

# Extraction of Elusive Information from Text

Elizabeth D. Liddy  
Center for Natural Language Processing  
School of Information Studies  
Syracuse University  
Syracuse, New York  
United States of America  
liddy@syr.edu

## ABSTRACT

This paper reports on techniques for extending Information Extraction capabilities beyond the recognition, tagging, and extraction of entities, events, and the relations amongst them that are reported in text to capture and represent the subtler aspects of content, whether in narratives, factual reports, or opinion pieces. As statements in text often exhibit subjective colorations that can be detected, analyzed, and interpreted by NLP algorithms for presentation to users for a more accurate understanding of what might otherwise be represented as straightforward information. This paper focuses on the temporal, certainty, and affective aspects, each of which have shown promise for greater sophistication in terms of what Information Extraction systems can glean from text.

## Keywords

Information Extraction, Natural Language Processing, Subjectivity, Affect, Certainty

## 1. Introduction

Standard Information Extraction (IE) is concerned with the computational processes for identifying and extracting useful information from massive volumes of digital textual data with the goal of detecting meaning expressed either explicitly or implicitly in text. The challenge is that text is unstructured, amorphous, and layered with meaning, and nowhere near as easy to deal with as structured content, such as that found in relational databases. While difficult, IE systems have advanced to the point where they can recognize, tag, and extract for storage, the entities, events, and relations amongst them that are reported in text, as well as thematic topicality of documents. Building on these capabilities, the current focus of our work is to use Natural Language Processing (NLP) to go beyond the more obvious, explicit aspects of content and to capture and represent the subtler aspects of content, whether in narratives, factual reports, or opinion pieces.

Our belief, supported by numerous observations in ours and others' studies of text [1], [2], [3] is that statements in texts are often accompanied by subjective colorations that we can be detected, analyzed, and interpreted by sophisticated NLP algorithms for presentation to users for a more accurate understanding of what might otherwise be represented as straightforward information. Additionally, multiple reports of the same event may have quite different colorations. These elusive aspects include the temporal, certainty, affective, emotive, opinion and evaluative dimensions of meaning expressed in text. IE work on recognizing, interpreting, and extracting, and representing each of these more elusive, less concrete, aspects of information is described below.

## 2. Temporal Aspects

We have been working to apply the theoretical concepts of temporal logic in order to automatically extract temporal relations from documents. Our work is based on, and utilizes, existing temporal logic systems wherever possible, with heaviest reliance on elements from Pustejovsky's TimeML [4]. Our earlier work on extracting temporal information was focused on temporal value extraction that associates a temporal value with an event. For example, in the text segment "...in a meeting with the Pakistani ambassador on June 25, 1999...", the OCCURS relation would link the meeting and the date. Similarly, in the text segment "...arriving in Pakistan between Aug. 5 and Aug. 7", the AFTER AND BEFORE relation would temporally situate the arrival.

Our current work on representing temporal aspects is extending the event-entity relations to recognition and analysis of the more difficult event-event temporal relations - which has the added benefit of enabling the system's sequencing of events. The system assign a temporal relation between events, as in the text segment "Bykov was released from prison on bail last month after being extradited from Hungary...", the event of the prison release will be accurately sequenced AFTER the extradition event. As a further example, in the text "...Gunmen opened fire as he got out of his car outside

his home...”, the two events will be tagged as CONCURRENT events.

### 3. Certainty Aspects

Certainty is the relative degree of being free from doubt, especially on the basis of evidence. We have observed that linguistic clues of certainty do provide evidence as to the assessment of the veracity of a textual report as writers tend to qualify their expressed statements, as do the participants and observers of an event. In a preliminary study, we found an average frequency of 0.53 certainty markers per sentence in a sample of news articles and editorials [1].

