An investigation into compositional features and feature merging for maximum entropy-based parse selection
Mullen, Anthony James

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date: 2002

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 04-03-2020
Chapter 1

Introduction

1.1 The emergent field of NLP

The use of computers to process and analyze natural language goes back almost as far as the existence of computers themselves. Indeed, computers evolved largely out of research into the language-related problem of cryptanalysis in World War II; already the frequencies and patterns underlying language were intimately tied to the development of computing machines. In 1949, Warren Weaver, director of the Natural Sciences Division of the Rockefeller Foundation, circulated a memorandum simply entitled “Translation” (Weaver, 1955) which suggested the possibility that languages themselves might be considered as encryption schemes, and that translating a message from one language to another be considered a task essentially similar to decryption. Thus the first pursuit of the field now known as natural language processing (NLP) was instigated: that of machine translation (MT).

Since then, the field has developed and changed along with the computers it is so closely linked to. The history of NLP is intertwined with the histories of information theory, computer science, and contemporary linguistics. It has been given new relevance in the last ten years by the maturation of the Internet and the increasing speeds of computers, two factors which make it possible and necessary to process very large amounts of natural language data in a reasonable amount of time.

In spite of a rich history, however, the path of NLP has not always been smooth. The problem of MT turned out to be far more complex than had been predicted, and early research attained dismal results, leading to a sharp drop-off in interest in the field for many years (ALPAC, 1966; Hutchins, 1986). Meanwhile, the field of linguistics underwent a revolution when Noam Chomsky introduced the notion of a generative grammar in Chomsky (1957). The idea that Chomsky espoused was that the syntactic properties of a language should be describable by a set of rules called a “grammar”. The set of sentences licensed by the gram-
mar should correspond with those sentences deemed well-formed based upon the
intuitions of native speakers of the language. While the idea of a grammar by
itself was not new, the formalization of its properties and the analysis of lan-
guage in such terms allowed for interesting questions to be investigated regard-
ning the human capacity for learning and using language, and also provided a frame-
work for investigating questions about the ability of computers to process human
language.

The linguistic paradigm shift brought about by Chomsky had two major con-
sequences. The most clearly positive was that those studying natural language
now had a mathematical framework in which to consider the structure of language
(Hopcroft and Ullman, 1979). Significantly, this framework of formal language
theory is tied crucially to computer science and complexity theory. The basic
ideas of generative grammar provided a powerful way to analyze and describe
language and are still fundamental in virtually all computational approaches to
syntax.

The other consequence of the Chomskyan approach was to subjectivize the
study of syntax. Whereas previous approaches to linguistics concentrated on the
empirical analysis of texts and transcriptions of utterances, with the intent of
deriving structure from actual instances of language in use, the Chomskyan ap-
proach preferred to look inward. The “observations” of Chomskyan linguistics are
fundamentally intuitive judgements on sentences. The assumption is that native
speakers’ intuitions are the sine qua non of grammaticality; the set of grammatical
sentences is defined simply as the set of sentences of a language which native
speakers feel to be grammatical. It is assumed in generative linguistics that these
native speaker intuitions provide a characteristic function for a set of utterances
which constitutes the language, although caveats are made that the “perfor-
man ce” of speakers will not necessarily reflect this (Chomsky, 1965). A corollary
of this subjectivization, incidentally, was the possibility of studying semantics as
a linguistic phenomenon. Previous empirical approaches to linguistics focused
entirely on the observable manifestations of language. The generative view held
that syntax was a related but separate linguistic phenomenon from semantics and
that the construction of meaningful utterances rested upon the interplay of the
underlying syntactic and semantic structures. As a result the study of linguistics
expanded its field of inquiry to questions of meaning previously considered to be
the domain of philosophy.

The basic tenets of generative grammar underlie most of mainstream con-
temporary linguistics. From a scientific standpoint, these assumptions and the
framework from which they arise are not without their critics\(^1\), and in the world
of practical NLP, they have sometimes been uneasy bedfellows with the statistical
approaches which have been increasingly influential in the field.

\(^1\) See also Nygve (1986) and Nygve (1996) for pointed critiques of the methodology of modern
linguistics on scientific grounds
1.1. THE EMERGENT FIELD OF NLP

1.1.1 Linguists versus statisticians?

Frederick Jelinek’s well known quote as head of the IBM speech group, “Anytime a linguist leaves the group the recognition rate goes up,” (reprinted in Jurafsky and Martin (2000)) reflected the feelings of many in the field of speech processing during the 1970’s and 80’s and into the 90’s. During this period, effective methods were evolving in the realm of speech recognition which relied almost entirely on statistical modeling of sound patterns and word sequences. Linguistics seemed to have little to offer to this branch of NLP. In the late 1980s, Jelinek and the research group at IBM turned their attention to the problem of MT, with the CANDIDE translation project (Berger et al., 1994). CANDIDE was pitted against SYSTRAN, the current state of the art of MT systems, and found to perform comparably. Among the statistical NLP community, this was encouraging, fostering hopes that strictly statistical models could eventually outperform hand-built “symbolic,” or linguistically motivated models.

