INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Linguistic Indicators

for Language Understanding:

Using machine learning methods

to combine corpus-based indicators

for aspectual classification of clauses

Eric V. Siegel

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

Columbia University

1998

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

Abstract

Linguistic Indicators

for Language Understanding:

Using machine learning methods

to combine corpus-based indicators

for aspectual classification of clauses

Eric V. Siegel

Linguistics as a field has provided enormous insights that describe how the thoughts behind language are reflected by the structure of sentences. For example, one writes a paper *in* one week, but rides a bicycle *for* one hour. This illustrates how prepositions (*in* and *for*) correspond to the type of event. Specifically, *in* modifies a completed process, while *for* modifies an ongoing process. The area explored by this thesis is, how can we best put our understanding of linguistics to use in order to tap into the vast knowledge encoded in texts?

The ability to distinguish *stative* clauses, e.g., "*She resembles her mother.*" from *event* clauses, e.g., "*She ran down the street.*" is a fundamental component of natural language understanding. These two high-level categories correspond to primitive distinctions in many domains, including, for example, the distinctions between *diagnosis* and *procedure* in the medical domain. *Stativity* is the first of three high-level distinctions that compose the *aspectual class* of a clause. These distinctions in meaning have been well motivated by work in linguistics and natural language understanding.

Aspectual classification is a necessary component for applications that perform certain natural language interpretation, natural language generation, summarization, information retrieval, and machine translation tasks. This is because each of these applications requires the ability to reason about time.

In this thesis, I develop a system to perform aspectual classification with linguistically-based, numerical indicators. These *linguistic indicators* make use of an array of aspectual *markers*, each of which has an associated constraint on aspectual class. For example, only clauses that describe an *event* can appear with the *progressive* marker, e.g., *"I was eating breakfast."* Therefore, the category of a verb or phrase is reflected by a numerical indicator that measures how often it occurs in the progressive. The values for such linguistic indicators are computed automatically across corpora of text. We develop and evaluate fourteen indicators over unrestricted sets of verbs occurring across two corpora. Our analysis reveals a predictive value for several indicators that have not previously been conjectured to correlate with aspect in the linguistics literature.

Then, machine learning is used to combine multiple indicators in order to improve classification performance. The models automatically derived by learning are manually examined, revealing several linguistic insights regarding the indicators and their interactions. Three machine learning techniques are compared for this task: decision tree induction, a genetic algorithm, and log-linear regression.

We conclude that linguistic indicators successfully exploit linguistic insights to provide a much-needed method for aspectual classification. Future work will

extend this approach to other semantic distinctions in natural language.

Contents

List of Figures

Acknowledgements

I have thoroughly enjoyed and benefited from my career as a graduate student due to the kind support and guidance of my advisor, Kathleen R. McKeown. I am indebted to Kathy for teaching me explicitly how to find the salient issues of a research project; she has the gift of always knowing exactly what question to ask next. Kathy encouraged me to explore various areas, and ultimately allowed me great flexibility in selecting research topics; I am fortunate to have worked on the natural language problem that interests me most, applying machine learning methods, my other primary field of interest, to boot. Only because of Kathy's devotion did my dissertation became a *culminated event.*

The other four members of my thesis committee logged many hours over the last few years, providing valuable insights and feedback for my work. Judith L. Klavans was my primary linguistic consultant and a great source of enthusiasm and motivational energy. Conversations about the linguistic side of this work with Philip Resnik were also extremely stimulating. The two of them kept me on target by providing many linguistic observations to further support my work, as well as teaching me the standard linguistic protocols for testing and evaluation. John R. Kender served two important roles in my thesis work. First, over the last 6 years John has been an influential role model for me, guiding me through many aspects of teaching, research, and academia. Second, John took me great steps forward in assessing and formally defending semantic distinctions set forth by linguists; the design of Section 2.2 resulted from discussions with John. Finally,

Salvatore Stolfo was my primary machine learning consultant. He was like the Judith and Philip of learning; only due to his gracious efforts and inspired insights is the machine learning side of my work meaningful from the vantage of that field.

What can I say about Jacques Robin, my graduate officemate of 5 years? Jacques became a close friend and a great basketball coach. He showed me how to balance the untamed ambitions of a new graduate student with the pragmatics of actual execution, and he showed me how to balance a research life with a night life. Jacques is an inspired, hilarious spark of positive action. He gave me pep talks and companionship. He's better than a pet dog.

The remaining members of Columbia's natural language processing group were incredibly important for feedback on this work. Vasileios Hatzivassiloglou was my primary statistical consultant: he is an encyclopedia of mathematical models, and was a great source of perspective on my work. He helped formalize one interpretation of my results, discussed in Section 2.2.4. Also, Vasileios designed and implemented the clustering of verbs described in Section 6.2.4. James Shaw first pointed out to me that *have* is statively ambiguous, resulting in the work reported in Section 4.3. Other helpful members of the NLP group include or included Michael Elhadad, Eleazar Eskin, Pascale Fung, Hongyan Jing, Min-Yen Kan, Diane Litman, Rebecca Passennau, Dragomir Radev, Ruth Reeves, Nina Wacholder, and Tony Weida.

