. 1991. Sublanguage grammar in natural language processing. In *Proceedings of RIAO '91 Conference*, Barcelona.

Maedche, A. and Staab, S. 2000. Discovering conceptual relations from text. In *Proceedings of ECAI-00*, Amsterdam. IOS Press.

Miller, G.A. 1995. WordNet: a lexical database for English. In: *Communications of the ACM* 38 (11), November, pp. 39 - 41.

Palmer, M.; Gildea D.; and Kingsbury, P. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles. *Journal of Computational Linguistics*, 31:1.

Pradhan, S.; Hacioglu, K.; Krugler, V.; Ward W.; Martin, J.; and Jurafsky, D. 2005. Support Vector Learning for Semantic Argument Classification. To appear in *Machine Learning Journal*, Special issue of Speech and Natural Language Processing.

Ozgencil, N.E. and McCracken, N. 2005. Semantic Role Labeling using libSVM. In Proceedings of CoNLL.

Riloff, E. 1996. Automatically generating extraction patterns from untagged text. In Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), Portland, Oregon.