12-2012

Domain Independent Ordering of Narrative Events

Lucian T. Silcox
University of Texas-Pan American

DOMAIN INDEPENDENT ORDERING OF NARRATIVE EVENTS

A Thesis
by
LUCIAN T. SILCOX

Submitted to the Graduate School of
The University of Texas-Pan American
In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2012

Major Subject: Computer Science

DOMAIN INDEPENDENT ORDERING OF NARRATIVE EVENTS

A Thesis
by
LUCIAN T. SILCOX

COMMITTEE MEMBERS

Dr. Emmett Tomai
Chair of Committee

Dr. Richard Fowler
Committee Member

Dr. Laura Grabowski
Committee Member

December 2012

ABSTRACT

Silcox, Lucian T., Domain Independent Ordering of Narrative Events. Master of Science (MS), December, 2012, 43 pp., 10 tables, 5 figures, 16 references, 26 titles.

A key aspect towards gaining a rich understanding of a text is the ability to order the events that occur in the narrative. For us, as readers, this is a simple task, accomplished without effort by recognizing subtle contextual and linguistic clues. For natural language systems, however, the task is considerably more difficult.

It is the intent of this research to automate the process of annotating and ordering temporal events in a narrative, and to measure the success of techniques trained on news articles when applied to less sequentially-structured texts. We build on the earlier work of Mani et al (2005) and Chambers et al (2007), applying machine learning techniques to the Timebank Corpus, and further, to a biographical corpus compiled for the purposes of this study.

DEDICATION

This milestone would not have been possible if it were not for the support and encouragement of my parents, Denise Silcox and Mike Adelman, who always emphasized the importance of education, even when I took them too seriously and stayed in school way too long.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Emmett Tomai, for all his support and advice during the thesis writing process. His insights into the entire thesis process and toward grad life in general were incredibly helpful to me, and helped to keep me motivated even when my own confidence began to flag.

I would also like to thank Dr. Richard Fowler, who was a frequent help and friendly ear, and Dr. Laura Grabowski, whose experience helped me to overcome some of the difficulties I faced.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

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, either in keywords (eg. yesterday, previously) or through deeper temporal inference, is 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 first corpus of temporally-annotated information opened up the possibility of applying machine learning techniques to solve these inherently difficult problems. Previous work on the topic varied widely in approach, and returned mixed results. For approaches not specifically built around a unique domain, the best reported results are only 59.43% accurate on a newswire corpus (Chambers et al., 2007). In this paper, we attempt to recreate that result, as well as apply the technique to a new domain of biographical accounts to

determine the efficacy of the pre-existing technique across domains. We then proceed to examine potential modifications to the technique to improve performance on the biographical corpus, and discuss any variance in the importance of specific linguistic features between the two domains.

Motivation

Time is important to us as a species, especially in modern societies. It pervades every aspect of our lives, from the moment when our alarm clocks wake us up, to the scheduled meetings and responsibilities throughout the day, to sitting down at dinner for some quality family *time*. In the life of any active individual, time management becomes a significant component of his or her day. We also define many aspects of our lives or of the world in terms of time and their duration.

It was this latter fact that eventually led to the adoption of this line of research. With an interest in automating the process of assisting with the identification of symptoms related to mental disorders, it quickly became apparent that a common trend that would need to be addressed was determining the duration of particular tendencies, or symptoms. Initial research showed that this problem itself was non-trivial, and a simplification of that issue eventually led to this research as it stands today.

Organization of Thesis

In the latter half of this chapter, we explore the previous work in the field of temporal extraction and annotation that have brought us to this point, including the works upon which we build our own research. In Chapter 2, we look in detail at the technical specifications of our implementation, and the reasoning behind the decisions made here. We wrap up the background

information in Chapter 3 by considering the specific corpora that we use for both training and evaluation.

In chapter 4, we examine the methodology behind the recreation of previous work, and include notes on its efficiency. This is broken down into two distinct phases to mirror previous work. We present the results of our experimentation, including those of the techniques as applied to the new domain, and begin to discuss the possible interpretations of those results. Finally, in Chapter 5, we discuss the implications of our work and expound on the interpretation of our data, before providing possible directions for future work in the area.

Previous Work

Previous work in the field has taken a number of different approaches. In domain-bounded tasks, a typical approach has been to classify events given their relationship to a number of key events that are known to always exist in the domain. This approach can be seen as applied to medical discharge summaries in the work of (Wang et al, 2008), where the key events included admission, pre- and post- operation, and discharge. In the same year, Chambers and Jurafsky (2008) demonstrated the applicability of event scripts to create what they called Narrative Event Chains for the purpose of ordering, adding some fluidity to the process but still linking to key moments in the domain.

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”). Further research conducted on news articles attempted to identify only the specific temporal relationships between two events, as seen in (Mani et al, 2006) and (Chambers et al,

2007). In both of the aforementioned studies, the events were classified by way of a simplified six-relation form of Allen's interval calculus (1984), and also presuppose an existing relationship between the two events. The latter study included the added difficulty of automating the entire process, whereas Mani et al use hand-annotated gold standards for feature detection.

In this paper, we are interested primarily in applying event ordering techniques to documents less structured than news articles, specifically biographies. It is the intention of our research to automate the process of annotating and ordering temporal events in a narrative, and thus we attempt to extend the work completed by (Chambers et al, 2007) and apply it to a less rigidly defined data set. Chambers reports best results of 59.43%, and also reports results versus the earlier work of (Mani et al, 2006) and (Lapata et al, 2007), upon which his study is built. We compare primarily to Chambers' work, but include partial results for completeness.

CHAPTER II

TECHNICAL SPECIFICATIONS

The specific requirements of the project necessitated the use of a number of previously established techniques in addition to proprietary code written to achieve our goals. To satisfy these requirements, third party implementations of the various techniques were selected which had proven success rates that were satisfactory to our needs. In this section, we explore in detail the various options that were chosen, as well as the justification for each. We begin with the selection of an implementation language, and progress from there to additional third-party libraries for said language, and on to the specific machine learning techniques that were chosen for evaluation.

Implementation Language

For the purposes of the implementation of our unique code solutions, we chose to utilize the Python programming language. Python is a general-purpose, interpreted language with an open and expressive design philosophy. The language was first designed by Dutch programmer Guido van Rossum in the early 1990's, and has progressed through several versions under his direction. Python is known for having a clear and readable syntax, extensive support for third-party modules and modules written for other languages such as C++ or Java, and a number of other advantages that led to it becoming increasingly popular in the late 2000's.

For our purposes, Python provided a number of advantages that made it an ideal choice. The first of these, and the simplest if not the least influencing, was a preexisting familiarity with the syntax of the language. Secondly, Python provides a number of native functionalities that were necessary for our implementation. First and foremost among these is the support for regular expressions, which were instrumental in the extraction of specific linguistic data from the corpora. While many languages offer regular expression support through external libraries, Python's native support reduced the complexity of installation, and thus provided an advantage over a different language, such as C++ which was also considered.

Our selection of Python was also influenced by our desire to utilize some of the functionality of a third-party platform called the Natural Language Toolkit, or NLTK, which was written for Python. The exact details of NLTK and its significance to our work are discussed later. Python's flexible support for additional platforms and modules, such as NLTK or the various machine-learning implementations we will review later, made it an ideal language for coping with any situations that may have arisen during the study.