Many other people in my life also lended a helping hand. Alexander Day Chaffee, my friend from childhood days of hacking, helped me enormously with

the interpretation of numerical results, and even had a couple helpful linguistic insights. Alex knows what is interesting about life and what is fun about computers; he will always be the Dalang of my Gamelan. Three professors from my undergraduate days at Brandeis University were extremely formative in my development as a researcher, and put in a lot of time to help me grow: Harry Mairson, Richard Alterman, and James Pustejovsky. Rick in particular was a great undergraduate thesis advisor (language generation), and inspired me to ride my bike to work every day. Other important contributors include: Peter Angeline, Ken Church, Ivan A. Derzhanski, Larry Eshelman, Jussi Karlgren, John Koza, Nelson Minar, Andreas L. Prodromidis, Conor Ryan, Astro Teller, Andy Singleton, David Schaffer, and Dakai Wu. Thanks also to Walter Alden Tackett and Aviram Carmi for the use of their SGPC software.

Thank you, my parents, Lisa Schamberg and Andrew Siegel, who are the ultimate best you could want. I appreciate their diligent struggle to find the optimal frequency with which to ask, "Is it done yet?" My sister, Rachel Siegel, is the center of the universe. Also, I'd like to thank my issues for remaining repressed (Alexander D. Chaffee, personal communication).

This research is supported in part by the Columbia University Center for Advanced Technology in High Performance Computing and Communications in Healthcare (funded by the New York State Science and Technology Foundation), the Office of Naval Research under contract N00014-95-1-0745 and by the National Science Foundation under contract GER-90-24069.

Chapter 1

Introduction

"To do is to be." - Descartes

"To be is to do." - Alexander D. Chaffee

"Do-be-do-be-do." - Frank Sinatra

Many challenges for natural language processing require the *classification* of words or phrases as one of a small number of categories. For example, *word sense disambiguation* is the process of finding the meaning of an ambiguous word from its context, e.g., *river bank* versus *Federal bank*. A second example, *aspectual* classification, is the problem of mapping a clause (e.g., a simple sentence) to one of a small set of primitive categories in order to reason about time. For example, *events*, such as, *"You called your father,"* are distinguished from *states*, such as, *"You resemble your father."*

Aspectual classification is necessary for interpreting even the most simple narratives in natural language. This is because, in general, the sequential order of clauses are not enough to determine the underlying chronological order. For example, consider:

"*John entered the room* (**event**). *Mary stood up* (**event**)."

In this case, the first sentence describes an event that takes place before the event described by the second sentence. However, in,

"*John entered the room* (**event**). *Mary was seated behind the desk* (**state**)."

the second sentence describes a **state**, which begins before the event described by the first sentence. Aspectual classification is a necessary step towards automatically identifying relationships in time between sentences.

The ability to distinguish **stative** clauses, e.g., "*She resembles her mother.*" from **event** clauses, e.g., "*She ran down the street.*" is a fundamental component of natural language understanding. These two high-level categories correspond to primitive distinctions in many domains, including, for example, the distinctions between *diagnoses* and *procedure* in the medical domain, and between *analyses* and *activity* in the financial domain.

Stativity is the first of three high-level distinctions that compose the *aspectual class* of a clause. Events are further distinguished along two other dimensions. First, *completedness* determines whether an event reaches a culmination or completion point in time at which a new state is introduced. For example, "*I made*

a fire" is **culminated**, since a new state is introduced – something is made, whereas, *"I gazed at the sunset"* is **non-culminated**. Second, *atomicity*, distinguishes *atomic* (instantaneous) events, such as, *"She noticed the picture on the wall."* from *extended* events, such as, *"She ran to the store."* By dividing events along the second and third dimensions we derive four classes of events

There is an array of semantic entailments related to aspectual category that linguistically motivates each of these three particular semantic distinctions. For example, one such entailment pertains to prepositional phrases that denote the duration of a state or event. *"For an hour"* can denote the duration of a **non-culminated event**, as in,

"I gazed at the sunset **for an hour.**"

In this case, *an hour* is the duration of the *gazing* event. However, when applied to a **culminated event**, it denotes the duration of the resulting state, as in,

"I left the room **for an hour.**"

In this case, *an hour* is not the duration of the *leaving* event, but, rather, the duration of what resulted from *leaving*, i.e., being gone.

Such *aspectual entailments* also illustrate the value of automatically classifying clauses according to aspect; once the category of a clause has been identified, they support an array of inferences pertaining to time. These inferences are crucial for natural language understanding and generation applications such as machine translation, processing medical reports, summarization, and augmenting