Exploring Higher Order Dependency Parsers

With special attention on feature engineering with Higher Order Parsing.

Masters Thesis

Sumbitted in the partiall fulfilment of requirements for the Degree of Masters Degree in Human Language Science and Technolgy.

Pranava Swaroop Madhyastha

European Masters Program in Language and Communication Technologies.

Supervisor:	Prof. Michael Rosner
Co-Supervisor:	RNDr. Daniel Zeman

This masters thesis is a part of European Masters Program in Human Language and Communication Techonologies Program.

The work in this thesis has been carried out at:

Department of Human Language Science and Technology
Department of Intelligent Computer Systems
University of Malta

&
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague

Abstract

Parsing is one of the most important steps in understanding of natural languages. In this thesis, we focus on the dependency grammar formalism since, the core concepts of dependency grammar, namely the relational view of head and modifier asymmetry, has proven useful for diverse set of languages, especially accounting for the explanation of word order and relation between surface structure and meaning. Most of the recent efficient algorithms for dependency parsing work by factoring the dependency trees. In most of these approaches, the parser loses much of the contextual information during the process of factorization. In this thesis we investigate how features (syntacto-semantic) affect the higher order discriminative learning methods for dependency parsing. We will show that linguistic features in most cases provide a significant improvement in the parsing accuracy.

We start by presenting a survey on several discriminative learning methods for graph based statistical dependency parsers and explain the concept of higher order that is the generalization of the work done by [Koo & Collins 2010] and [McDonald et al. 2006]. This leads us to the core of the thesis - feature engineering in higher order dependency parsers. Here, we experiment with several syntacto-semantic features then try to explain the theoretical foundation of these features. The experiments are done on two diverse languages - English and Czech, we have compared the several results obtained with different parsing algorithms.

Keywords: Dependency Parsing, Discriminative Learning, Higher order, Semantic features.

Contents

Introduction

1

One of the major challenges in the field of computational linguistics is to transform text from the native natural language representation to representations which can be fed as an input to the computer. The computer would then be using this to perform various tasks. The transformation of representations from natural language to well-defined formal languages involves several layers of processing. Amongst these, parsing is one of the most important and the most difficult of them.

Parsing of a natural language can be defined as the process of mapping sentences in the natural language to their syntactic representations. Parsing also lays an important foundation for understanding the natural language syntax and semantics. Recently, statistical parsing has taken precedence over other forms of parsing due to its highly efficient parsing capabilities ([Marcus et al. 1993]). While parsing accuracy of these parsers are mostly rising, this is still not enough for integration with the practically implementable natural language processing applications and hence there is a pressing need for better accuracy [Merlo et al. 2011]. This is also due to the highly ambiguous nature of the natural language. High accuracy natural language parsing would be very useful for modern NLP applications which include machine translation, question answering systems, information extraction, text summarization, semantic role labeling, etc..

The syntactic structure of the natural language is formalized into a certain syntactic representation, this is also known as the grammatical formalisms. There are several syntactic representations which are used in computational linguistics. In this thesis, we would be focusing on one of the most important representations which is based on the notion of dependency [Tesnière 1959]. This formalism is now formally known as the "Dependency Grammar" (DG) Framework. Also, we would be concerned with a particular type of parsing for the DG called the data-driven discriminative graph based dependency parsing, also known more frequently as graph based dependency parsing formalism which is a type of statistical parsing.The rest of the thesis will set the very basic premise of the thesis.