While the current implementation has progressed to Python version 3.3.0, the version we chose to use was the 32-bit implementation of version 2.6, for compatibility reasons with third-party libraries, specifically the aforementioned NLTK. Attempts to upgrade to later versions or the 64-bit implementation were met with significant difficulties in setup, and with no real benefit for doing so, we decided against the upgrades.

Natural Language Toolkit

The Natural Language Toolkit, or NLTK, is a Python-specific platform for handling natural human language data. It provides access to a number of preexisting corpora, as well as

library functionalities for everything from tokenization and stemming to statistical classification and semantic reasoning. It was originally released in 2001 by Steven Bird, Ewan Klein, and Edward Loper, and the NLTK Project continues development to this day. With extensive documentation, including a comprehensive reference guide (Bird et al., 2009), NLTK is ideal for research conducted in natural language or related areas.

For our purposes, NLTK provided access to both a Naïve Bayes classifier, which will be described in more detail later, as well as access to Wordnet (Miller, 1995), a lexical database for English that deals with recognizing words and their synonyms and various disambiguation senses, which they call synsets and lemmas. This creates a combination dictionary/thesaurus that is unrivaled in its purpose. The figure below depicts an example result returned from Wordnet's online browsing functionality (chosen for readability over the results returned by Wordnet through NLTK in Python.)

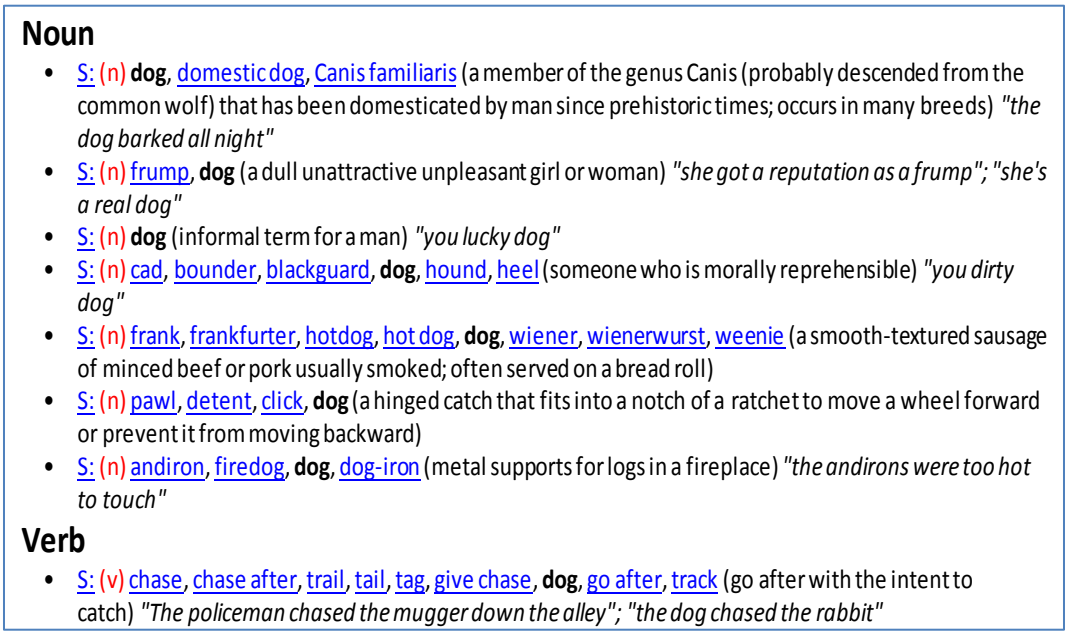

Figure 1: Wordnet Synsets for the English word "dog" (wordnet.princeton.edu, 2012)

Stanford Parser and the Penn Tagset

Part-of-speech tagging is a vital part of many natural language systems, and parsers to work out the grammatical structure of sentences can be complicated tools. As part-of-speech tagging was necessary for our work, we chose to use the Stanford Parser (Klein and Manning, 2003) for parse tree formulation and tagging. The Stanford Parser is a widely accepted and widely supported statistical parser implemented in Java and first released in 2002. The parser can take plain text and return either a phrase structure tree or a part-of-speech tagged copy of the original line, both formats which we leverage to extract crucial information. Figures 2 and 3 demonstrate both formats of result from the Stanford Parser for a sample sentence.

The Stanford Parser chooses tags for English part-of-speech tagging from a set defined and popularized by the Penn Treebank. A Treebank is a text corpus with all sentences annotated and tagged with syntactic information. The Penn Treebank, maintained by the University of Pennsylvania, provides perhaps the best known standard for part-of-speech tagging in English. For purposes of interpretation of the parsed data seen in our work, the following table contains a partial listing of the part-of-speech tag representations as seen in the Penn Tagset.

```
The/DT strongest/JJS rain/NN ever/RB recorded/VBN in/IN India/NNP
shut/VBD down/RP the/DT financial/JJ hub/NN of/IN Mumbai/NNP ,/,
snapped/VBD communication/NN lines/NNS ,/, closed/VBD airports/NNS
and/CC forced/VBD thousands/NNS of/IN people/NNS to/TO sleep/VB in/IN
their/PRP$ offices/NNS or/CC walk/VB home/NN during/IN the/DT night/NN
,/, officials/NNS said/VBD today/NN ./.
```

Figure 2: Stanford Parser Tagged Text. (nlp.stanford.edu, 2012)

