Recent Developments in Temporal Information Extraction

Inderjeet Mani
Georgetown University

Abstract

The growing interest in practical NLP applications such as text summarization and question-answering places increasing demands on the processing of temporal information in natural languages. To support this, several new capabilities have emerged. These include the ability to tag events and time expressions, to temporally anchor and order events, and to build models of the temporal structure of discourse. This paper describes some of the techniques and the further challenges that arise.

1 Introduction

Natural language processing has seen many advances in recent years. Problems such as morphological analysis, part-of-speech tagging, named entity extraction, and robust parsing have been addressed in substantial ways. Hybrid systems that integrate statistical and symbolic methods have proved to be successful in particular applications. Among the many problems remaining to be addressed are those that require a deeper interpretation of meaning. Here the challenges in acquiring adequate linguistic and world knowledge are substantial.

Current domain-independent approaches to extracting semantic information from text make heavy use of annotated corpora. These approaches require that an annotation scheme be designed, debugged, and tested against human annotators provided with an annotation environment, with inter-annotator reliability being used as a yardstick for whether the annotation task and guidelines are well-defined and feasible for humans to execute. A mixed-initiative approach that combines machine and human annotation can then be used to annotate a corpus, which is in turn used to train and test statistical classifiers to carry out the annotation task.

The above corpus-driven methodology is expensive in terms of engineering cost. There are various ways of lessening the expense, including trading off quality for quantity. For example, a system can be trained from a very large sample of fully automatic annotations and a smaller sample of human-validated annotations. Nevertheless, the total cost of putting together an

annotation scheme and applying it to produce a high-quality annotated corpus is still high.

Temporal information extraction offers an interesting case study. Temporal information extraction is valuable in question-answering (e.g., answering 'when' questions by temporally anchoring events), information extraction (e.g., normalizing information for database entry), summarization (temporally ordering information), etc. Here, as we shall see, a system has to strive for a relatively deep representation of meaning. However, the methodology outlined above breaks down to some extent when applied to this problem. This in turn suggests new approaches to annotation by humans and machines.

2 Temporal information extraction

To illustrate the problem of Temporal Information Extraction, consider the following discourse:

(1) Yesterday, John fell. He broke his leg.

A natural language system should be able to **anchor** the falling event to a particular time (yesterday), as well as **order** the events with respect to each other (the falling was *before* the breaking). We can see here that a system needs to be able to interpret events (or more generally, events and states, together called eventualities), tense information, and time expressions. The latter will be lumped under temporal adverbials, including temporal prepositions, conjunctions, etc. Further, in order to link events to times, commonsense knowledge is necessary. In particular, we infer that the breaking occurred the same day as the falling, as a result of it, and as soon as the fall occurred. However, this is a default inference; additional background knowledge or discourse information might lead to an alternative conclusion.

Consider a second example discourse (2):

(2) Yesterday Holly was running a marathon when she twisted her ankle. David had pushed her.

Here we need to understand that the use of the progressive form (i.e., aspectual information) indicates that the twisting occurred *during* the 'state' of running the marathon. Knowledge of tense (past perfect) suggests that the pushing occurs *before* the twisting (at least). Commonsense knowledge also suggests that the pushing occurs *before* and caused the twisting. We can see that even for interpreting such relatively simple discourses, a system might

require a variety of sources of linguistic knowledge, including knowledge of tense, aspect, temporal adverbials, discourse relations, as well as background knowledge. Of course, other inferences are clearly possible, e.g., that the running stopped *after* the twisting, but when viewed as defaults, these latter inferences seem to be more easily violated.

Consider now the problem of representing the structure of the extracted information. It is natural to think of this in terms of a graph. For example, a graph for (1) is shown in Figure 1; here we assume the document publication date is 18 February 2004:

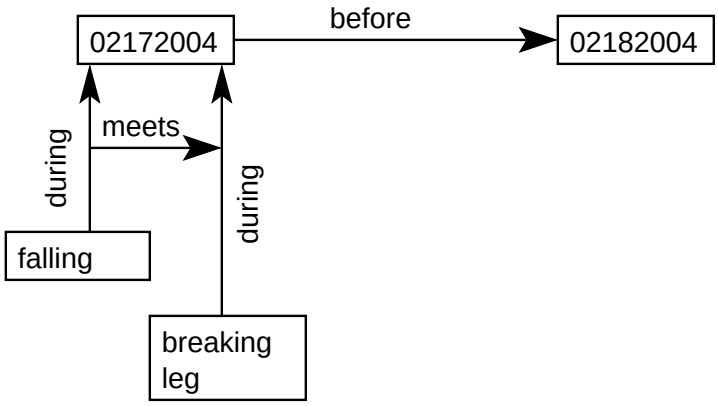

Fig. 1: *Graph for a simple story*

Here we have assumed that the falling culminates in the breaking of the leg, i.e., that there is no time gap in between.

