Identification of Temporal Event Relationships in Biographical Accounts

Lucian Silcox
University of Texas PanAm
1201 W. University Dr.
Edinburg, Tx 78539
lucian.silcox@gmail.com

Emmett Tomai
University of Texas PanAm
1201 W. University Dr.
Edinburg, Tx 78539
tomaie@utpa.edu

Abstract

This paper examines the efficacy of the application of a pre-existing technique in the area of event-event temporal relationship identification. We attempt to both reproduce the results of said technique, as well as extend the previous work with application to a newly-created domain of biographical data. We find that initially the simpler feature sets perform as expected, but that the final improvement to the feature set underperforms. In response, we provide an analysis of the individual features and identify differences existing between two corpora.

1 Introduction

As natural language systems continue to grow, so too does the importance of extracting temporal information from text. Narratives often contain a wealth of temporal information, linking specific events to each other and to individual named entities of importance, but such information is often implicitly conveyed, rather than explicitly stated. The continued interest in Question Answering and other data extraction systems has emphasized the need to better understand these relations to move past superficial understanding to a level of deeper comprehension. For native speakers, the temporal clues hidden in the text are relatively simple to comprehend. However, even for human annotators, the task of identifying and classifying the specific relationship between two events can be problematic. This complexity, of course, only exacerbates the problem of trying to automate the process for any information extraction system.

The creation of the TimeBank Corpus (Pustejovsky et al, 2003a), a fully-annotated newswire domain, opened up the possibility of applying machine learning techniques to the task of automatically extracting temporal relations. We look to the standards of the TimeBank Corpus to create a corpus of biographical accounts, and apply techniques that have been shown to work on TimeBank to the new domain.

2 Related Work

Domain-independent approaches have often focused on events that can be bound to a global timeline (Mani et al, 2003). This includes dates and times, but often neglects phrases that indicate events occurring in relative time (e.g. "during school," "before the crash," or "recently"). Research conducted on news articles attempted to identify the specific temporal relationships between two events, as seen in (Mani et al, 2006). Further work in that domain extended this start by identifying additional features that better predicted those temporal relations. (Lapata & Lascarides 2007; Chambers et al, 2007).

In this work, we are primarily interested in applying event ordering techniques to documents less structured than news articles, specifically biographies. It is the intention of our work to validate the efficacy of previous techniques in a different domain, and thus we attempt to extend the work completed by Chambers et al through application to a newly created corpus of biographical data. In the previous work, Chambers reports best results of 59.43% accuracy with gold standard features on TimeBank. We attempt to reproduce these results, and also adopt the policy of including incremental results against features selected based on the work

Proceedings of NAACL-HLT 2013, pages 529–533,
Atlanta, Georgia, 9–14 June 2013.

of Mani et al (2006), and Lapata & Lascarides (2007).

3 Data

For purposes of validation of our implementation, we adopt the use of the TimeBank corpus (v1.1), which consists of 186 newswire documents and 3,406 identified event pairs with temporal relationships. The number of identified event pairs differs slightly from the previous work, which reports only 3,345. We cannot account for this discrepancy.

Furthermore, we oversee the creation of the Bio Corpus, consisting of 17 biographical accounts and annotated with 1,594 event pairs. Despite the small size of the corpus, we feel that the greatly increased event relationship density of our samples compared to a similar number of TimeBank documents offsets the disadvantage of the small document count.

The accounts are drawn from those available at Biography.com, and describe multiple aspects of the subject's life. Because the style of the biographies tends to explore one aspect of life fully, before moving on to another, we frequently see references to events contained in previous sections. These relations, which are not only across sentence boundaries but often in entirely different paragraphs, are one of the most striking differences between TimeBank documents and those of the new corpus.

To prepare the corpus, each document was automatically event tagged through the adoption of EVITA, the Events in Text Analyzer (Sauri et al, 2005). EVITA was previously found to perform with 80.12% accuracy, a result comparable to the accuracy of graduate student annotators with basic training. The temporal relations between event pairs were then hand-annotated according to the TimeML standard (Pustejovsky et al, 2003b).

4 Methodology

In an attempt to reproduce the event relationship classification techniques of the previous work, we first implement the approach and test it on our version of the TimeBank corpus. We then demonstrate that the validated techniques are applicable to the biographical domain, and that where discrepancies do occur, the specific feature set can be modified to elicit improvements not seen in the TimeBank data. In all possible cases we utilize the same techniques and tools as the earlier work, except where sufficient information is lacking, such as in the specific implementation of the machine learning techniques. In such situations, assumptions are made as deemed necessary.