1.1 DEPENDENCY GRAMMAR: DEFINITION AND CURRENT STATUS

A dependency tree can be defined in the most basic way as a directed acyclic graph in which all the words in the given sentence are connected together by grammatical relations. For example, the subject and object depend on the main verb; adjectives depend on the nouns that they modify; etc.. In each pair of connected words, one is called the 'dependent' which is basically a modifier and the other is called the 'head'. That is, the modifier modifies the head. Simply, an analysis based on DG can be explained as a tree where, each token in the sentence is a node in the tree, and each arc connects a head to its modifier. A more detailed discussion on the dependency grammar will be made in the next chapter. DG is an increasingly important grammar representation in modern computational linguistics. It is particularly well-suited for languages with approximately free word order [Covington 2001]. Also, dependency representations are emerging as the standard for comparing the result of syntactic analysis across different grammar formalisms and parsing approaches. In a way, the DG formalizes the syntactic structure as a directed tree of dependencies. The classical phrase-structure [Chomsky 1956] models have been of less help in exploring the joint 'syntactic and semantic' phenomena, especially with a cross-linguistic perspective. [Mel'čuk et al. 1987] and [Covington 2001] claimed that one of the advantages of DG over approaches based on phrase based or constituent structures is that it allows for a more adequate treatment of languages with variable word order, where discontinuous syntactic constructions are more common than in languages like English. Also note, dependency links are close to the description of semantic relationships needed for the next stage of interpretation in the linguistic hierarchy.

There are two dominant and mostly studied approaches to dependency parsing: *graph-based and transition-based*, where graph-based parsing is understood to be slower but exhaustive (global optima based approach), and often more accurate.

DGs have been at the forefront of computational linguistics since last two decades. This can be seen by its application to functional description of grammar [Sgall 1984], possibilities of extracting rich lexical information from corpora [Bangalore et al. 2009], applications related to semantic graphs [Marneffe et al. 2007] and adaptability to various languages with the same formalism [Bourdon et al. 1998].

1.2 DEPENDENCY PARSING: STATUS QUO

Highly efficient parsers have made DG to be one of the most explored grammar formalisms in the last decade [Merlo et al. 2011]. One of the major hurdles in understanding natural languages is mostly concerned with producing an optimized natural language system. Implementations of efficient grammar formalisms form one of the basic components of these systems. Current data driven dependency parsing formalisms can be divided into three different types:

- Local-and-greedy transition based parsers (e.g., MALTPARSER and similar parsers [Nilsson et al. 2006], [Yamada & Matsumoto 2003]),

- Globally optimized graph-based parsers which are also known as discriminative graph based dependency parsers. (e.g., MSTPARSER [McDonald et al. 2005] ; [Koo & Collins 2010] and [Carreras et al. 2006]), and

- Hybrid systems (e.g., ([Sagae & Lavie 2006] and [Nivre & McDonald 2008])), which combine the output of various parsers into a new and improved parse.

Transition based parsers basically scan the input from left to right. They usually have linear complexity and mostly make use of a big list of features. Most of their their decisions are local. Some of the transition based parsers have the restriction of sticking to the 'left to right' direction [Nilsson et al. 2006]. [Nilsson et al. 2006] also states that the transition based parsers have $O(n)$ complexity, that is, it has a 'linear complexity'. Also note, even if they can use the big list of features, which can basically include the rich structural information, it is basically restricted, as only the next two or three lexemes are available to the parser. This implies that it has a very small look-ahead window, and hence, it is right to predict the relatively less rich contextual information. This usually results in error propagation and relatively bad performance on root and long distance dependencies when compared to graph based or discriminative dependency parsers [McDonald & Nivre 2007].

The transition based parsers are also known to an extent as 'history-based models'. History-based models basically incorporate arbitrary information from the prediction history to estimate the possible decisions at each choice point in its search space. Transition based parsers make an independence assumption, which actually is not a theoretical necessity, in order to decrease the search space. Hence, performing a applying a local optima approach. These approaches in general involve a concrete search for the minimum loss structure. This would help put an upper-bound on the inference time almost for every sentence structure. Basically in a 'transition-based parser' the history based model is applied at every state transition. These are known to apply deterministic parsing time as they employ greedy inference.