In assessing and assigning certainty tags, the system utilizes various lexical expressions (e.g. *allegedly*, *supposedly*, *surely*, *we suspect*), as well as citing or referring to authority (e.g. *ITAR TASS reports*). In addition, we are currently exploring how to utilize rhetorical devices that are beneficial to a writer’s credibility (e.g. showing knowledge of subject, ability to cite authority), as well as rhetorical devices that are detrimental to a writer’s credibility (e.g. correcting prior inaccuracies / errors).

We have defined four dimensions of certainty that can be automatically identified in text, namely:

- **PERSPECTIVE** – whose certainty is involved;
- **FOCUS** – the object of the certainty;
- **TIME** – what time the certainty is expressed, and;
- **LEVEL** – the degree of certainty indicated.

### 4. Affective Aspects

The affective aspects of text involve the measurement of the polarity of text, whether it be the negative or positive attitude of a reporter or subject, a favorable or unfavorable review of a product, the right or left political leaning of a speaker, or the relative strength or weakness of views of a speaker. Today, the affective aspects are of key importance, given the increasing amounts of text available, whether in blogs, message boards, discussion groups, chat rooms, or web sites.

Detection of affective aspects is accomplished via machine learning techniques (e.g. n-grams, support vector machines, naïve Bayesian networks) based on annotated data (typically 50 to 100 documents) where 1 to 20 features of text are used (e.g. target verb, syntactic phrase type, voice, count of ‘affect’ words, association with a known set of words, etc.). The most frequently used metric for computing Semantic Orientation is PointWise Mutual Information (SO-PMI) - an algorithm that estimates the positive or negative Semantic Orientation of a given word. The metric computes the SO of each word

in a text based on its relative statistical association with positive and negative paradigm terms.

Studies use widely varying numbers of paradigm words, ranging from 1,300 for Hatzivassiloglou & McKeown, [5] to just 14 for Turney & Littman [6]. After initial training, specific domain applications may retrain with specialized positive and negative paradigm words. Then the semantic orientation of a document / posting / web site is determined by word co-occurrence counts using Information Retrieval techniques. That is, each word in a ‘text’ is put as a query to the search engine using the NEAR operator to determine the frequency with which that word co-occurs within N words of each paradigm term. So, if following the Turney & Littman approach, 28 queries would be constructed for each word in the text being analyzed. Various algorithms are then used to combine the + or - valences of all words in the text to determine its overall orientation.

### 5. Conclusion

While IE has made valuable contributions to-date in terms of its ability to recognize, categorize and extract entities, relations, and events reported in text, we are focused on pushing the boundaries by using NLP to extend IE capability to recognize the more elusive, subjective aspects of text. We believe that recent and ongoing research on temporal, certainty, and emotive aspects of text shows great promise for greater sophistication in terms of what can be gleaned from texts. The urgency to add this level of interpretability to IE systems is supported by the number of web sites, blogs, and chat-rooms that are becoming increasingly important sources to track.

### References:

- [1] V.L. Rubin, E.D. Liddy, & N. Kando, Certainty identification in texts: Categorization model and manual tagging results. J. Shanahan, J. Qu, & J. Wiebe, (Eds.), *Computing attitude and affect in text*. Springer, Dordrecht, The Netherlands, 2005.
- [2] V.L. Rubin, N. Kando, & E.D. Liddy, Certainty categorization model. *Proceedings of AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, CA, 2004.
- [3] J. Wiebe, R. Bruce, M. Bell, M. Martin, & T. Wilson, A corpus study of evaluative and speculative language. *Proceedings of 2nd ACL SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark, 2001.
- [4] J. Pustejovsky & I. Mani, Annotation of temporal and event expressions. *Proceedings of HLT-NAACL Conference*, Alberta, Canada, 2003.

[5] V. Hatzivassiloglou & K. McKeown, Predicting the semantic orientation of adjectives. *Proceedings of the 35<sup>th</sup> ACL Annual Conference*. Madrid, Spain, 1997.

[6] P.D. Turney & M.L. Littman, Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4), 2003, 315-346.