On the other side, linguists found the pursuit of functioning NLP systems using statistical means to be increasingly vacuous in terms of the light it shed on the human language faculty.\(^2\) Furthermore, it seemed that practical advantages should be derivable from the generalizations and intuitions about language that humans have. Human language use, after all, is the standard against which any NLP application must ultimately be compared, and the generalizations which come naturally to people often require considerable effort to approximate with statistical methods.

1.1.2 Interdisciplinary approaches to NLP

Some of these disagreements underscored a divergence of motivation. NLP applications are, generally speaking, well-defined and domain specific. Seeking to solve such problems is not the same as seeking a deep understanding of the nature of language, although the two pursuits, it seems, should in principle be related. Given such a practical perspective, elegance of theoretical underpinnings is outweighed in importance by effectiveness in achieving results. For practitioners of NLP, it was and remains uncontroversial that statistical modeling techniques may be employed to aid the task, and that whatever insights from linguistics prove advantageous should be exploited as well. The question has only been how.

In fact, while empiricists and generative linguists are still in dispute about the correct way to proceed in the science of linguistics, the NLP community has developed into a relatively harmonious one. The combination of statistics and linguistic insight has proved to be quite fruitful. In the past decade, the inter-

\(^2\)Abney (1996) presents entertaining and compelling arguments against this viewpoint, citing the phenomena of language acquisition, language change, and language variation as several which benefit from stochastic analyses, as well as addressing issues related to individual adult native speakers, including, among others, that of syntactic disambiguation.
CHAPTER 1. INTRODUCTION

disciplinary nature of NLP has begun to cohere, as is evidenced in the approach taken by an increasing number of graduate and undergraduate textbooks, such as Charniak (1993), Manning and Schütze (1999), and Jurafsky and Martin (2000), all of which emphasise the interaction between statistical methods and linguistic ones. The relationship between the two approaches is increasingly seen to be a natural and mutually supportive one. Chanod et al. (1999) compares and contrasts “symbolic” (linguistic) and statistical approaches to NLP in methodology and technology, concluding that both approaches have unique strengths and weaknesses and calling for a concerted effort to bring the two approaches together into a “coherent set of methods, resources, and tools.” As work on practical NLP continues, it is increasingly obvious that there are realms in which statistics alone are insufficient (a common claim of symbolic linguists) and likewise, there are places where symbolic approaches, on their own, fail. Appreciating the ways in which the two subfields can benefit each other has become crucial to the continued development of the discipline of NLP.

1.2 The problem of ambiguity

In fact, the purely symbolic approach falls short of practical applicability in what might at first seem to be a rather surprising place: that of assigning grammatical structure to sentences, or parsing. The reason this may appear counterintuitive, of course, is that a grammar is itself a symbolic construction. The analysis of sentences in terms of a grammar is very much the realm of generative linguistics. However, when it is necessary to parse a sentence automatically, given no other knowledge than the grammar (as opposed to all the real-world knowledge available to a linguist, or indeed any native speaker), the thorny problem of ambiguity arises.

The term “ambiguity” in this thesis refers to grammatical or syntactic ambiguity, although different varieties of ambiguity are usually closely related to one another. Syntactic ambiguity is a formal property of certain grammars. Grammars which are ambiguous allow certain strings to be derived in several different ways, yielding more than one possible underlying grammatical structure.

A very simple example may be seen in a grammar such as the following:

\[
\begin{align*}
S & \rightarrow AB \\
A & \rightarrow ab \\
A & \rightarrow a \\
B & \rightarrow bc \\
B & \rightarrow c
\end{align*}
\]

With this grammar, the string \textit{abc} may be associated with two possible trees, as seen in figure 1.2. Neither, according only to the grammar, is any more or less legitimate. Both are licensed by the rules.