Chambers' work attempts to identify the relationships between event pairs according to a previously defined set consisting of *Before*, *iBefore*, *Includes*, *Begins*, *Ends*, and *Simultaneous*. The set of event pairs are pre-selected and chosen for preexisting relationships, so a classification of *No Relation* is not required. In order to achieve classification, a support vector machine (SVM) is implemented via the Library for Support Vector Machines (Chang & Lin, 2011) and is trained on an extensive set of thirty-five features, as detailed below.

(1) Mani	Tense, Aspect, Class, Tense_Agree, Aspect_Agree, Event Words
(2) Lapata	Subord., Before, Synsets, Lemmas
(3) Chambers	POS, Class_Agree, Temporal Bigrams, Dominance, Prepositions, Same_Sentence

Table 1. Features of classification at each stage.

The feature set was incrementally built by a number of previous experiments, as detailed in Table 1, above. Initially, five temporal attributes originally identified by TimeML as having temporal significance, are adopted. These include the tense, aspect, and class of each event, as well as the modality and polarity of each. However, per the previous work, which demonstrated modality and polarity performing with high majority baselines, we exclude them from consideration. While Chambers et al include the task of automating the identification of these features, we report results versus the gold standards taken from TimeBank.

Mani et al (2006) added features indicating an agreement between the two events in the case of tense and aspect, and Chambers extends this to include a class agreement variable. In addition to simple agreement, bigrams of tense, aspect, and class are first included by Chambers to more fully represent the relationship between the event attributes (e.g. "Past Present," "Perfect Prog").

Next to be included are the event strings themselves, extracted verbatim, and the corresponding

	Baseline	Mani	Lapata	Chambers
TimeBank – Chambers	37.22	50.97	52.29	60.45
TimeBank – New	37.11	51.97	53.79	58.22
Bio Corpus	45.67	53.14	52.89	56.65

Table 2: Accuracy of SVM classification for Temporal Relationships.

	Baseline – (Lapata)	Part-of-Speech	Prepositional Head	Class Agreement	Temporal Bigrams
TimeBank	53.79	55.99	56.48	55.02	54.84
Bio Corpus	55.40	54.77	57.34	55.71	55.49

Table 3: Accuracy of feature subset analysis. Includes all features attributed to Mani and Lapata.

Wordnet (Fellbaum, 1998) synsets and lemmas. Also included are the parts-of-speech for both event words, the two words immediately preceding each event, and that of the token immediately following the events. Bigrams for part-of-speech from each event and its preceding token are also included, as well as a bigram for the part-of-speech of the two events as related to each other.

Lapata and Lascarides (2006) first added a feature indicating whether or not two events were in a subordinate relationship, which Chambers' includes, and extends it with the addition of one indicating a dominating relationship. This information is extracted by considering the parse tree as defined by an intermediate stage of the Stanford Parser. Similar to these two linguistic ordering features, we include another feature indicating the textual ordering of the two events (true if Event 1 is before Event 2, and false if not), and one indicating whether the two events are intra- or inter- sentential (same sentence or different sentences). Finally, we adopt Chambers' use of a feature for identifying whether or not each event is a part of a prepositional phrase.

All of these features are extracted from the text via regular expressions and application of the aforementioned third-party tools (such as WordNet and the Stanford Parser). With the features extracted, the first experiment on TimeBank uses only those features identified by Mani et al. Experiments two and three incrementally grow the feature set with those identified by Lapata & Lascarides and Chambers, respectively. The feature sets can be seen in Table 1, above. Results of this reproduction of the previous work are used as a point of comparison to the results of classification on our own Bio Corpus, using the same incremental growth classification scheme as before.

Furthermore, we provide independent feature analysis of a selection of the new features added by Chambers over the Mani+Lapata set, leveraging the results to draw some conclusions as to the linguistic differences existing between the two corpora.

5 Results

We first perform classification on TimeBank with the feature set attributed to Mani, the results of which can be seen in Table 2. Our system returns an accuracy of 51.97%, outperforming Chambers' reported result by one full point. This over-performance is extended to the Lapata feature set in a 1.82 point increase over our results for Mani's features, versus the 1.32 increase seen in Chambers' reported results, which at least maintains a similar magnitude of improvement.