Turning to the structure of such graphs, it should be clear that the events will not necessarily be totally ordered in time, so we should consider the events and times in the graph to be partially ordered. Let us assume that events and times are represented as intervals marked by pairs of time points, and let us adopt the thirteen relations that Allen (1984) proposes in his interval-based temporal logic. Then, we can consider how to map NL texts to such graphs by an automatic procedure, and then use the graphs to answer questions, produce summaries, timelines, etc. The focus in this paper is on the mapping, rather than the use.

3 Previous research

Until recently, most of the prior research on temporal information extraction had drawn inspiration from work in linguistics and philosophy, as well

as research on temporal reasoning in artificial intelligence. The early work of Moens & Steedman (1988) and Passonneau (1988) focused on linguistic models of event structure and tense analysis to arrive at temporal representations. For example, in Moens & Steedman (1988), "Harry hiccupped for three hours" would be analyzed as a process of iteration of the point event of hiccupping. Passonneau (1988) developed an information extraction system that could temporally locate events in texts, processing sentences like "The compressor failed before the pump seized". Much of the early work also adopted Allen's temporal relations, and used meaning representations augmented with temporal variables (Reichenbach 1947) or temporal operators (Prior 1968).

Earlier work also devoted a lot of attention to temporal aspects of discourse. A default assumption that runs through the literature (see, especially Dowty (1986)) is that a simple past tense sentence, if it describes an event, advances the narrative, so that the event occurs after the eventuality in the previous sentence. This is the *narrative convention* of narrating events in the order they occur. If the eventuality is a state, a default assumption is that it *overlaps* with the eventuality of the previous sentence. Work by Webber (1988) related the ordering principles to a general model of discourse processing where tense was treated anaphorically, specifying a number of rules governing the temporal relationships among successive clause pairs in a discourse. Later work by Song & Cohen (1991) extended Webber's work in an implemented system that hypothesized that only certain kinds of tense shifts were coherent. They went on to suggest various heuristics to resolve ambiguities in temporal ordering. Hwang & Schubert (1992) implemented a system based on a framework of compositional semantics, showing why compositionality was a crucial property in temporal interpretation, especially for handling subordinated events.

In parallel, developments in formal semantics led to the evolution of Discourse Representation Theory (Kamp & Reyle 1993). Here the semantic representation of a sentence in a discourse context includes temporal ordering and inclusion relations over temporal indices. However, the focus was on the default narrative convention above, along with the *states overlap* assumption. Clearly, discourse relations like causality, as in (2), violate this convention. This point was taken up by work by Lascarides & Asher (1993), who developed a theory of defeasible inference that relied on a vast amount of world knowledge. Hitzeman et al. (1995) argued convincingly that reasoning in this way using background knowledge was too computationally expensive. Instead, their computational approach was based on assigning

weights to different ordering possibilities based on the knowledge sources involved, with semantic distance between utterances, computed based on lexical relationships, standing in for world knowledge.

The widespread use of large corpora in NLP allowed work on temporal information extraction to advance forward quite dramatically. Wiebe et al. (1998) used a corpus-based methodology to resolve time expressions in a corpus of Spanish meeting scheduling dialogs at an overall accuracy of over 80%. Other work on resolving time expressions in meeting scheduling dialogs include Alexandersson et al. (1997) and Busemann et al. (1997). In the meantime, community-wide information extraction tasks had started to show beneficial results. The MUC-7 (1998) task tested accuracy in flagging time expressions, but did not require resolving their values. In flagging time expressions, however, at least 30% of the dates and times in the MUC test were fixed-format ones occurring in document headers, trailers, and copyright notices, thus simplifying the task.

Another area of work in temporal information extraction involves processing temporal questions. Androutsopoulos (2002) allowed users to pose temporal questions in natural language to an airport database, where English queries were mapped to a temporal extension of the SQL database language, via an intermediate semantic representation that combined both temporal operators and temporal indices. For example, the question "Which flight taxied to gate 4 at 5:00 pm?" would result in an interpretation where the taxiing started or ended at 5 pm. Although this effort was focused on databases, the emphasis on mapping a representation of NL meaning to a formal language that can support inference is inherent in approaches to temporal information extraction.

4 TimeML

The body of previous work suggests the need for an annotation scheme that can capture the kind of graph structure shown in Figure 1. TimeML (Pustejovsky et al. 2004) is a proposed metadata standard for markup of events and their temporal anchoring in documents that addresses this. It has been applied mainly to English news articles. The annotation scheme integrates together two annotation schemes: TIDES TIMEX2 (Ferro et al. 2000) and Sheffield STAG (Setzer & Gaizauskas 2000), as well as other emerging work (Katz & Arosio 2000). It identifies a variety of event expressions, including tensed verbs, e.g., "has left", "was captured","will resign"; stative adjectives "sunken", "stalled", "on board"; and event nominals "merger", "Military

Operation", "Gulf War".