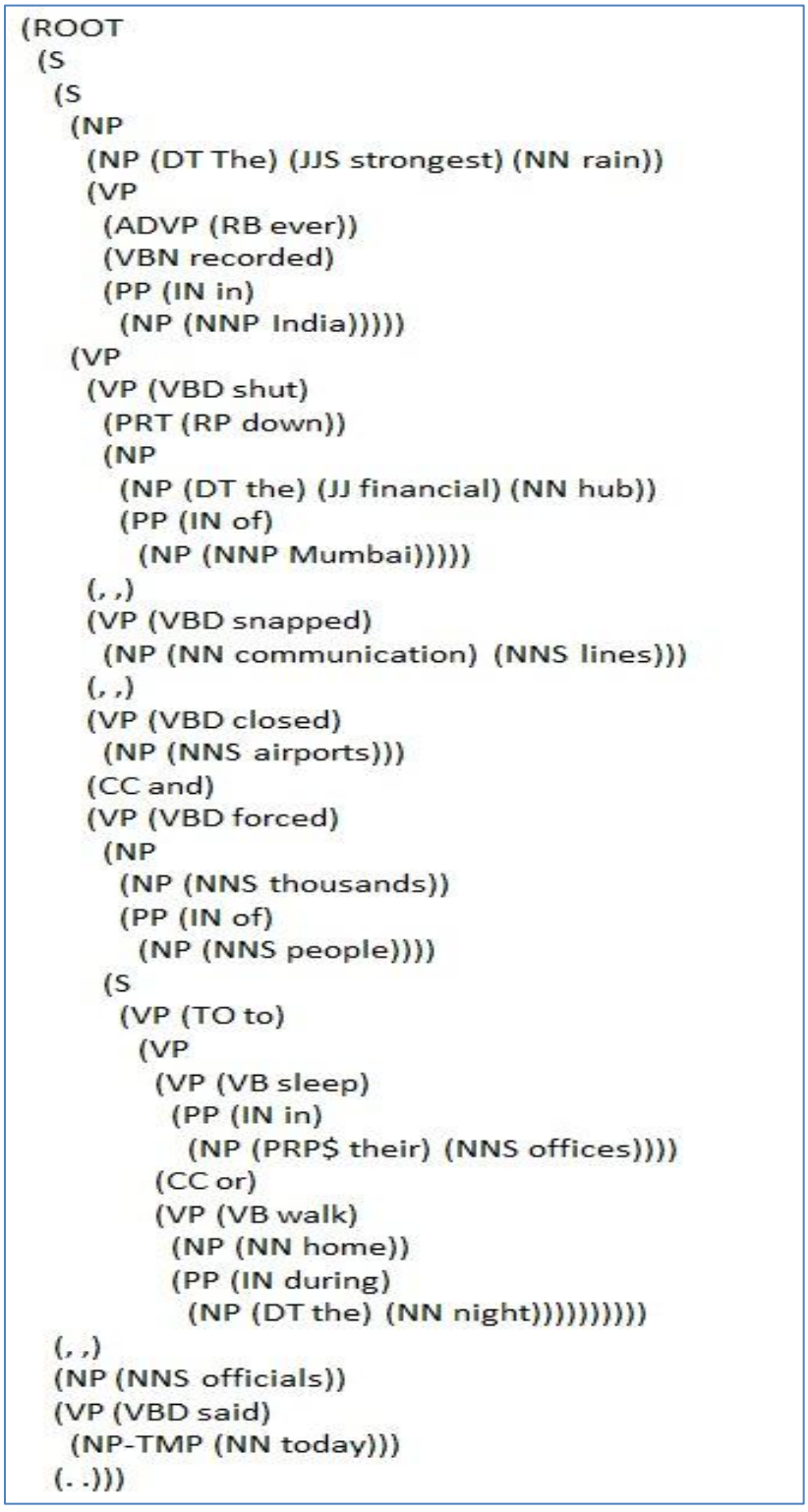

Figure 3: Stanford Parser phrase structure tree example. (nlp.stanford.edu, 2012)

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Figure 4: The Penn Tagset

Machine Learning Classifiers

In machine learning, classification is the task of deciding to which of a set of categories a new observation should be assigned. It does this by first analyzing a set of training data, which consists of a set of previously acquired observations for which the proper classification is already known. These cases are analyzed for a number of variables or features that are expected to influence the classification. Beyond this simple abstraction, classifiers vary greatly in approach and technique. In this study, we use two different classification techniques, for purposes of comparison to previous work. Below, we examine the specifics of both a Naive Bayes classifier, and a Support Vector Machine.

Naïve Bayes Classifiers

Naive Bayes classifiers are probabilistic classifiers based on applying Bayes' Theorem to the feature set. In addition to this, the features are assumed (correctly or not) to be strongly independent of each other, so the presence of any given feature is considered to be unrelated to the presence (or absence) of any other feature. For instance, you may consider that a dog is covered in fur, gives birth to live young, and is a mammal. A Naive Bayes classifier would look to all of these features as being independent of each other, and would classify further examples based on each one by itself (despite the fact that one, being a mammal, is likely linked to the other two). This assumption of independence may seem flawed (and, rightfully, can be), but classifiers of this type have repeatedly shown themselves to be extremely effective, even in situations with relatively small training sets (the independence assumption removes the need for a large training set to identify interrelations.)

We implement the Naive Bayes classifier defined in the Natural Language Toolkit in nltk.classify.naiveBayes. In addition, we apply a technique called Laplace smoothing, provided by nltk.probability. Laplace smoothing, sometimes known as additive or add-one smoothing, is a technique used to reduce variations in categorical data. We use it to correct situations where data does not occur in a sample, and instead assign a non-zero probability to the number. It is frequently used in Naive Bayes classification, especially in the area of natural language to correct this exact problem. We present results from the classification both with and without Laplace smoothing applied.

Support Vector Machine Classification

Support Vector Machines (or SVMs) are another machine learning technique for classification of observations in a supervised learning environment. Unlike Naive Bayes, which applies probabilistic techniques to determine likelihood for any given category, SVMs represent each example in the training set as a point in space, and attempt to map them in such a way that the categories are separated by the maximum sized gap. Inherently, this leads SVMs to be two-class classifiers, where any given data item will map to either category A or category B. However, multiclass support vector machines can and have been created by mapping the inputs into high-dimensional feature spaces, and performing classification on vectors.

For our specific implementation, we have chosen LIBSVM (Chang & Lin, 2011). LIBSVM, or Library for Support Vector Machines, is a multi-language, full SVM implementation which is designed to be accessible to everyone from beginners to seasoned veterans. As such, it provides a simple interface and allows for automatic, hands-free operation, but is still ultimately tunable to improve results.

CHAPTER III

DATA SOURCES AND CORPORA

TimeBank and TimeML

For training and initial evaluation, we adopt the Timebank Corpus (Pustejovsky et al., 2003a), version 1.1, preferring it to a more recent update solely for purposes of comparison to the earlier work of Chambers. The Timebank 1.1 corpus is a newswire domain consisting of 186 news articles annotated to the TimeML specification (Pustejovsky et al., 2003b), and including 3406 event pairs with classified relationships. This creates a point of contention in our intended comparison, as the work of Chambers et al (2007) reports only 3345 event pairs. Attempts to explain the discrepancy through analysis of possibly excluded subgroups failed to produce an identical number of event pairs. Our method for event pair extraction has been confirmed to be identical to previous research via personal communication with the original author (Chambers, 2011). However, the difference of 61 events, a 1.82% increase in dataset size over the previous work, is not expected to impact the results so greatly as to invalidate our own findings, and thus additional time was not spent in trying to justify the discrepancy.

Timebank documents are provided in XML format, tagged in line with much information pertaining to the categorization of events and temporal signals, among other tags. In addition, the inclusion of tags to the end of the document provide key details on the aforementioned tagged events (MAKEINSTANCE), as well as any identified relationships between events (TLINK), as

