Learning Temporal Information from Text

Feng Pan
University of Southern California, USA

INTRODUCTION

As an essential dimension of our information space, *time* plays a very important role in every aspect of our lives. Temporal information is necessarily required in many applications, such as temporal constraint modeling in intelligent agents (Hritcu and Buraga, 2005), semantic web services (Pan and Hobbs, 2004), temporal content modeling and annotation for semantic video retrieval (QasemiZadeh et al., 2006), geographic information science (Agarwal, 2005), data integration of historical stock price databases (Zhu et al., 2004), ubiquitous and pervasive systems for modeling the time dimension of the context (Chen et al., 2004), and so on.

Extracting temporal information from text is especially useful for increasing the temporal awareness for different natural language applications, such as question answering, information retrieval, and summarization. For example, in summarizing a story in terms of a timeline, a system may have to extract and chronologically order events in which a particular person participated. In answering a question as to a person's current occupation, a system may have to selectively determine which of several occupations reported for that person is the most recently reported one (Mani et al., 2004).

This chapter presents recent advances in applying machine learning and data mining approaches to automatically extract explicit and implicit temporal information from natural language text. The extracted temporal information includes, for example, events, temporal expressions, temporal relations, (vague) event durations, event anchoring, and event orderings.

BACKGROUND

Representing and reasoning about temporal information has been a very active area in artificial intelligence and natural language processing. In general there are two approaches to temporal ontology for natural language. The first one is the descriptive approach (Moens and Steedman, 1988; Smith, 1991), which most concerns the descriptive properties of tense and aspect in natural language. The logical and computational approach (Allen and Ferguson, 1997; Hobbs and Pan, 2004) is the other approach that tries to formalize this quite complex ontology, for example, in first-order logic.

More recently, there has been much work on automatically extracting temporal information from text, especially on recognizing events, temporal expressions and relations (Boguraev and Ando, 2005; Mani et al., 2006; Ahn et al., 2007; Chambers et al., 2007). Pan et al. (2006, 2007) shows that implicit and vague temporal information (e.g., vague event durations) is also useful for temporal reasoning and can be learned by standard supervised machine learning techniques (e.g., Support Vector Machines (SVM), Naïve Bayes, and Decision Trees). The TimeBank corpus annotated in TimeML (Pustejovsky et al., 2003) has become a major resource for providing the annotated data for learning temporal information from text. TimeML is a rich specification language for event and temporal expressions in natural language text; unlike most previous attempts, it separates the representation of event and temporal expressions from the anchoring or ordering dependencies that may exist in a given text.

MAIN FOCUS

Machine learning approaches have become more and more popular in extracting temporal information from text. This section focuses on the most recent efforts on extracting both explicit and implicit temporal information from natural language text.

Learning Explicit Temporal Information

The TimeBank corpus currently contains only 186 documents (64,077 words of text), which is a very small size corpus for machine learning tasks. In order to address this challenge of data sparseness, especially for lacking temporal relation annotations, Mani et al. (2006) proposed to use temporal reasoning as an oversampling method. It takes known temporal relations in a document and derives new implied relations from them. There are a total of 745 derivation rules created based on Allen's interval algebra (Allen, 1984). For example, if it is known from a document that event A is before event B and event B includes event C (i.e., event C occurs during event B), then we can derive a new relation that event A is before event C. This oversampling method dramatically expands the amount of training data for learning temporal relations from text. As a result, they have achieved a predictive accuracy on recognizing temporal relations as high as 93% by using a classic Maximum Entropy classifier.

Boguraev and Ando (2005) proposed another approach to the data sparseness problem of the current TimeBank corpus. They developed a hybrid system that combines a finite-state system with a machine learning component capable of effectively using large amounts of unannotated data for classifying events and temporal relations. Specifically, temporal expressions are recognized by the finite-state system; events and temporal relations, especially relations between events and temporal expressions, are learned by the machine learning component with features extracted from the local context and the finite-state system outputs.

Modeling and Learning Implicit Temporal Information

Compared with much work of learning *explicit* temporal information, there has been little work on how to model and extract *implicit* temporal information from natural language text. For example, consider the sentence from a news article:

George W. Bush met with Vladimir Putin in Moscow.

How long did the meeting last? Our first inclination is to say we have no idea. But in fact we do have some idea. We know the meeting lasted more than ten seconds and less than one year. As we guess narrower and narrower bounds, our chances of being correct go down. As part of our commonsense knowledge, we can estimate roughly how long events of different types last and roughly how long situations of various sorts persist. For example, we know government policies typically last somewhere between one and ten years, while weather conditions fairly reliably persist between three hours and one day. There is much temporal information in text that has hitherto been largely unexploited, implicitly encoded in the descriptions of events and relying on our knowledge of the range of usual durations of types of events.

Pan et al. (2006, 2007) presented their work on how to model and automatically extract this kind of implicit and vague temporal information from text. Missing durations is one of the most common sources of incomplete information for temporal reasoning in natural language applications, because very often explicit duration information (e.g., "a five-day meeting", "I have lived here for three years") is missing in natural language texts. Thus, this work can be very important in applications in which the time course of events is to be extracted from text. For example, whether two events overlap or are in sequence often depends very much on their durations. If a war started yesterday, we can be pretty sure it is still going on today. If a hurricane started last year, we can be sure it is over by now.

They have added their new annotations of the implicit and vague temporal information to the TimeBank corpus, and proposed to use normal distributions to model the judgments (i.e., annotations) that are intervals on a scale. The inter-annotator agreement between annotations is defined as the overlapping area between normal distributions. In their corpus every event to be annotated was already annotated in the TimeBank corpus, and annotators were instructed to provide lower and upper bounds on the estimated duration of the event excluding anomalous cases, and taking the entire context of the article into account. A logarithmic scale is used for the annotated data because of the intuition that the difference between 1 second and 20 seconds is significant, while the difference between 1 year 1 second and 1 year 20 seconds is negligible. A preliminary exercise in annotation revealed about a dozen classes of systematic discrepancies among annotators' judgments. So they developed annotation guidelines to make annotators aware of these cases and to guide them in making the judgments.