Eventualities in TimeML have various attributes, including the type of event, its tense, aspect, and other features. Temporal adverbials include *signals*, i.e., temporal prepositions ("for", "during", "on", "at", etc.) and connectives ("before", "after", "while", etc.). TimeML also represents time expressions, adding various modifications to TIMEX2, yielding an annotation scheme called TIMEX3. The main point of TimeML, however, is to link eventualities and times; for example, anchoring an event to a time, and ordering events and/or times. This is done by means of TLINK, or temporal links labeled with Allen-style temporal relations. Linking also take into account actual versus hypothetical events, e.g., (3), where the leaving is subordinated to a modal "may", and (4), where the leaving is subordinated to the saying/denying. These latter situations are addressed by means of SLINKs, or *subordinating* links. Thus, in (5) below, the saying subordinates the other events, which are in turn subordinated in the order found in the sentence.

(3) John may leave tomorrow.

(4) John said/denied that Mary left.

(5) The message to the chief of staff was meant to be taken as a suggestion that Sununi offer to resign, one highly placed source said.

Finally, TimeML also annotates *aspectual* verbs like "start (to cough)", "continue lazing about", etc. These verbs, rather than characterizing a distinct event, indicate a particular phase of another event; as a result, the aspectual verb is linked by a *aspectual* link (ALINK) to the event.

Recent work by Hobbs & Pustejovsky (2004) maps the structure of TimeML to a formal theory of time (the DAML small Time Ontology), which in turn allows formal queries to be posed to a reasoning system.

5 TIMEX2

TIMEX2 is the historically oldest segment of what is now TimeML. Although the guidelines are fairly complex, it is the relatively most robust part of the TimeML scheme. As a result, it has been applied more extensively than TIMEX3 or the rest of TimeML. It was developed originally by the DARPA TIDES program and has since been adopted by the U.S. Government in the Automatic Content Extraction (ACE) program's Relation Detection and Characterization (RDC) task, and in two ARDA TimeML summer workshops (NRRC 2004).

TIMEX2 is an annotation scheme for marking the extent of English time expressions (with TIMEX2 tags) and normalizing their values in ISO-8601 (1997) format (with a few extensions). The TIMEX2 scheme represents the meaning of time expressions expressed as time points, e.g., "yesterday" with the value 20040217, or "the third week of October":$2000W42$. It also represents durations, e.g., "half an hour long":$PT30M$. TIMEX2 also handles fuzzy times such as "Summer of 1990":$1990SU$, where a primitive SU is used. It also distinguishes between specific and non-specific uses (the latter being a catchall for indefinite, habitual, and other cases) e.g., "*April* is usually wet":$XXXX04; non\_specific$. Sets of times are represented to some extent, e.g., "every Tuesday" has a value $XXXXWXX2$ with *periodicity F1W* and *granularity G1D*, where *F1W* means once a week, and *G1D* means a grain size of one day.

Annotators can be trained for TIMEX2 tagging very quickly (usually half a day of training followed by a few homework exercises). Inter-annotator accuracy, on the average, across 5 annotators annotating 193 news documents from the (TDT2 1999) corpus, is .86 F-measure in identifying time values. The F-measure for identifying tag extent (where tags start and end) is .79. The reason the value F-measures are higher than the extent F-measures is because the scorer flags occurrences of tags in a candidate annotation that occur in almost but not exactly the same position in the reference annotation as errors of extent, but nevertheless compares the values of such overlapping tags, scoring the values correct if the candidate and reference values are equivalent.

However, inter-annotator reliability on two features is low: F-measure on *granularity* is .51, and on *non-specificity* it is .25. While there were only a small sample of these latter features in the corpus (200 examples compared to 6000 examples of time values), these do indicate a problem, leading to a number of modifications, including the revised specification for sets in TIMEX3 (see below). Error analyses confirm that annotators do deviate from the guidelines and produce systematic errors, for example, annotating "several years ago" as PXY (a period of unspecified years, a valid time expression) instead of $PAST\_REF$; or annotating "all day" as *P1D* rather than $YYYYMMDD$.