well as more obscure information. Figure 5 provides a look at a sample Timebank document (with some information removed for formatting purposes.)

```
<TEXT>
<s><ENAMEX TYPE="LOCATION">MOSCOW</ENAMEX> (<ENAMEX TYPE="ORGANIZATION">AP</ENAMEX>) _ The mayor of <ENAMEX TYPE=
"LOCATION">Moscow</ENAMEX>  has <EVENT eid="e1" class="OCCURRENCE" ><EVENT eid="e18" class="OCCURRENCE" >allocated
</EVENT></EVENT> funds  <SIGNAL sid="s66" >to</SIGNAL> <EVENT eid="e2" class="I_ACTION" >help</EVENT>
<EVENT eid="e3" class="OCCURRENCE" >build</EVENT> a museum in honor of <ENAMEX TYPE="PERSON">Mikhail Kalashnikov
</ENAMEX>, the Russian who <EVENT eid="e6" class="OCCURRENCE" >gave</EVENT> his name to the world's most widely wielded
weapon, <EVENT eid="e91" class="REPORTING" >according</EVENT> to a news agency <EVENT eid="e55" class="REPORTING" >
report</EVENT> <TIMEX3 tid="t15" type="DATE" temporalFunction="true" functionInDocument="NONE" value="1998-02-27"
anchorTimeID="t13" >Friday</TIMEX3>.</s> <s><ENAMEX TYPE="PERSON">Kalashnikov</ENAMEX> <EVENT eid="e7" class=
"OCCURRENCE" >designed</EVENT> the AK-47 automatic rifle, famed for its reliability and effectiveness.</s> <s>Together
with its derivatives, the <ENAMEX TYPE="PERSON">Kalashnikov</ENAMEX> rifle  has  been <EVENT eid="e8" class="OCCURRENCE"
 >used</EVENT> by the police and armies of <CARDINAL>55</CARDINAL> countries and an assortment of guerrillas,
terrorists and just plain thugs around the globe.</s>
<s><CARDINAL>Seventy-five million</CARDINAL> copies of the rifle  have  been <EVENT eid="e9" class="OCCURRENCE" >built
</EVENT> <SIGNAL sid="s88" >since</SIGNAL> it <EVENT eid="e10" class="ASPECTUAL" >entered</EVENT> <EVENT eid="e79" class
="STATE" >production</EVENT> <SIGNAL sid="s97" >in</SIGNAL> <TIMEX3 tid="t16" type="DATE" temporalFunction="false"
functionInDocument="NONE" value="1947-02" >February 1947</TIMEX3>, making it history's most widely distributed weapon.
</s> <s>UR The <ENAMEX TYPE="ORGANIZATION">ITAR-Tass</ENAMEX>, 4th graf pvs (pvs/ji)</s>
</TEXT>

<MAKEINSTANCE aspect="PERFECTIVE" eiid="ei104" tense="PRESENT" eventID="e1" />
<MAKEINSTANCE aspect="NONE" eiid="ei105" tense="NONE" eventID="e18" />
<MAKEINSTANCE aspect="NONE" eiid="ei106" tense="NONE" eventID="e2" />
<MAKEINSTANCE aspect="NONE" eiid="ei107" tense="NONE" eventID="e3" />
<MAKEINSTANCE aspect="NONE" eiid="ei108" tense="PAST" eventID="e6" />
<MAKEINSTANCE aspect="NONE" eiid="ei109" tense="NONE" eventID="e91" />
<MAKEINSTANCE aspect="NONE" eiid="ei110" tense="NONE" eventID="e55" />
<MAKEINSTANCE aspect="NONE" eiid="ei111" tense="PAST" eventID="e7" />
<MAKEINSTANCE aspect="PERFECTIVE" eiid="ei112" tense="PRESENT" eventID="e8" />
<MAKEINSTANCE aspect="PERFECTIVE" eiid="ei113" tense="PRESENT" eventID="e9" />
<MAKEINSTANCE aspect="NONE" eiid="ei114" tense="PAST" eventID="e10" />
<MAKEINSTANCE aspect="NONE" eiid="ei115" tense="NONE" eventID="e79" />
<MAKEINSTANCE aspect="NONE" eiid="ei116" tense="NONE" eventID="e2" />
<TLINK relatedToEventInstance="ei110" eventInstanceID="ei109" relType="IDENTITY" />
<TLINK relatedToEventInstance="ei112" eventInstanceID="ei111" relType="BEFORE" />
<TLINK relatedToTime="t15" eventInstanceID="ei110" relType="BEFORE" />
<TLINK relatedToEventInstance="ei104" eventInstanceID="ei108" relType="BEFORE" />
<TLINK signalID="s97" relatedToTime="t16" eventInstanceID="ei111" relType="BEFORE" />
<TLINK signalID="s97" relatedToTime="t16" eventInstanceID="ei114" relType="IS_INCLUDED" />
<TLINK signalID="s97" relatedToTime="t16" eventInstanceID="ei113" relType="BEGUN_BY" />
<TLINK relatedToEventInstance="ei111" eventInstanceID="ei113" relType="AFTER" />
<TLINK relatedToEventInstance="ei114" signalID="s88" eventInstanceID="ei113" relType="AFTER" />
<SLINK subordinatedEventInstance="ei107" eventInstanceID="ei106" relType="MODAL" />
<SLINK subordinatedEventInstance="ei104" eventInstanceID="ei110" relType="EVIDENTIAL" />
<SLINK signalID="s66" subordinatedEventInstance="ei116" eventInstanceID="ei104" relType="MODAL" />
<ALINK relatedToEventInstance="ei115" eventInstanceID="ei114" relType="INITIATES" />
</TimeML>
```

Figure 5: Sample selection from a Timebank document (APW19980227.tml.xml).

The Bio Corpus

For evaluation in a new domain, we have compiled a corpus consisting of seventeen biographical narratives, hand tagged with 1594 event pairs with confirmed temporal relationships. The documents were chosen at random from those available at Biographies.com,

and were not screened or altered in any way. While we realize the problems that arise from working with a small dataset, we believe this disadvantage will be offset by the large number of annotated event pairs, which far exceeds the event pair content of a similar number of Timebank documents. The small size of the corpus was strongly influenced and necessitated by the time requirement of annotating each document by hand, and ideally would be increased given additional resources.

As part of the process of annotating the Bio Corpus, we first adopted the use of the Events in Text Analyzer (Evita), as proposed by Sauri, Knippen, Verhagen, & Pustejovsky (2005). Evita was demonstrated to perform with an F-measure of 80.12% accuracy, which was found to be comparable to that of graduate linguistic students with basic training in annotation, which was judged for the purposes of our research to be effective enough to use to forego hand annotation in the task of event identification. Evita is included as a part of the toolkit associated with the Temporal Awareness and Reasoning Systems for Question Interpretation (TARSQI) Project (Verhagen et al., 2005), which is a series of applications for automating the annotation of temporal data in narratives, based on the aforementioned TimeML standard.

The annotation of the temporal relationships between events was hand-specified according to the TimeML 1.1 standard, to match the Timebank corpus. The relationships were based on Allen's interval calculus and its thirteen identified categories (1984). For purposes of comparison to the previous work, we condense this superset into six tighter categories. Six of the original thirteen relations are inverses of others, so we begin by merging the inverses into their respective matches. The *Identity* relationship, similarly, maps into Simultaneous. It is important to note that Timebank identifies a *During* relationship that is not discussed in documentation. Here we accept the handling of Chambers et al and merge *During* into *Includes* (and with its

inverse *Included_by*). Our final set of categories, which mirror those of Chambers, are *Before*, *iBefore, Includes, Begins, Ends,* and *Simultaneous.*

CHAPTER IV

METHODOLOGY

We attempt to replicate both the automatic identification of the temporal attributes, as well as the event relationship classification of the previous work. In both cases, we apply the techniques as previously defined to replicate the results of the previous work. We then show that the validated techniques are applicable to the biographical domain, but that the specific feature set can be modified to elicit improvements not seen in the TimeBank data.

Specifically, we first show that the learning of temporal attributes in the Bio Corpus is effective for the tense and class features, but suffers in the identification of aspect, presumably due to a higher baseline performance. In an attempt to improve performance of aspect, we introduce a new Have-Be Bigram feature to its classification. We also include the consideration of Wordnet lemmas for class, a feature considered but not selected in the previous work. We show that in the Bio Corpus, its inclusion elicits an improvement over classification without it.

In the second phase, we demonstrate that despite discrepancies in the accuracy of the first, the technique is still applicable to the Bio Corpus, often exhibiting improvements of a magnitude similar to those seen on TimeBank. In the case of discrepancy, we perform an independent feature analysis of the features of the varying set to identify the linguistic tendencies that set the two corpora apart.

