Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (Doctoral Consortium), Chicago, IL, July 2008

Combining Global Relevance Information with Local Contextual Clues for Event-Oriented Information Extraction

Siddharth Patwardhan
School of Computing
University of Utah
Salt Lake City, UT 84112
sidd@cs.utah.edu

Abstract

Existing IE systems tend to focus on a tight window of context surrounding the desired information to be extracted. This research addresses shortcomings of these systems by introducing a two-phase approach to IE that incorporates global relevance information with local contextual evidence, to effectively extract information from free text.

Introduction

Event-oriented Information Extraction (IE) is the task of extracting pieces of information pertaining to specific events from free text. For example, a system designed to extract information about *disease outbreaks* would locate disease outbreak events described in free text and would generate a template, such as the one shown in Figure 1, for each event detected. IE systems, therefore, process unstructured textual information and convert it to a more structured representation, viz. *event templates* or *database entries*.

Disease:	Ebola
Victims:	Five teenage boys
Location:	Mercer County, New Jersey
Date:	February 26, 2002

Figure 1: Example event template for a *disease outbreak*.

A structured representation of text generated by IE systems enables various forms of data analysis that would be impossible with free text. The numerous benefits of such systems have stimulated researchers to explore various techniques to accurately extract information from text. This research presents an approach for IE that aims to overcome some inadequacies of current systems.

Motivation & Goals

Generally speaking, current approaches to IE fall into two categories: classifier-based approaches, which use machine learning techniques to locate information, and pattern-based approaches, which use explicit sets of patterns or rules to find relevant information. One commonality behind the various approaches to IE is that they simultaneously decide whether a context describes a relevant event and whether a word or phrase in this context is a desirable extraction. Classifier-based systems rely on features that examine both the word or phrase, and its immediate context. Similarly, pattern-based systems typically use patterns or rules that match a candidate extraction and words in its close vicinity.

A limitation of simultaneously locating a relevant event and identifying the text to be extracted is that the rules or patterns used for this task need to be highly event-specific. As a result, systems typically tend to miss relevant information embedded in contexts that are not event-specific. For example, consider the domain of terrorist event reports, where one of the goals is to extract the weapons that were used in the event. Existing systems generally require patterns to recognize the context in which a weapon is explicitly linked to an event or its consequences (e.g., "*attack with* <noun phrase>", or "<noun phrase> *caused damage"*). However, weapons are not always directly linked to an event in text, but they may be inferred from the discourse. A news story could mention that a weapon was "found" nearby without explicitly stating that it was involved in a terrorist event.

Likewise, some patterns may seem to be relevant locally, but they can be deemed irrelevant when the global context is considered. For example, consider the news excerpt in Figure 2. Locally, a pattern such as "*attack against* <noun phrase>" seems likely to identify the targets or victims of a terrorist attack. But within the global context, it becomes clear that it is not related to a physical attack at all.

> IT REPORTS THAT AMBASSADOR BRUGUES HAD POINTED OUT THAT "THE U.S. PHILOSOPHERS AND THEORETICIANS WHO TRIED TO THROW SOCIALISM INTO THE GARBAGE CAN OF HISTORY HAVE LAUNCHED A NEW ATTACK AGAINST CUBA."

Figure 2: Excerpt from a news document.

The goal of this research is to combine "global" relevance information with "local" contextual clues to effectively extract relevant information from text. The strategy proposed in this research addresses the shortcomings of current IE systems by introducing a separate *relevant region identification* phase, used in combination with evidence from local features in the given context.

Proposed Work

This research introduces a decoupled approach to IE, which includes a *relevant region identification* phase followed by a *text extraction phase*. The relevant region identifier locates sentences or groups of sentences discussing relevant events. The text extraction phase then uses contextual clues to extract words or phrases from these relevant regions.