6 TIMEX2 tagging

A variety of approaches have been developed to tag TIMEX2 expressions. I discuss one method here; others are briefly summarized later. The TIMEX2

tagger TempEx (Mani & Wilson 2000) handles both absolute times (e.g., "June 2, 2003") and relative times (e.g., "Thursday") by means of a number of tests on the local context. Lexical triggers like "today", "yesterday", and "tomorrow", when used in a specific sense, as well as words which indicate a positional offset, like "*next* month", "*last* year", "this *coming* Thursday" are resolved based on computing direction and magnitude with respect to a reference time, which is usually the document publication time. Bare day or month names ("Thursday", or "February") are resolved based on the tense of neighboring past or future tense verbs, if any. Signals such as "since" and "until" are used as well, along with information from nearby dates.

TempEx has been applied to different varieties of corpora, including broadcast news, print news, and meeting scheduling dialogs. The performance on all of these is comparable. On the 193-document TDT2 subcorpus, it obtained a .82 F-measure in identifying time values and .76 F-measure for extent.

In conjunction with work on tagging TIMEX2, word-sense disambiguation has also been carried out. For example, deciding whether an occurrence of "today" is non-specific or not can be carried out by a statistical classifier at .67 F-measure (using a Naïve Bayes classifier), which is significantly better than guessing the majority class (.58 F-measure for specific). Other types of sense temporal disambiguation have also been carried out. For example, deciding whether word tokens like "spring", "fall", etc. are used in a seasonal sense can be carried out at .91 F-measure (using decision trees), whereas just guessing seasonal all the time scores .54 F-measure.

7 TIMEX3 extensions

As mentioned earlier, the SET specification in TIMEX2 proved to be problematic for annotators. In TIMEX3, SET has been simplified to have two attributes in addition to the value: *quant* quantification over the set, and *freq* frequency within the set. Thus, we have examples like "three days every month": *P1M;quant=every;freq=P3D* and "twice a month": *P1M; freq=P2X*.

TIMEX3 also allows event-dependent time expressions like"three years after the Gulf War" to be tagged, since, unlike TIMEX2, events are tagged in TimeML. TIMEX3 in addition allows a functional style of encoding of offsets in time expressions, so that "last week" could be represented not only by the time value but also by an expression that could be evaluated to compute the value, namely, that it is the predecessor week of the week

preceding the document date.

However, at the time of writing, automatic tagging of TIMEX3 has not yet been attempted, nor has inter-annotator reliability on TIMEX3 been studied, so we cannot as yet assess the feasibility of these extensions.

8 Challenges in TimeML link annotation

The annotation by humans of links in TimeML is a very challenging problem. Ordering judgments, as indicated by discourses (1) and (2) above, can be hard for a variety of reasons:

- The annotation of events other than tensed verbs. Since states are included, deciding which states to annotate can also be difficult, since the text may not state when a state stopped holding (this is an aspect of the AI *frame problem*). For example, given (6), we infer that Esmeralda was no longer hungry after the eating event, and that as far as we know nothing else changed. The guidelines call for just annotating those states which the text explicitly indicates as having changed, but specifying this is difficult.

  (6) Esmeralda was hungry. She ate three Big Macs.

- The difficulty of deciding whether a particular relation is warranted. For example, in (2) above, we recommended against committing to the twisting as *finishing* the marathon running. Determining what inference to commit to can be fairly subtle.
- The possibility of ambiguity or lack of clear indication of the relation. In such a case, the user is asked not to annotate the TLINK.
- The granularity of the temporal relations. A pilot experiment (Mani & Schiffman 2004) with 8 subjects providing event-ordering judgments on 280 clause pairs revealed that people have difficulty distinguishing whether there are gaps between events. The 8 subjects were asked to distinguish whether an event is (a) *strictly before* the other, (b) *before and extending into* the other, or (c) is *simultaneous* with it. These distinctions can be hard to make, as in the example of ordering "try on" with respect to "realize" in (7):

  (7) In an interview with Barbara Walters to be shown on ABCs "Friday nights", Shapiro said he tried on the gloves and realized they would never fit Simpson's larger hands.

  Not surprisingly, subjects had only about 72% agreement (corresponding to a low Kappa score of 0.5) on these ordering distinctions. Ignoring the (a) versus (b) distinction raises the agreement to Kappa

0.61, which is (arguably) acceptable. This experiment shows that a coarse-grained concept of event ordering is more intuitive for humans.

- The density of the links. The number of possible links is quadratic in the number of events. Users can get fatigued very quickly, and may ignore lots of links.

To date, no inter-annotator study has been carried out on linking. However, analyses of a preliminary version of the Timebank Corpus, a collection of news documents annotated with TimeML at Brandeis University (NRRC 2004), reveal a number of interesting aspects of annotator behavior. In this corpus there were 186 documents, with 8324 eventualities and 3404 TLINKs, about 45 eventualities per document but only 18 TLINKs per document. This means that less than half the eventualities are being linked. Further, the vast majority (69%) of the TLINKs are within-sentence links. Sentences in news texts are generally long and complex, and many of these links involve an eventuality in a subordinate clause being linked to another in some other clause. Similarly, links between subordinate clauses of one sentence and a main clause of another are also found.