To accomplish this we adopt the two phase model of the previous work. Phase 1 is concerned with learning the temporal attributes of an event, as previously identified by

TimeBank, while Phase 2 deals with identifying the relationships between pairs of events with a pre-existing relationship assumed. In all possible cases we utilize the same techniques and tools, except where sufficient information is lacking, such as in the specific implementation of machine learning techniques. In such situations, assumptions are made as deemed necessary.

Phase 1: Learning Temporal Attributes

According to the TimeML specification, events are associated with five features of temporal importance - *Tense*, *Aspect*, *Class*, *Modality*, and *Polarity*. Grammatical *Tense* and *Aspect* are mandatory to temporal understanding as they locate the event in time, and provide information as to whether the event was on-going or already completed at the time of mention. *Modality* and *Polarity* are important for recognizing mentioned events that were only hypothetical, or explicitly did not occur. The final attribute, *Class*, is the type of event, which Timebank classifies as one of seven, seen in Table 1, which is also found in (Chambers et al., 2007) and (Pustejovsky et al., 2003).

TENSE	None, Present, Past, Future
ASPECT	None, Prog, Perfect, Prog_Perfect
CLASS	Occurrence, Reporting, I_Action, Perception, Aspectual, State, I_State
MODALITY	None, To, Should, Would, Could, Can, Might
POLARITY	Positive, Negative

Table 1: Possible Values for Classification of Temporal Attributes

In order to learn the above attributes, we consider the importance of a number of easily extractable linguistic features. We begin by extracting the event word itself via regular

expression, as well as the two words immediately preceding the event. We then identify the part-of-speech (POS) of the three words via use of the Stanford Parser. In some circumstances, sentence (or occasionally document) boundaries prevented the extraction of one of more preceding words. While we do not know how this situation was handled in previous work, the decision was made to leave the values as null, rather than cross over to a different sentence.

Next, the event word was used to find synsets (lists of synonyms), and lemmas (canonical stemmed words) via Wordnet (Fellbaum, 1998). Regular expression extraction was then used to identify the presence of certain keywords existing before the event word. First, attention was paid to any variation of *Have* or *Be*, which are both auxiliary verbs that often help to convey tense or aspect (e.g. "have completed," "was completing"). Next, modal words were identified, such as *Could* or *Might*, which would be of importance in identifying modality, and finally the presence of *Not*, which would indicate polarity.

TENSE	POS-2-event, POS-1-event, POS-event, have_word, be_word
ASPECT	POS-event, modal_word, be_word
CLASS	Synset
MODALITY	None
POLARITY	None

Table 2: Feature Selection for Temporal Attributes in Naïve Bayes.

Using these features, we train a Naïve Bayes classifier as provided by NLTK to identify the five temporal attributes in the TimeBank v1.1 data. For each attribute, we assign a subset of features according to the previously selected feature sets of Chambers et al and seen in Table 2. To ensure accuracy, results are calculated via 10-fold cross-validation. The results of this

classification are compared to the published results of Chambers et al to demonstrate the validity of the technique first-hand, and, furthermore, identify any discrepancies in the result that could bias further experimentation.

The second classification attempt utilizes the same experimental setup as the first, but applies the previously validated technique and feature sets to the biographical corpus. Here we demonstrate the efficacy of the technique when applied to the new domain, and, through comparison with the TimeBank results, identify the variances in linguistic features that lead to discrepancies in the efficiency of the technique across domain. We show that tense and class follow similar enough patterns to benefit from the previously selected features, but identify aspect as an outlier that requires alteration of the selected features to improve.

Finally, we leverage the information garnered in the second classification to propose the addition of new features in the hopes of improving performance in the new domain. Specifically, the inclusion of a Have-Be Bigram is found to demonstrate improvements for tense and aspect classification, and the previously removed lemmas are found to positively impact the performance for class. Classification with the new, expanded feature set is trained and tested separately on both TimeBank and the Bio Corpus, and the results compared to determine the utility of the new features in each case.

Phase 2: Event Pair Classification

In this stage, we attempt to identify the temporal relationship that exists between two events from the previously defined set 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 trained on an extensive set of 35 features, as detailed below.

The first features to be considered are those temporal attributes first identified in Phase 1, with the exclusion of modality and polarity due to high majority class baselines which led to a lack of effective feature selection in previous work. (1,2) *Tense*, (3,4) *Aspect*, and (5,6) *Class* are included independently for both events. In addition to this, Mani et al (2006) added features indicating an agreement between the two events in the case of (7) tense and (8) aspect, and Chambers extended this to include an additional (9) class agreement variable. In addition to simple agreement, (10,11,12) 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 (13,14) event strings themselves, extracted verbatim, and the corresponding Wordnet (15,16) synsets and (17,18) lemmas, as previously identified during phase one. Also from the previous phrase, the Stanford Parser-identified parts-of-speech are included for (19,20) both event words, the (21-24) two words immediately preceding each, and, as a new addition at this stage, the part-of-speech of the (25,26) token immediately following each event. In addition, bigrams for part-of-speech from each (27,28) event and its preceding token are included, as well as a (29) 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 (30) subordinate relationship (wherein one event is in an independent clause and the other occurs in a subordinate clause). Chambers' includes this feature, and extends it with the addition of one indicating a (31) dominating relationship (that is, one event exists in a phrase that

is, in the syntactic parse tree, in a daughter node of the phrase containing the other event). This information is extracted by considering the parse tree as determined by an intermediate stage of the Stanford Parser. Similar to these two linguistic ordering features, we include another feature indicating the (32) 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 (33) 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 (34,35) prepositional phrase, the values of which range over 34 English prepositions.

With the aforementioned features identified and extracted, we progress to classification. In all cases, the features are represented as binary values. In the case of previously multi-valued features, such as event word, the feature is exploded into a larger set of binary representations. Classification is done via LIBSVM, utilizing the provided grid search script for parameter selection.

The first set of experimentation is trained and tested on the TimeBank corpus to validate previous results. We perform three classifications at this stage, consisting of the independent feature sets of Mani, Mani and Lapata, and finally, Mani, Lapata, and Chambers. These feature listings are identified in Table 3. For each set of features, we compare to the previously published results to confirm the validity of the technique, as well as determining baselines for the magnitude of improvement leveraged by each addition to the feature set.

The second set of classifications are identical in form to the first, only applied to the Bio Corpus. In addition to demonstrating the efficacy of the technique across domains, we utilize the incremental nature of the classifications to identify the discrepancies between results. We look to

Mani	Tense, Aspect, Class, Tense_Agree, Aspect_Agree, Event Words
Lapata	Subordination, Before, Synsets, Lemmas
Chambers	POS, Class_Agree, Temporal Bigrams, Dominance, Prepositions, Same_Sent.

Table 3: Feature Listings for SVM classification

this information to justify performing a series of classifications with incremental addition of features from Chambers' set. In so doing, we emphasize the linguistic differences between the two corpora, and reveal modified feature sets that lead to improvements in the Bio Corpus.

CHAPTER V

RESULTS

We present the results of classification below, divided into Phase 1 and 2 and further based on the primary corpus of consideration at each stage.