With the full set of features, including Chambers' additions, our system exhibits a reversal in the previous trend of over-performance. As seen in Table 2, when Chambers' reported results of 60.45%, our own system returns results of only 58.22%. Not only does this leave a void of over two percent between the expected and actual accuracies, but it represents a much smaller increase in performance between Lapata's and Chambers' feature sets on Bio. In an effort to identify an underperforming feature, although without point of comparison from previous work, we explore an independent analysis of the new features, and found all features to be performing with at least some measure of improvement, as can be seen in Table 3.

Mani's feature set, when applied to the Bio Corpus, returns similar results as on TimeBank, with slightly higher accuracy at 53.14%. This

translates to a smaller improvement over the baseline than we see in the newswire domain, but maintains approximately the same level of accuracy. Also following the same trend that is exhibited on TimeBank, the new features attributed to Lapata yield results with a small degree of improvement over the expected values at 55.4% versus TimeBank's 53.79%.

The application of the full feature set returns the expected reversal of trend, but underperforms by an even greater degree at 56.65%, leading us to suspect linguistic differences between the two corpora. In an effort to confirm this, we perform the same independent feature analysis as we performed on TimeBank. Notable results of re-classification (seen in Table 3) came from the part-of-speech features, as well as from the prepositional phrase heads. Part-of-speech was found to degrade performance and drop accuracy from 55.40% to 54.77%. Omission of the part-of-speech from a full feature set classification does not, however, improve performance over the initial classification. The prepositional phrase feature, on the other hand, returned the opposite result from part-of-speech – an improvement over the full feature set accuracy at 57.34%, strongly suggesting the importance of prepositional phrases in classification in the Bio Corpus.

6 Discussion

On TimeBank, results of temporal relationship classification return results similar to what was expected. In the simpler feature sets of Mani and Lapata, our own experiments over-perform by a small margin in each case, maintaining a similar magnitude of improvement at each step. This small but interesting variation is likely the result of the 61 additional event pairs in our version of the TimeBank corpus. Given our lack of justification for the difference, this claim is merely speculative.
On the final feature set, with the inclusion of all features set out by Chambers, we still see a small improvement over the prior feature sets, but a small magnitude of change, coming in at a high of 58.22% compared to Chambers' 60.45%. While still reasonable, a sudden underperformance compared to the previous slight over-performances is unusual. Justification for this discrepancy could be attributed to the differences in the data set, but there is also a possibility that ambiguity in the description of the features led to improper extraction techniques. Our analysis of the individual feature fails to return what we can identify as an underperforming feature, however.

In the case of the Bio Corpus, we initially see a similar trend in performance, with the feature sets attributed to Mani and Lapata performing as expected, while the full Chambers set returns a less than impressive result. Additional analysis of the individual improvements from Chambers' new features, however, identifies two outliers to performance on Bio. The underperformance of part-of-speech, and the surprising improvement based solely on the prepositional phrase feature, would suggest different linguistic trends between the two corpora.

In future explorations of this topic, we would like to expand the size of the biographical corpus and reaffirm its correctness through the use of cross-validation between multiple annotators. This would help to ensure that no unintentional biases have skewed our results. In addition, we would like to further investigate feature selection to find a best-case subset for performance on the Bio corpus. While we initially began such an analysis, the sheer number of potential combinations rendered it outside of the scope of this work.

References

Chang, C., Lin, C. (2001). LIBSVM : a library for support vector machines. Software available <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Chambers, N., Wang, S., Jurafsky, D. (2007). Classifying Temporal Relations Between Events. ACL-07, Prague.

Fellbaum, C. (1998, ed.). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Lapata, M., Lascarides, A. 2006. Learning sentence-internal temporal relations. In Journal of AI Research, volume 27, pages 85–117.

Mani, I., Verhagen, M., Wellner, B., Lee, C. M., Pustejovsky, J. (2006). Machine Learning of Temporal Relations. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (2006): 753-60. TimeML Publications.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferrom L., Lazo, M. (2003). The TIMEBANK Corpus. Proceedings of Corpus Linguistics 2003: 647-656.

Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. IWCS-5, Fifth International Workshop on Computational Semantics.

Saurí, R., Knippen, R., Verhagen, M., Pustejovsky, J. (2005). Evita: A Robust Event Recognizer for QA Systems. Proceedings of HLT/EMNLP 2005: 700-707.