Overall, we expect that inter-annotator consistency is a hard-to-reach ideal as far as TLINKs are concerned. However, the following steps can improve consistency within and across annotators:

1. Adding more annotation conventions. For example, it might be helpful to have annotation conventions for dealing with links out of subordinate clauses. Clearly, TimeML needs a certain level of training, more than would be required for TIMEX2, so adding specific conventions can make for tighter and more consistent annotation.
2. Constraining the scope of annotation. The goal here is to restrict the number of decisions the human has to make. This could involve restricting the types of events and states to be annotated, as well as the conditions under which links should be annotated. Thus, efforts on a 'TimeML Lite' are important.
3. Expanding the annotation using temporal reasoning. Since temporal ordering and inclusion operators like *before* and *during* are transitive and symmetric, it is possible to expand two different annotations by closure over transitive and symmetric relations, thereby increasing the possibility of overlap. This also boosts the amount of training data for link detection.
4. Using a heavily mixed-initiative approach. Here automatic tagging and human validation go hand-in-hand, so that the annotator always starts from a pre-existing annotation that steadily improves.

5. Providing the user with visualization tools during annotation. This can help them produce more densely connected graphs. This is borne out by results with a graphical visualization tool called TANGO (NRRC 2004) that we have helped develop for annotation. This in turn has led to more complete annotations using temporal reasoning as above.

Figure 2 shows a sample TANGO screen.

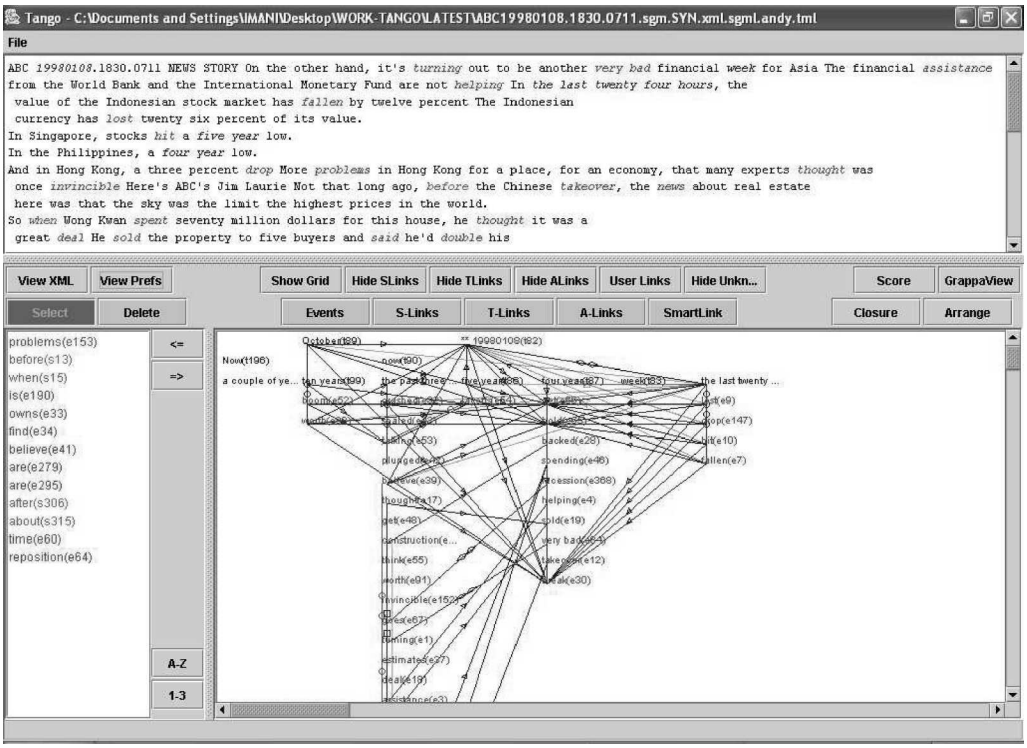

Fig. 2: *TANGO: A graphical tool for annotating links*

The right-hand window shows a graphical annotation pallette, onto which events and times from the pending window on the left can be moved. The top of the pallete automatically sorts the times. The user can link events and other events or times by drawing links, with pop-up menus being used to specify non-default attributes. The system can auto-arrange the display, or rely on the user arrangement. The Closure button applies temporal reasoning rules to expand the annotation with additional links; such an expanded annotation is shown in the figure. At any point, the annotation can be dumped in XML or scored against a reference annotation.