Phase 1: Identification of Temporal Attributes

Results of Naïve Bayes Classification on TimeBank

In comparison to Chamber's original work, majority class baselines were calculated for our copy of the Timebank v1.1 corpus, and were found to differentiate from Chambers' reported results by ~1% in the categories of tense, aspect, and class. Given the aforementioned variance in the number of reported event pairs, we assume a similar variance in the number of classified events in the corpus, although we cannot verify this as that number was not previously published by Chambers. The exact discrepancy can be seen in Table 4, below, and does not does not seem to impact modality and polarity, which still are found to be in excess of 98% (specific numbers were not reported in the previous work).

When we perform the classification with the inclusion of the previously selected features, we find that the results also skewed somewhat from the reported numbers. All three attributes suffer penalties to their expected performance, with aspect exhibiting the greatest shortcoming,

TimeBank Corpus	TENSE	ASPECT	CLASS	MODALITY	POLARITY
Baseline – Chambers	52.21	84.34	54.21	~98+	~98+
Baseline – New	52.63	85.21	53.02	98.32	99.52
NB w/ Laplace Smoothing - Chambers	88.28	94.24	75.2	--	--
NB w/ Laplace Smoothing - New	83.62	88.72	71.65	--	--
NB w/o Laplace Smoothing - New	83.63	88.83	73.02	--	--

Table 4: Results of Naïve Bayes classification on TimeBank

falling approximately 5.5 percentage points shy of the expected outcome at 88.72%. Tense was close behind at 83.62% versus 88.28% expected, a 4.66 point difference.

Class suffered the least, at only an approximate 3.5 point drop in expected performance. Interestingly, performing the classification without the additive Laplace smoothing returned better results, albeit the only substantial increase was in Class, which saw an improvement of almost 1.4 points.

This result demonstrates that our understanding of the previous work is accurate enough to return similar if not identical results. We discuss potential causes of the discrepancy in Chapter 5, and the effect that it may have on further experimentation. Despite the variance in results, we find that the accuracy is sufficient enough to proceed to analysis on the Bio Corpus, where results are compared versus our own results on TimeBank, and not against the gold standard set by Chambers.

	TENSE	ASPECT	CLASS	MODALITY	POLARITY
Baseline – Bio Corpus	42.99	96.35	89.22	98.34	99.34
NB w/ Laplace Smoothing	80.27	95.84	90.48	--	--
NB w/o Laplace Smoothing	80.16	95.85	93.03	--	--

Table 5: Results of Naïve Bayes classification on Bio Corpus.

Results of Naïve Bayes Classification on the Bio Corpus

When applied to the Bio Corpus, we identify several differences from the TimeBank data. Initially, the majority class baselines for the Bio Corpus are found to be much different, with tense exhibiting a lower baseline, and aspect and class both coming in significantly higher, as can be seen in Table 5. Class experiences the greatest magnitude of variance, at 89.22% versus the 53.02% of TimeBank, while aspect exhibits the most shocking change, rising from an already high 85.21% to 96.35%, where it is nearly as high as modality and polarity. These latter two unfortunately both perform in the expected range of greater than 98%. The greatly increased baseline of aspect gives rise to an interesting situation in the second phase, as at such a high baseline, it qualifies to be excluded, as modality and polarity were in previous work. For sake of completeness, we include it in the initial classification, but also explore the ramifications of its removal from the feature set in Phase 2.

Classification using Chamber's features, as may be expected from the variance in baseline, differs wildly from the results on TimeBank. Tense increases by a large amount over its baseline, improving from 42.99% to 80.27%. While still underperforming, the amount of improvement over its baseline is comparable to that of the tense feature on TimeBank, with the two corpora improving by within 1.5 points of each other (35.99 TimeBank versus 37.22 Bio).

Regrettably, however, the loss of 8% accuracy when compared to Chambers' reported results on TimeBank raises questions as to the efficacy of the automatic features in Phase 2.

Slight improvements in tense classification can be elicited with the addition of a new feature – a bigram of the Have and Be word features – and the removal of the individual Have and Be features from consideration. On both the TimeBank and Bio corpora, these feature changes resulted in an approximate quarter of a point increase over the standard Chambers features. The effect of these modifications to the feature set on both the Bio Corpus and TimeBank can be seen in Table 6.

Aspect actually suffers a performance hit when we attempt to classify with Chambers' features, dropping from the previously noted high baseline by half a point to 95.84%. This degradation further indicates that, with such a high baseline, the consideration of aspect should be removed from Phase 2. However, improvements can be found. Inclusion of the aforementioned Have-Be bigram feature in aspect consideration reveals a 1.6 point increase over even the baseline, claiming 97.95% accuracy. TimeBank classification does not seem to benefit from the same consideration, gaining only .36 points of improvement over the standard features. Given the difference that aspect plays in the two corpora, as evidenced by the wildly differing baselines, this variance in improvement based on a single feature is unsurprising.

In the area of class, the already high baseline improves slightly with the standard features and evaluates to 90.48% accuracy. Notably, the inclusion of Laplace smoothing hurts class performance, as removing it notes an improvement of over 2.5 points (with negligible changes to tense and aspect). As with the previous features, we can further improve with some variance in the feature set, this time including lemmas along with the standard synsets. This inclusion boasts an improvement of another 3 points on the Bio Corpus, bring it to over 96% accuracy. That this

already examined feature was not selected for the TimeBank corpus is unsurprising when we see that including it there actually degrades performance to 69.17%.

	Bio Corpus	TimeBank
TENSE w/ Have-Be Bigram	80.54	83.80
ASPECT w/ Have-Be Bigram	97.95	89.08
CLASS w/ Lemmas	96.16	69.17

Table 6: Feature Improvements for Temporal Attributes

With these initial results, we demonstrate that the previous approach for the identification of the three temporal attributes works as expected for the cases of tense and class, but fails to exhibit an equal performance for aspect, which actually suffers for the attempt at classification. We further show that with the inclusion of new features, specifically a bigram of the Have and Be word features, we can reverse the performance of aspect and tease out a not insubstantial improvement over the baseline. We also demonstrate that the previously excluded lemmas feature can be selected for class consideration in the Bio Corpus, eliciting a 3% improvement. These features are shown to be principally unique to the Bio Corpus, as the same degree of improvement is not seen in TimeBank.

Phase 2: Event Pair Relationships

As we progress towards event-pair relationships, we choose to present results of classification using the gold standard for the three relevant temporal attributes. This decision was based on the much higher classifications of both aspect and class (after classification, both

scored less than a 4% shy of 100), and because the remaining attribute, tense, while exhibiting a good degree of improvement, still underperformed the published results of the previous work by 8%. These considerations, in combination with the known discrepancies in the source data that already provide a somewhat unstable point of comparison, led us to reduce the number of potential variables, and base our reported results on the best possible case.

Results of SVM Classification on TimeBank

For Phase 2, we once again begin by comparison of the baselines between Chambers' results on TimeBank and our own experience. Unlike Phase 1, where even the baseline showed a slight variance, which was attributed to a potential variance in the number of reported events, in Phase 2, the baselines match almost exactly, despite the known discrepancy in number of event pairs. Compared to Chambers' 37.22%, we see an accuracy of 37.11% on TimeBank, which provides a confident starting point for further comparison. As in previous work, we compare not only against the full feature set previously reported, but also against the results of subsets accounting for the original work of Mani and the additional features added by Lapata.