The graph based dependency parsers on the other hand, have a better contextual information. They usually perform an 'exhaustive' search over all the probable parse trees for a given sentence, and hence are globally optimized. Once they do an 'exhaustive' search they sum all the possible tree structures (this will be discussed in length in the following chapters) to find the best scoring tree for a given sentence. But then, an increase in the feature sets actually makes it a hard problem, hence the feature sets are mostly restricted to the single edges - (which is the first order parsing model) or edge-pairs (second order parsers - ([McDonald et al. 2006]; [Carreras 2007]) or edge-triplets (third order [Koo & Collins 2010]). There have been efforts on incorporating arbitrary tree-based features, but these adversely affect the overall complexity of the parser. These models have at least $O(n^3)$ complexity. In contrast, transition-based parsers have the opportunity to extract non-local features by examining the stack, but they are intractable when they exploit such non-local features.

In this thesis, we will concentrate on this graph-based family of parsers and explore mostly the higher order parsers - second and higher.

1.3 RESEARCH OBJECTIVE

The major focus of the research presented in this thesis is to investigate the effects of semantic and morphological features on the discriminative data driven dependency parsing. Especially, we will be concentrating on the graph based dependency parsing. One powerful aspect of discriminative models is their ability to incorporate rich sets of highly dependent features. The question that we seek an answer for in this thesis is: *"How do these features effect parsing?"*. [Bikel 2004a] has done a detailed analysis on each class of distribution of features for generative models. But a similar analysis seems to be missing for discriminative models, especially the graph based parsing models. This thesis is a step ahead into a similar analysis. In this thesis we would be concerned about the affects of features on the learning and prediction of the graphical dependency parsing algorithm, especially using averaged perceptron. More details about the same are discussed in detail in the next sections. We come to a conclusion that adding the semantic features improves the parsing accuracy and these are also dependent on the type of algorithms and the amount of information about the structure exploited by algorithms.

Premise

2

In this section, we review the background concepts of Dependency Grammar and dependency parsing. We will then provide a succinct description of the research orientation adopted. We will also present a holistic view of the graph based dependency parsing models.

2.1 WHY PARSING?

Natural languages like English are hard to define in exact terms and is ambiguous in many situations, while a formal language is mostly well defined and is less ambiguous. A natural language has often evolved during thousands of years and yet it continues to evolve. This makes it impossible to state an exact definition at a given time. It is also hard to draw boundaries between natural languages, and whether a particular language is counted as an independent language is usually dependent on historical events and culture, seldom only on the linguistic criteria. These properties not only make natural language processing a challenging task but also a very interesting research topic. Especially with the increasing use of information technology in combination with natural languages. Many computer applications that involve natural languages like machine translation, question answering and information extraction are dependent on modeling natural language in an easier representation. Moreover, these applications usually have to deal with unrestricted text, including grammatically correct text, ungrammatical text and foreign expressions.

Thus, parsing of natural languages can be seen as the process of mapping an input string or a sentence to its syntactic representation. We assume that every sentence in the given set of sentences (or in other words the corpus) has a single correct analysis which the speakers of the language generally agree that this analysis is preferable. We do not necessarily assume that a formal grammar defines the relationship between sentences and their preferred interpretations. [Nilsson et al. 2006] in his paper uses the concept of text analysis to characterize this problem that can only be evaluated with respect to the empirical evidence of a language text.

Several attempts have been made to formalize the grammar of the language over a long period. The first records of such an attempt dates back to 400 BC, when Panini described and formalized the Sanskrit grammar. The first computational study of grammar could be dated back to the early 1950s with the seminal work of CFG [Chomsky 1956]. Since then, there have been a very large set of grammatical formalisms, which have existed and which have been used and implemented in several domain of computational linguistics. These plethora of grammar formalisms help the objective of parsing in several ways for different applications. In the next section we would solely concentrate on the Dependency Grammar formalism that forms the backbone of the formal structure in our work.

2.2 THE DEPENDENCY GRAMMAR

Syntactic representations based on word-to-word dependencies have a long tradition in general linguistics. The basic assumptions behind the notion of dependencies are summarized in the following sentences from the seminal work of [Tesnière 1959]: (translated from French verbatim)

"The sentence is an organized whole; its constituent parts are the words. Every word that functions as part of a sentence is no longer isolated as in the dictionary: the mind perceives connections between the word and its neighbors; the totality of these connections forms the scaffolding of the sentence. The structural connections establish relations of dependency among the words. Each such connection in principle links a superior term and an inferior term. The superior term receives the name governor; the inferior term receives the name dependent"

The very basic dependency structure can be viewed as shown in 2.1.