9 Empirical constraints on temporal discourse

The availability of empirical data from experiments and corpora allow one to test to a certain extent the theories of temporal ordering discussed earlier. The tests to date have mainly been on news. As Bell (1999) has pointed out, the temporal structure of news is dictated by perceived news value rather than chronology. Thus, the latest news is often presented first, instead of events being described in narrative order. So, one would not expect the narrative convention to be strong.

This is borne out in the experiment of Mani & Schiffman (2004) cited above, where it was found that the narrative convention applied only 47% of the time in ordering events in 131 pairs of successive past-tense clauses. Interestingly, 75% of clauses lack explicit time expressions, i.e., the 'anchor time of most events is left implicit, so that simply anchoring events to times explicitly associated with them in the text will lead to extremely poor TLINK recall. Clearly, therefore, document- and narrative-based inference could be crucial in automatic tagging.

In support of the 'states overlap principle, the TimeBank data shows that the overall percentage of links involving an *overlap* relation is 9% on the average, but 21.8% when one or both eventualities are states, a significant increase.

10 Automatic TLINK tagging

Mani et al. (2003) address the problem of implicit times by using document-level inference. Their algorithm computed a reference time (Reichenbach 1947, Kamp & Reyle 1993:594) for the event in each finite clause, defined to be either the time value of an explicit temporal expression mentioned in the clause, or, when the explicit time expression is absent, an implicit time value inferred from context, using a naive algorithm which is only 59% correct. A set of 2069 clauses from the North American News Corpus was annotated with event-time TLINK information by a human (after correcting the reference times produced by the above propagation algorithm), and then turned into feature vectors and used as training data for various machine learning algorithms. A decision rule classifier (C5.0 Rules) achieved significantly higher accuracy (.84 F-measure) compared to other algorithms as well as the majority class, where the event is simultaneous with the temporal anchor (most news events occur at the time of the explicit or implicit temporal anchor). Next, the anchoring relations and sorting of the times

were used to construct a (partial) event ordering, which was evaluated by a human for document-level event orderings. The machine achieved a .75 F-measure in event ordering of TLINKs.

In comparison, Mani & Wilson (2000) used a baseline method of blindly propagating TIMEX2 time values to events based on proximity. On a small sample of 8,505 words of text, they obtained 394 correct event times in a sample of 663 verb occurrences, giving an accuracy of 59.4%. Filatova & Hovy (2000) obtained 82% accuracy on 'timestamping' clauses for a single type of event/topic on a data set of 172 clauses. However, fundamental differences between the three evaluation methods preclude a formal comparison.

While the approach of Mani et al. (2003) is corpus-based, it suffers from several serious disadvantages, including lack of training data, very few predictive features, and rules which cover just a small number of examples. In addition, it lacks an adequate representation of discourse context in the feature vector, except for features that track shifts in tense and aspect. In future, to address this problem successfully, one would need to carry out more annotation, improve machine learning approaches, and try out a variety of other features motivated by corpus analysis.

11 Multilinguality

While the TimeML scheme in itself has been confined to English, there have been several efforts aimed at temporal information extraction for other languages. In terms of link extraction, Schilder & Habel (2000) report on a system which takes German texts and infers temporal relations from them, achieving 84% accuracy, and Li et al. (2000) take Chinese texts, and using a number of somewhat complex rules, achieves 93% accuracy on extracting temporal relations. However, these approaches are few and far between, and are hard to compare.

The problem of time expression tagging, being simpler than link extraction, has also been carried out on a number of languages. Research on time expression resolution for meeting scheduling dialogs has addressed German (Alexandersson et al. 1997, Busemann et al. 1997) as well as Spanish (Wiebe et al. 1998). The latter Spanish dialogs (from the Enthusiast Corpus of Rose et al. (1995), collected at CMU) have been translated into English and annotated with TIMEX2 tags by a bilingual annotator, based on tagging the English portion and adapting it to the Spanish. There has also been some initial work on a Hindi tagger for the TIDES Surprise Language experiment

(TIMEX2 2004).

At Georgetown, we have also completed work on TIMEX2 tagging of Korean. We have annotated a corpus of 200 Korean news articles (from Hankook and Chosun newspapers) with TIMEX2. The main difference, in comparison to English TIMEX2, is in terms of morphology. Korean has agglutinative morphology, and this has implications for some of the rules for tag extent. For example, English temporal annotation guidelines state that temporal prepositions like "from" (as in "from 6 p.m.") are not part of the extent. Since Korean instead uses postpositions that are bound morphemes, we allow sub-word TIMEX2 tags that exclude the postposition. Likewise, the English guidelines require vague conjoined expressions like "three or four days", to be annotated with two tags, whereas "three or four" is a single word in Korean. Apart from this, however, the annotation scheme carries over very well. Inter-annotator reliability of 2 annotators on 30 documents shows .89 F-measure for values and extent.