Such a simple example does not begin to give an idea of how much of a problem
1.2. THE PROBLEM OF AMBIGUITY

![Diagram](image.png)

Figure 1.1: A very simple instance of grammatical ambiguity.

ambiguity is, however. With natural language, ambiguity creeps in at all levels: lexical, semantic, and grammatical, and all of these varieties of ambiguity are interrelated. For example, a word which is lexically ambiguous between two parts of speech may yield entirely different grammatical structures for a sentence depending on which interpretation of the word is assumed. The wider variety of linguistic phenomena may be described by the grammar (including the variety of lexical nuances encoded in the lexicon, and, if semantics are included, the finer and more comprehensive the distinction between semantic concepts), the greater the chance that multiple structures may be assigned to a given utterance. A common rule of thumb is that as the coverage of a grammar increases—that is, its ability to grant a correct structure to a wider variety of possible sentences—so too does its inherent ambiguity. Furthermore, the ambiguity, as measured in terms of the number of possible parses for a sentence, has been shown to be, in the worst cases, worse than exponential in the length of the sentence (Carroll, 1994). This is discussed in greater detail in chapter 2.

The Alpino grammar of Dutch (Bouma, van Noord, and Malouf, 2001), which will be discussed in greater detail in chapter 7 of this thesis, provides some vivid examples of the problem. The worst case of ambiguity seen so far in Alpino occurs in the sentence:

Aan Charles Masterman, een collega uit de eerste ministeriele jaren, was die neiging al eerder opgevallen.

which yields a rather astounding 521,472 unique parses in the Alpino grammar. The average number of parses, with respect to sentence length, can be seen in figure 1.2. The sources of ambiguity are varied. Multiple possibilities for prepositional phrase attachment is a major factor, which humans often resolve through recourse to real-world knowledge. Words may be ambiguous with regard to their parts of speech or, with verbs, their transitivity, leading to a sort of domino effect of ambiguity, where each ambiguous element opens up new possibilities for whole structures. For humans, the ability to rule out simple absurdities immediately—such as the possibility that the word “to” earlier in this sentence might be a preposition—allows the space of possible interpretations to be drastically cut down. It is a considerably different case for an automatic parser which has recourse only to a grammar.
Figure 1.2: The average number of parses assigned by the Alpino grammar as a function of the length of sentences.
1.2. THE PROBLEM OF AMBIGUITY

This, then, is clearly a hurdle for computational parsing. The example above, while extreme, is, typical in the respect that its meaning is entirely clear to a human speaker of Dutch. While it is, of course, sometimes the case that utterances may have ambiguous interpretations for humans, the problem here lies in the massive amount of ambiguity which is generally not perceived by humans. In this thesis, ambiguity is used to refer to the proliferation of different structures which can be associated with a string given a grammar. In natural language, this does not necessarily mean that multiple meanings will be derivable, by a human, from a sentence. In the majority of cases, in fact, the grammar licenses many structures which humans would not perceive or find at all sensible.

The reason for this discrepancy is that humans have access to resources which are not available to a computational grammar of reasonable complexity. Syntactic decisions are made on the basis of deep semantic knowledge of the world. Perhaps this information could, in theory, be encoded in the form of a huge semantic grammar. The amount of knowledge generally considered to be “non-linguistic” which is employed to make these decisions, however, would suggest that the notion of a grammar would have to be extended to cover all of human cognition, rather than merely linguistic behavior. In short, humans’ ability to disambiguate points to a very close relationship between language and thought. Computers, unable to think, are thus at a considerable disadvantage.

1.2.1 Statistical parsing

Statistical methods are well suited to the task of selecting an object from a set of possible objects, each with different characteristics, given that a training corpus exists of previously selected objects in similar circumstances. This is not by any means specific to linguistic applications. Any such ranking or selection task (in fact a specific case of classification, with only one class) for which a characteristic corpus of data is available is an appropriate domain for statistical modeling. Natural language parsing provides an excellent example of a linguistic problem well-approached by statistical means (Charniak, 1997). It is reasonable to think of parsing in this way, as being a question of choosing a single parse from a number of possible parses. Based upon characteristics of each parse, the model should rank the parses according to the likelihood that they are correct, based upon similar instances seen in the training data. This, if certain issues are resolved correctly, allows the model to simulate the “real world” knowledge of human speakers to the extent that such knowledge is implied by the parsing choices made in a training corpus of humanly parsed sentences.

The issues I refer to are the following: firstly, the question of what characteristics of a parse are relevant to the task, and secondly, the question of what underlying statistical framework best allows the information from the data to be exploited in a model.

This thesis will be occupied largely with the first of these issues, having de-
cided upon the maximum entropy framework, described in detail in chapter 3, with regard to the second issue. The experiments presented in chapters 6 and 7 emphasize the importance of the feature sets used for modeling. Specifically, the work in those chapters is aimed at getting the best possible modeling out of a given set of features by eliminating noise within the set. This thesis describes a novel method of doing this which is demonstrated to be helpful in some cases. The cases in which it is not helpful prove to be of equal interest, however, and the reasons why it improves some models and fails to improve others are the subject of the concluding discussion.