In the first case, Mani included the five temporal attributes (of which we exclude modality and polarity), the event words themselves, and the Tense and Aspect agreement features. Using gold standards for the temporal features, we see a return of 51.97%, outperforming Chambers' reported result by one full point.

Lapata further added features representing the subordinating and textual ordering (before) relationships, as well as the inclusion of lemmas and synsets of the event words. These additional features result in a 1.82 point increase over our results for Mani's features, versus the 1.32 increase seen in Chambers' reported results.

TimeBank Corpus	Baseline	Mani	Lapata	Chambers
SVM – Chambers (Gold)	37.22	50.97	52.29	60.45
SVM – New (Gold)	37.11	51.97	53.79	58.22

Table 7: Results of SVM classification on TimeBank

TimeBank Corpus	Baseline (Lapata)	Part-of-Speech	Prepositional Head	Class Agreement	Temporal Bigrams
SVM Performance	53.79	55.99	56.48	55.02	54.84

Table 8: Feature subset analysis on TimeBank. Includes all features of Mani and Lapata.

The final stage involves adding all the remaining features, which massively increases the size of the feature set. At this stage, Chambers reports an accuracy of 60.45% with gold standard temporal attributes, which reverses the previous trend by outperforming our own results of 58.22%. Not only does this leave a void of over two percent between the expected and actual accuracies, but it demonstrates a much more impressive increase in performance between Lapata's and Chambers' feature sets on TimeBank, almost doubling our improvement on Bio with the same features. In an effort to identify an underperforming effort, although without point of comparison from previous work, we performed further subset 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 8.

These results confirm that despite the discrepancies in the dataset, and, furthermore, those in the classification of the temporal attributes in Phase 1, the standard approach for identification of the temporal relations between events still performs admirably. In the subsets associated with Mani and Lapata, we see results that over-perform the previously published numbers by a slight

margin, while the Chambers set reveals an underperformance but that is still reasonably close to the reported accuracy. The cause of this sudden shift in the performance trend is discussed later in Chapter 6.

Results of SVM Classification on the Bio Corpus

For the Bio Corpus, our baseline for performance was found to be much higher than that of Timebank, evaluating to 45.67%. This change is, in itself, interesting, at is speaks to a different distribution of event relationships in our biographical narratives, compared to the newswire documents of TimeBank.

Mani's feature set, when applied to the Bio Corpus, return 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.

At this stage, in the interest of testing the effect of features with baselines as high as 96%, we experimented with classifying Mani without the addition of the aspect-related features (aspect for both events, and the aspect agreement feature). We also applied the same classification to TimeBank, to compare the performance degradation in both high-baseline and mid-baseline environments. As expected, removing aspect from the Bio Corpus resulted in a very minor penalty, dropping performance by a quarter of a point to 52.89%. Conversely, TimeBank without aspect suffered a 4.5 point drop to 47.42%.

Classification with Lapata's additional features yield results close to the expected values, with both accuracy and magnitude of increase being commensurate with TimeBank. In terms of accuracy, we reach 55.4% versus TimeBank's 53.79%, which translate to a 2.26 and 2.01 point

Bio Corpus	Baseline	Mani	Mani – No Aspect	Lapata	Chambers
SVM – New (Gold)	45.67	53.14	52.89	55.40	56.65

Table 9: Results of SVM classification on BioCorpus

increase over Mani, respectively. This similarity would suggest that the new features included at this stage likely share similar patterns in both corpora.

With the final addition of the new features from Chambers' work, we see the most unexpected change in classification. Performance still improved over the Lapata feature set, but only to 56.65%, which is only a 1.25 point increase. The same features on TimeBank gave rise to a 4.43 point improvement, notable even given the discrepancy between our results and those that Chambers' reported. This change in both accuracy and magnitude of improvement likely suggests notable changes in the structure of the source material from that of the newswire documents of TimeBank.

In an effort to identify the cause of the change in expected performance, we re-classified the data using subsets of features consisting of the Lapata+Mani base with each additional feature from Chambers, as we did to identify the discrepancy in TimeBank itself. Class agreement and the bigrams of Tense, Aspect, and Class were both found to produce minor improvements over our reported results for Lapata, as can be seen in Table 10. The variance in magnitude of these improvements over those changes seen in TimeBank also suggests interesting changes in linguistic patterns in the two corpora.

Notable results of re-classification came from the part-of-speech features, as well as from the prepositional phrase heads. Part-of-speech was by far the most shocking result, actually

hurting performance and dropping accuracy from 55.40% to 54.77%. Omission of the part-of-speech from a full feature set classification does not improve performance over the initial 56.65%, however, instead returning a lower 55.71%. The prepositional phrase feature, by itself, returned the opposite result from part-of-speech - a significant improvement over the full feature set accuracy at 57.34%, strongly suggesting the importance of prepositional phrases in classification in the Bio Corpus.

With these results, we demonstrate the expected performance of the technique on the Bio Corpus, with each feature set returning results comparable to what was previously seen on TimeBank. Again, the largest discrepancy is found in the Chambers set, and so an independent feature analysis reveals linguistic trends in the biographical data. Notably, the highest results returned at this stage of our research were found when considering Lapata's feature with the inclusion of only the prepositional head feature.

Bio Corpus	Baseline (Lapata)	Part-of-Speech	Prepositional Head	Class Agreement	Temporal Bigrams
SVM Performance	55.40	54.77	57.34	55.71	55.49

Table 10: Feature subset analysis on the Bio Corpus. Includes all features of Mani and Lapata.

CHAPTER VI

EVALUATION AND DISCUSSION

We have previously shown that the techniques laid out by Chambers et al, an extension of the works of Mani et al and Lapata and Lascarides, are sound enough in their approach that they are capable of producing reasonable results in a new, unanticipated domain of biographical data. Discrepancies do occur, but can often be mitigated with alterations to the established feature set, as we demonstrated in phase 1 with the inclusion of the have-be bigram and lemmas, or in phase 2 with the identification of prepositional phrase head as the most meaningful feature of the Chambers additions.

While our results are somewhat tempered by unexplained discrepancies in both data and performance, valid conclusions can still be drawn pertaining to the efficacy of the pre-existing techniques on a new domain. We begin by examining the replication efforts and results, before advancing to an investigation of results on the Bio Corpus, and how improvements can be made in this new domain.

A failure to accurately replicate the results of Chambers' previous work is initially disappointing, and is made more so by being unable to supply reasonable justification for the failures. This important issue ultimately begins to arise from an unexplained variance in the numbers of reported event pair relationships pre-existing in the TimeBank Corpus. A lack of supporting information fails to reveal if the discrepancy is only in the event-relation count, or

whether it extends to the number of identified events in general. A brief personal communication with the principal author of the previous study (Chambers, 2011) failed to identify a cause for this discrepancy, in either source material or data extraction technique. It is the belief of this author that a number of identified events and, by extension, the event-pair relationships they were involved in, were removed from consideration, likely due to being of a form that makes extraction of the noted features difficult or impossible, such as cardinal numbers or multi-word event phrases. Such features are suspect based wholly on our own difficulty in attempting to handle their extraction. While this removal cannot be confirmed or denied, it is a reasonable and perhaps likely assumption.