Several automatic taggers have been developed at Georgetown. The first, KTX (TIMEX2 2004), is a memory-based tagger that uses a dictionary of temporal expressions and their values derived automatically from a training corpus. Relative times in the test data are resolved using hand-created heuristics based on offset length in the training data. KTX achieves a F-measure of .66 on tagging extents and .86 F-measure for values on 200 documents. While KTX has Korean-specific morphological knowledge, it doesn't perform any prediction, being confined to just memorizing instances seen before. Another tagger, TDL (Baldwin 2001), has been developed that performs a degree of generalization. In this approach, a time expression and its TIMEX2 tag information form a training example for learning the mappings between strings and the values of temporal attribute variables. For example, a collection of similar date examples like "February 17, 2001":20010217 will generate a rule of the form $Pattern(?M, ?D, comma, ?Y) \rightarrow Value(Year(?Y), Month(?M), Day(?D))$, with a confidence based on the frequency of the pattern. TDL however lacks specific knowledge of Korean (or any other language, though it makes assumptions about the maximum word length of a time expression). TDL achieves .56 F-measure for extent and .83 F-measure for time values on 71 English documents.

Our experience with multilingual annotation suggests that the TIMEX2 scheme ports well to a variety of languages, and that a corpus-based approach with at least some language-specificity to handle morphology is, so far, the most cost-effective.

12 Conclusions

Overall, temporal information extraction offers many opportunities to tie together natural language processing and inference based on formal reasoning. The work reported here has made considerable progress due in part to the twin emphases of a corpus-based approach and evaluation. The strategy has been to develop semantic representations that can be used for formal inference in support of various practical tasks. These representations are motivated to some extent by work in formal semantics and symbolic AI. Once the representations are formally specified, the goal is then to automatically construct such representations using corpus-based methods. A similar strategy can be taken to advance the field of spatial information extraction.

However, it should be borne in mind that annotating data with relatively more complex representations is expensive and difficult to carry out. As a result, the emphasis shifts towards tools to help the human efficiently produce annotated corpora. Some of these corpora and tools are available at NRRC (2004) and TIMEX2 (2004).

At Georgetown, we are continuing to push ahead with temporal information extraction (including TLINK extraction) for different languages, including Chinese. We have also developed a new approach to modeling discourse structure from a temporal point of view, on which annotation will begin in due course. Finally, we have started to apply this work to both summarization and question-answering.

Acknowledgements. We would like to thank Georgetown's Seok Bae Jang, Jennifer Baldwin and Alan Rubinstein for their work on KTX, TDL and analyses of the Timebank corpus, respectively.

REFERENCES

Alexandersson, Jan, Norbert Riethinger & Elisabeth Maier. 1997. "Insights into the Dialogue Processing of VERBMOBIL". *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 33-40. Washington, D.C.

Allen, James F. 1984. "Towards a General Theory of Action and Time". *Artificial Intelligence* 23:2.123-154.

Androutsopoulos, Ion. 2002. *Exploring Time, Tense and Aspect in Natural Language Database Interfaces*. Amsterdam & Philadelphia: John Benjamins.

Baldwin, Jennifer. 2001. "Learning Temporal Annotation of French News". Master's Research Thesis. Dept. of Linguistics, Georgetown University.

Bell, Alan. 1999. "News Stories as Narratives". *The Discourse Reader* ed. by A. Jaworski & N. Coupland, 236-251. London & New York: Routledge.

Busemann, Stephan, Thierry Declerck, Abdel Kader Diagne, Luca Dini, Judith Klein & Sven Schmeier. 1997. "Natural Language Dialogue Service for Appointment Scheduling Agents". *Proceedings of the Fifth Conference on Applied Natural Language Processing* (*ANLP-97*), 25-32. Washington, D.C.

Dowty, David R. 1986. "The Effects of Aspectual Class on the Temporal Structure of Discourse: Semantics or Pragmatics?". *Linguistics and Philosophy* 9.37-61.

Ferro, Lisa, Inderjeet Mani, Beth Sundheim & George Wilson. 2001. "TIDES Temporal Annotation Guidelines Draft - Version 1.02". MITRE Technical Report MTR MTR 01W000004. McLean, Virginia: The MITRE Corporation.

Filatova, Elena & Ed Hovy. 2001. "Assigning Time-Stamps to Event-Clauses". *Workshop on Temporal and Spatial Information Processing at the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL'2001*), 88-95. Toulouse, France.

ISO-8601. 1997. — `ftp://ftp.qsl.net/pub/g1smd/8601v03.pdf` [Source checked in March 2004]

Hitzeman, Janet, Marc Moens & Clare Grover. 1995. "Algorithms for Analyzing the Temporal Structure of Discourse". *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics* (*EACL'95*), 253-260. Dublin, Ireland.