1.2.2 Maximum entropy and NLP

The maximum entropy framework (maxent) is an approach to statistical modeling which is currently employed in a wide variety of statistical NLP contexts; in addition to its early uses in speech recognition (Jelinek, 1998), it has been used in MT applications (Berger, Della Pietra, and Della Pietra, 1996), text classification (Nigam, Lafferty, and McCallum, 1999), part-of-speech tagging (Ratnaparkhi, 1996) and dialog systems (Koeling, 2001), to name but a few. Its usefulness in syntactic disambiguation is also well-recognized and will be discussed in more detail in chapter 5.

1.3 Where does this work stand?

This thesis is concerned with several different instances of the particular task of statistical parse selection. The methods of evaluation used for the various models in chapters 6 and 7 are described in detail and performance as measured by those metrics is the primary goal. The deeper questions of language and parsing, as related to the mechanisms of human cognition, are not the subject of the present inquiry. NLP and linguistics are closely entwined, however, and it is worth pointing out that presenting the work herein requires maintaining a certain degree of balance between the two disciplines, and will furthermore involve a certain degree of taking-on-faith. A primary assumption of this work, for example, is that human language is structured according to a grammar and that retrieving the meaning of an utterance requires that a grammatical structure be assigned to the utterance. That is to say, it is assumed that parsing is a task worth doing. This is unlikely to be contentious to linguists, while in the world of application development for practical NLP, it may be dubious whether such a deep analysis is generally worthwhile. In any case it is an assumption rather than an assertion of the present work that grammatical structure is worth deriving, albeit an assumption shared by a large community of researchers on the subject. Certainly no endorsement is implied of theory-specific details of the grammatical frameworks dealt with in the present work. Likewise, it is assumed that the
evaluation metrics used are meaningful. Assessing the goodness or badness of a parse structure is not a straightforward matter, and considerable thought has been put into arriving at useful metrics for doing so. Nevertheless, these metrics are not perfect, and are well acknowledged as being only an approximation to what is really intended, namely human speaker preference. More will be said on this issue in chapters 6 and 7.

1.4 The structure of this thesis

Chapter 2 introduces and describes the related tasks of parsing and parse selection. Basic assumptions of grammatical structure are illustrated and standard algorithms for parsing sentences are described.

Chapter 3 describes the statistical framework of maximum entropy. This is a general purpose framework for statistical modeling with many advantages in the field of NLP. The problem of overfitting and noise is described.

Chapter 4 follows upon the previous two chapters, wrapping up a survey of background work by describing applications of maximum-entropy style modeling to the task of parsing and grammar modeling.

Chapter 5 describes the approach to statistical parse selection using the maximum entropy method which will be used in the current thesis. This approach follows similar lines as much of the state-of-the-art work in the field, and is guided by the same intuitions. However, certain qualities of the statistical features used are emphasized in such a way as to lay the groundwork for a novel approach to noise reduction, feature merging, which is also introduced in this chapter. A toy example is presented by way of motivation for the experiments conducted in subsequent chapters.

Chapter 6 describes experiments done in parse selection with parses generated by the Alvey Natural Language Tools grammar. The problem of overfitting due to noise is illustrated clearly, and the feature merging technique introduced in chapter 5 is employed to help to counter the effects of the noise. Experiments are described which demonstrate the effectiveness of feature merging in conjunction with the more conventional noise reduction approach of a frequency-based feature cutoff, in certain cases. The work in this chapter may be thought of as a preliminary study of the potential of the feature merging technique. It is shown that models which suffer from certain deficits may be significantly improved using a combination of the two noise reduction methods. Other models, however, whose performance is better to begin with, show no improvement using the merging technique.

Chapter 7 takes the insights gained in chapter 6 and attempts apply them to the real world parsing problem of the Alpino grammar environment for Dutch. The implementational details are different, but the underlying idea of feature merging is identical. Certain problems arising in the evaluation of the results
from chapter 6 are avoided, and the conclusions from that chapter are questioned and new results presented in a clearer and more consistent way. Initial results on held-out data appear to suggest an improvement gained by employing the merging technique, but this is not borne out in subsequent tests. Several varieties of feature set are used, and the results of all tests fail to indicate any improvement to be gained from the merging process.

In chapter 8 I conclude the inquiry and re-iterate certain key points about the feature merging technique and its applicability. The discussion focuses on the difference between the cases in which merging might be of help and the cases in which it appears not to be useful. I also discuss possible avenues for further research in the field.