This two-phase approach allows us to be more aggressive with the extraction of information. In other words, by knowing that we are in a relevant region of text, we could pick up on less significant indicators to perform the extraction. For example, if the region classifier indicates that a particular sentence contains the description of a terrorist event, and we come across a weapon, such as *gun*, in that sentence, then we could reasonably infer that a gun was most likely used as the weapon in this event. However, the word *gun*, on its own, is an unreliable indicator of a terrorist event. Thus, in the absence of relevant regions, we would require more specific rules to reliably perform this extraction.

The use of relevant region identification, in addition, prevents the extraction of words or phrases in situations that are not event-specific in the global sense, but appear to be so within a small window of context. This typically happens because of the use of idioms or metaphors. For instance, a phrase like "*Clinton unleashed harsh attacks on Obama*", most likely refers to "verbal attacks" and not to a terrorist attack. This would become apparent to our system, which also considers the global relevance of the text as part of the process.

Based on these findings, the general plan for this research is (a) to do a thorough study of different techniques for identifying relevant regions in text, and (b) to explore various local contextual clues for performing extractions of words or phrases. Further, as currently envisioned, the two modules of the system are independent of one another. But this need not be the case. The contextual clues could be used to improve the accuracy of the relevant region identification. Similarly, information from the relevant region identifier could benefit the extraction of text using the contextual clues. This will be explored as part of the research.

There are numerous ways in which a relevant region identifier could be constructed. The research plan is to start with a straightforward sentence classifier, like the one described in our prototype system (Patwardhan and Riloff 2007). Alternatively, lexical chains or coreference chains containing some known relevant items can be used to link different segments of text and define relevant regions. Likewise, an approach like *TextTiling* (Hearst 1997) to segment text, followed by the use of a classifier to identify relevant segments is also appealing. Each of these techniques will be investigated in this research.

Since this system enables the more aggressive use of local contextual clues for extracting relevant words or phrases, various different types of contextual clues will be explored in this research. These include lexico-syntactic patterns learned from training documents or other sources (Patwardhan and Riloff 2006). Further, we observe that certain words inherently define an event-role in their meaning and can be learned using a bootstrapping approach (Phillips and Riloff 2007). Semantic properties of words are also great clues in locating extractions for certain event-roles. For instance, knowing that a word describes a weapon is a strong clue for terrorist events. Resources such as WordNet could be exploited to get this information. This research will explore techniques to automatically learn these different contextual clues for extracting event information.

Current Progress

To test the feasibility and plausibility of these ideas, we built a prototype IE system (Patwardhan and Riloff 2007), which consists of a self-trained relevant sentence classifier along with extraction patterns learned using a measure of *semantic affinity*. We found that, despite using a simple sentence classifier (the relevant region identification phase), we see better overall performance of the extraction patterns (the contextual clues) in a majority of the cases. In addition, this was achieved without sentence-level annotation of relevant sentences. The system was trained using only a set of relevant documents, a set of irrelevant documents, and a small set of seed extraction patterns.

Additionally, we have explored the use of the Web (Patwardhan and Riloff 2006) to increase the coverage of our lexico-syntactic contextual patterns. The Web-based pattern learner introduces a measure of *semantic affinity* to learn new patterns from the Web. The measure provides an estimate of the tendency of a pattern to extract words that fill a specific event-role. This measure of *semantic affinity* was used in our prototype IE system to learn contextual extraction patterns from training documents. This research will employ techniques inspired by these ideas, to aggressively gather local contextual clues for IE.

In summary, a simple prototype system was used to show that the basic research idea has promise and there is some credible evidence to continue along this research path.

Acknowledgments

This work has been supported by NSF Grant IIS-0208985, and Department of Homeland Security Grant N0014-07-1-0152.

References

Hearst, M. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics* 23(1):33–64.

Patwardhan, S., and Riloff, E. 2006. Learning Domain-Specific Information Extraction Patterns from the Web. In *Proceedings of the ACL 2006 Workshop on Information Extraction Beyond the Document*, 66–73.

Patwardhan, S., and Riloff, E. 2007. Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 717–727.

Phillips, W., and Riloff, E. 2007. Exploiting Role-Identifying Nouns and Expressions for Information Extraction. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, 165–172.