Whatever the cause of the discrepancies between our observations of the TimeBank 1.1 data and that previously reported, the effects are hard to predict. We would expect to see some variance in the results of classification due solely to having a slightly different amount of data to train and test on, but the results could be more wildly skewed if the change is, as we postulate above, due to the removal of a group of outliers that defy common trends. Without being able to provide a definitive reason for the discrepancy, it is impossible for us to estimate the effect that this will have on the classification efforts. This inevitability is thus included as a potential justification for any discrepancy in classification on TimeBank, although other justifications are provided as well, where applicable.

Classification of the five temporal attributes on TimeBank was found to underperform predicted results, often by a considerable margin. In addition to the aforementioned data discrepancy, this unanticipated fluctuation could be the result of a lack of clarity in the exact intention behind the feature extractions. This is somewhat supported by the magnitude of the

discrepancy in the three classified features (excluding Modality and Polarity, as described earlier).

In the case of Class, the discrepancy between Chambers' result and our highest reported accuracy (in this case, with the removal of Laplace smoothing), is only 2.18% and based exclusively on the event word Wordnet synset feature, which is unambiguous in its intent. On the other hand, Tense and Aspect both experience a much greater magnitude of performance deterioration, at 4.65% and 5.41%, respectively. In the feature selection for both of these attributes, however, we see some subset of the Have, Be, and Modal features (those that identify the existence of such words preceding the event itself). These features lack the same degree of clarity in how they should be extracted, as previous work does not indicate what scope of the sentence should be examined for their existence. It was assumed for our purposes that we only consider the same scope already examined for part-of-speech (up to two words before the event.) Other potentialities include examining the entirety of the sentence preceding the event, or only those parts of the sentence between any preceding events and the one in question. Exploration of these alternatives was considered, but was ultimately judged to be outside of the scope of this study, as the enumeration of possibilities between all features made finding a "correct" implementation unlikely, if it were identifiable as such at all.

In our new biographical domain, the temporal attributes initially differed from TimeBank in the majority class baseline, which demonstrated unexpected levels of variation in all examined features (Tense, Aspect, and Class). Tense began with a baseline almost 10% lower than on TimeBank, but exhibited an improvement via classification of an even greater magnitude (37.28% Bio versus 30.99% TimeBank). The final highest accuracy for Tense on the

biographical domain was still 3.35% lower than our best results on TimeBank, but is close enough to demonstrate the efficacy of the technique for that particular feature.

In the case of Aspect, however, and, to a lesser degree, Class, the initial baseline calculation was found to be abnormally high in comparison, with Class at 89.22%, and Aspect at 96.35%, the latter almost as high as Modality and Polarity, which were expressly excluded due to their high baselines. With such high baselines, classification was unlikely to return greatly improved results, and experimentation confirmed this. Aspect actually suffers from application of classification with the previously selected features, dropping by half a percentage point, while Class improved by only 1.26%, as opposed to a 20% improvement on TimeBank. Such a lack of improvement is not surprising given the high baseline, but it makes it difficult to judge the efficacy of the features.

In an effort to improve performance despite the high baselines, we began experimenting with the inclusion of additional features to the classification, as well as different subsets of features. Improvements were found for all three of the classified attributes, although the improvements to Tense were small enough to be almost inconsequential. In the case of Aspect, however, the addition of a bigram feature based on the Have and Be features, improved performance to 97.95%, which is a 1.6% improvement over the baseline, and an even larger improvement over previous classification attempts. TimeBank also improved with the inclusion of the feature, but did not exhibit the same magnitude of improvement. In the case of Class, we include the Wordnet lemmas in consideration. This penalizes performance on TimeBank, so it is not surprising that a feature that was initially extracted was not chosen as an important feature for classification, but on the Bio Corpus, lemmas improve performance by over 3%.

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. This cannot be confirmed given our lack of justification for the discrepancy, but given the relatively small fluctuation in the result, we find it to be a reasonable assumption.

On the final feature set, with the inclusion of all features set out by Chambers, we still see an 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 a step in the right direction, a sudden underperformance compared to the previous slight over-performances is unusual. It may be again a case of unclear interpretation of the intention of certain features, or, perhaps, the discrepancy in the data set represents a collection of relationships that exhibit a significant deviation from the norm for one or more of Chambers' new features. Regardless, we feel that this performance is similar enough to indicate that our technique is viable if not ideal, and provide us with a reasonable point of comparison for application to our own biographical corpus.

In the case of the Bio Corpus, we initially see a higher baseline, but classified results are very similar to the degree of accuracy seen on TimeBank, being within a few percentage points of those results. The magnitude of performance is lower, of course, to account for the higher baseline, but these results indicate that the pre-existing technique is still, ultimately, viable for classification in this new domain. Our highest reported performance on the previously published feature sets is 56.65%.

A more detailed analysis of the individual features confirms some of what was previously expected, and also some surprising new revelations. Classification without the inclusion of Aspect or the Aspect-linked features (Aspect agreement and the Aspect bigram) proved to cause a negligible decrease in performance in the Bio Corpus, while TimeBank in the same situation was penalized more severely. This is in line with what was expected, given the high class baseline.

Additional analysis, specifically of the individual improvements from Chambers' new features, proves that it is, however, possible to elicit some improvement in the results. Our highest reported accuracy, quite surprisingly, comes from the addition of only the preposition phrase feature over the Lapata feature set, which results in an almost 2% increase to 57.34%. Part-of-speech features actually hurt performance from the simpler feature set, but their exclusion in the full set does not otherwise elicit an improvement.

Future Work

While this is a start, much more work can be done in this area. First and foremost, we would like to expand on the size of the Bio Corpus, and be able to take the time to hand-annotate the event detection (replacing EVITA), as a better gold standard would help remove the uncertainty of our baselines for Aspect and Class, which could be high from a tendency of the automatic classification to favor a certain type of event.

We would also like to reevaluate the annotation of the event-pair relationships, since earlier work has shown it to be inherently difficult even for graduate student annotators. Ideally, annotation would be done by at least three trained annotators, and then subjected to an analysis of the agreement between them to determine our gold standard. It is possible, with only a single

annotator, that a specific class of relationship could be unintentionally favored and thus bias the results.

REFERENCES

Allen, J. (1984). Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154.

Bird, S., Loper, E., Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

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

Chambers, N. (2011). Email interview.

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

Chambers, N., Jurafsky, D. (2008). Unsupervised Learning of Narrative Event Chains. *ACL 2008*.

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

Klein, D., Manning, C. D. (2003). Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

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

Mani, I., Schiffman, B., Zhang, J. (2003). Inferring Temporal Ordering of Events in News. *ACL 2003*.

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.

Verhagen, M., Mani, I., Saurí, R., Knippen, R., Littman, J., Pustejovsky, J. (2005). Automating Temporal Annotation with TARSQI. *Demo Session. Proceedings of the ACL 2005.*

Wang, D. T., Plaisant, C., Quinn, A. J., Stanchak, R., Schneiderman, B. (2008). Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records. *SIGCHI 2008 Proceedings*.

BIOGRAPHICAL SKETCH

Lucian Silcox received his Bachelor degree in 2007 from the University of Texas Pan American with a double major in Computer Science and English Literature. He returned and received his Masters of Science in Computer Science in 2012, with a focus in natural language processing.

While pursuing his degree, he had the opportunity to intern at the University of Southern California's Institute for Creative Technology, where he was able to contribute to projects in emotional modeling and analogical reasoning. He currently lives in McAllen, Texas and is considering enrollment in a doctoral program.

Lucian Silcox

10054 N. Shary Rd

Mission, Tx 78573