CS 224 U: Project Milestone

Aju Thalappillil Scaria, Rishita Anubhai, Rose Marie Philip
{ajuts, rishita, rosep}@stanford.edu

Project goals

In the literature review, we focused on research publications in the areas of event and entity extraction with semantic role labeling and temporal ordering of events. Even though there has been a lot of work on one (or some) of these areas in isolation, there is no approach that seamlessly integrates all these components together. Without a system that can do all these tasks together, it is impossible to extract any meaningful information from the enormous amount of text content we have. For instance, most of the work on entity extraction and semantic role labeling assumes that the trigger word indicating an event is given. While these papers give us a good indication that these tasks can be done with reasonable performance, the task of automatic information extraction from text (specifically those describing a phenomenon or process) remains largely unsolved. Through this project, we envision to build a system that takes a paragraph of text as input and does the following:

1. Identify the events by locating the trigger words.
2. For each event, identify its arguments (only entities).
3. For each argument that is associated with a specific event, label the association with a semantic role, like Agent, Destination, Theme, Location etc.
4. Identify event-event relations that denotes the temporal ordering between events, like Cotemporal, NextEvent etc.

Previous approaches

Most of the work in the area of event and entity extraction can be analyzed by different perspectives:

- **Coverage and domain.** Most of the prevoius work dealt only with a subset of the four tasks listed as the project goals and dealt with very specific domains. For instance, Toutanova et al. deals with argument extraction and semantic role labeling assuming that the trigger words are provided. While Bjorne et al.(2009) solves the problem of event and argument extraction almost completely, their event categories and arguments are closely tied to the BIONLP task and hence, would not generalize to event extraction from domain-independent text.
- **Parsing scheme: Constitueney vs Dependecy parse.** Some of the previous work relied on the constituency parse structure of sentences, while others used the dependency parse structure. For instance Toutanova et al. approaches semantic role labeling as a joint task of argument identification and labeling on the parse tree of the sentence. Bjorne et al.(2009) and McCloskly et al.(2011) focus more on the dependency parse structure of the sentence.
- **Modeling: Graph vs Tree.** The work of Bjorne et al.(2009) and McCloskly et al.(2011) are based on graphs and deal with edge prediction, while Toutanova et al. uses tree structure with classification of nodes as an entity or not.

Current approach

In this project, we combine the learnings from the different methodologies to build a model that can be used for event and entity extraction with classes that are not specific to any domain. Even though our dataset is based on paragraphs from a biology textbook, we belives our models would generalize well to deal with more general content as our features and event/entity classes are not tied to the biological domain in anyway. The key modeling decisions are as follows:

- We model events and entities as nodes in the constituency tree. Each sentence is assumed to be independent of each other as far as entity-event relationships are concerned. Events are denoted by their trigger word and are hence pre-terminals in the parse tree. Entities are denoted by a parse tree-node that covers the whole span of text of the entity. In some cases when there is no single node that covers the entire entity (mostly because of parser errors, for e.g., PP attachment), we use some approximation by repeatedly removing tokens from the end or beginning of the span of text to identify a node that covers it. We manually verified that this heuristic works well in practice and results entities that convey almost the full meaning of original span and are well-formed.
- Since the dependency parse of a sentence has a lot of information about the dependencies between tokens, we also use features based on the dependeny parse in conjuction with the constituency parse. This is done by identifying the position of the head word of an entity or event from the dependency tree and analyzing the relations.

- We handle Task 1 as an independent task. Task 2 and 3 are done jointly. Task 4 is done as a separate task.
- For the classification tasks, we use maximum entropy model based on an implementation of L-BFGS for Quasi Newton unconstrained minimization.

Dataset

In this project, the dataset was prepared by annotating 125 paragraphs from different chapters from the text book *Biology (Eighth Edition)* by *Neil A. Campbell* and *Jane B. Reece*. Each paragraph is a text file and has an associated annotation file that indicates the different events and entities (by their character offsets in the original paragraph) and the event-entity and event-event relationships. The annotations were done by experts in the field (they were employees of a company named Vulcan).

Progress

This document talks about progress.

References

Nathanael Chambers and Dan Jurafsky 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08*.

Nathanael Chambers and Dan Jurafsky 2009. Unsupervised Learning of Narrative Schemas and their Participants In *Proceedings of ACL-09*.

Nathanael Chambers and Dan Jurafsky 2011. Template-Based Information Extraction without the Templates In *Proceedings of ACL-11*.

Jari Bjorne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, Tapio Salakoski 2009. Extracting Complex Biological Events with Rich Graph-Based Feature Sets. In *Proceedings of the Workshop on BioNLP: Shared Task,*.

Kristina Toutanova, Aria Haghighi, Christopher D. Manning 2005. Joint Learning Improves Semantic Role Labeling. In *Proceedings of ACL 2005*.

Sebastian Riedel Andrew McCallum 2008. Fast and Robust Joint Models for Biomedical Event Extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Makoto Miwa, Rune Saetre, Jin-Dong D. Kim, and Junichi Tsujii 2010c. Event extraction with complex event classification using rich features In *Journal of bioinformatics and computational biology*.

David McClosky, Mihai Surdeanu, and Christopher D. Manning 2011. Event Extraction as Dependency Parsing In *Proceedings of ACL-11*.

Nathanael Chambers and Dan Jurafsky 2008. Jointly Combining Implicit Constraints Improves Temporal Ordering In *EMNLP-08*.

Quang Xuan Do, Wei Lu, Dan Roth 2012. Joint Inference for Event Timeline Construction In *EMNLP-12*.