Hobbs, Jerry R. & James Pustejovsky. Forthcoming. "Annotating and Reasoning about Time and Events". To appear in *The Language of Time: Readings in Temporal Information Processing* ed. by Inderjeet Mani, James Pustejovsky, & Robert Gaizauskas. Oxford: Oxford University Press.

Hwang, Chung Hee & Lenhart K. Schubert. 1992. "Tense Trees as the Fine Structure of Discourse". *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics* (*ACL'92*), 232-240. Newark, Delaware.

Kamp, Hans & Uwe Reyle. 1993. *From Discourse to Logic (Part 2)*. Dordrecht: Kluwer Academic.

Katz, Graham & Fabrizio Arosio. 2001. "The Annotation of Temporal Information in Natural Language Sentences". *Workshop on Temporal and Spatial Information Processing at the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL'2001*), 104-111. Toulouse, France.

Lascarides, Alex & Nicholas Asher. 1993. "Temporal Relations, Discourse Structure, and Commonsense Entailment". *Linguistics and Philosophy* 16.437-494.

Li, Wenjie, Kam-Fai Wong & Chunfa Yuan. 2001. "A Model for Processing Temporal References in Chinese". *Workshop on Temporal and Spatial Information Processing at the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL'2001*), 33-40. Toulouse, France.

Mani, Inderjeet & George Wilson. 2000. "Robust Temporal Processing of News". *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (*ACL'2000*), 69-76. Hong Kong.

Mani, Inderjeet, Barry Schiffman & Jianping Zhang. 2003. "Inferring Temporal Ordering of Events in News". *Proceedings of the Human Language Technology Conference* (*HLT-NAACL'03*), 55-57. Edmonton, Canada.

Mani, Inderjeet & Barry Schiffman. Forthcoming. "Temporally Anchoring and Ordering Events in News". To appear in *Time and Event Recognition in Natural Language* ed. by James Pustejovsky & Robert Gaizauskas. Amsterdam & Philadelphia: John Benjamins.

Moens, Marc & Mark Steedman. 1988. "Temporal Ontology and Temporal Reference". *Computational Linguistics* 14:2.15-28.

MUC-7. 1998. *Proceedings of the Seventh Message Understanding Conference.* Washington, D.C.: DARPA.

NRRC. 2004. — `http://nrrc.mitre.org/` [Source checked in March 2004]

Passonneau, Rebecca J. 1988. "A Computational Model of the Semantics of Tense and Aspect". *Computational Linguistics* 14:2.44-60.

Prior, Arthur N. 1968. "Tense Logic and the Logic of Earlier and Later". *Papers on Time and Tense* ed. by A. N. Prior, 116-134. Oxford: Oxford University Press.

Pustejovsky, James, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Robert Gaizauskas, Andrea Setzer, Graham Katz & Inderjeet Mani. Forthcoming. "The Specification Language TimeML". To appear in *The Language of Time: Readings in Temporal Information Processing* ed. by Inderjeet Mani, James Pustejovsky, & Robert Gaizauskas. Oxford: Oxford University Press.

Reichenbach, Hans. 1947. *Elements of Symbolic Logic.* London: Macmillan.

Rose, C.P., Barbara Di Eugenio, L. Levin, & C. Van Ess-Dykema. 1995. "Discourse Processing of Dialogues with Multiple Threads". *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (*ACL'95*), 31-18. Cambridge, Mass., U.S.A.

Schilder, Frank & C. Habel. 2001. "From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages". *Workshop on Temporal and Spatial Information Processing at the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL'2001*), 65-72. Toulouse, France.

Setzer, Andrea & Robert Gaizauskas, 2000. "Annotating Events and Temporal Information in Newswire Texts". *Proceedings of the Second International Conference On Language Resources And Evaluation* (*LREC-2000*), 1287-1294. Athens, Greece.

Song, Fei & Robin Cohen. 1991. "Tense Interpretation in the Context of Narrative". *Proceedings of the Ninth National Conference on Artificial Intelligence* (*AAAI'91*), 131-136. Anaheim, Calif.

TDT2. 1999. — `http://morph.ldc.upenn.edu/Catalog/LDC99T37.html` [Source checked in March 2004]

TIMEX2. 2004. — `http://timex2.mitre.org/` [Source checked in March 2004]

Webber, Bonnie. 1988. "Tense as Discourse Anaphor". *Computational Linguistics* 14:2.61-73.

Wiebe, Janyce M., Thomas P. O'Hara, Thorsten Ohrstrom-Sandgren & Kenneth J. McKeever. 1998. "An Empirical Approach to Temporal Reference Resolution". *Journal of Artificial Intelligence Research* 9.247-293.