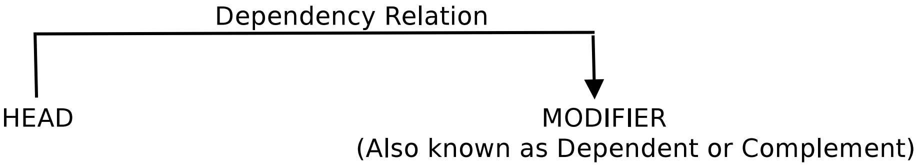

Figure 2.1 Basic dependency structure - A head and a modifier

As we have explained before, we usually represent the dependency relations among the words of a sentence as a graph. A dependency representation is a labeled directed graph, where the nodes are the lexical items and the labeled arcs represent dependency relations from heads to dependents.

It is these binary dependents - *the head and the modifier* that play a major role in the structure. Let us have a quick glance at a dependency structure in 2.2.

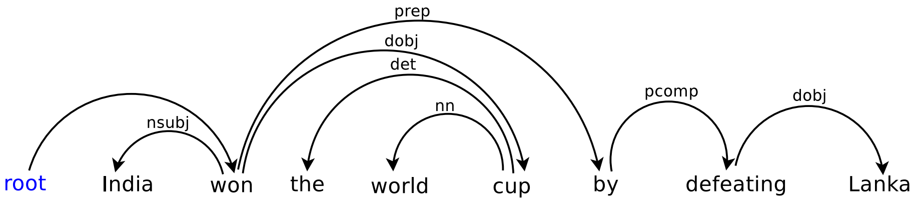

Figure 2.2 A simple example;please note that we are adding an extra root here

The dependency structure for a sentence **x** with words $\vec{x} \in [x_1 ... x_n]$, is the directed graph on the set of positions of $\vec{x}$ that contains an edge from i to j if and only if x_j depends on x_i . In this way, the dependency structures can capture information about certain aspects of the linguistic structure of a sentence. This notion allows us to express linguistic concepts as structural constraints on graphs. In practice, the dependencies are usually required to form a well defined and well formed tree. There are well formed rules for the retention of the tree order for a dependency tree.

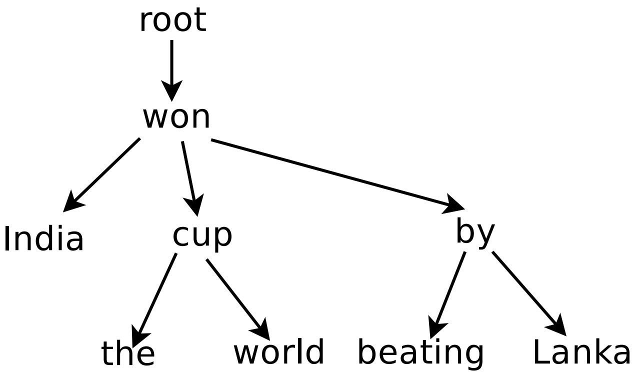

Figure 2.3 The directed graph for previous example

2.3 PROJECTIVITY AND NON-PROJECTIVITY CONSTRAINTS

Projectivity is concerned with the restriction of the span of the dependency relation. It requires each dependency subtree to cover a contiguous region of the sentence, hence, making sure that there is no crossing in dependency relations. Or in other words, the dependency spans don't cross each other. This is a very important constraint and it makes sure that there is no shuffle in the word order, which acts as a boon for the dependency algorithms, since, it adds a constraint. These have been strongly exploited with dependency parsers like [Eisner & Satta 1999].

However, there are many languages where the dependency subtree may be spread out over a discontinuous region of a sentence, which usually results in the crossing of the spans. [Kuboň et al. 1998] have mentioned that such representations mostly occur due to the linguistic phenomena such as extraction of entailment, topicalization and extrapolation. These are particularly common in languages with flexible word-order. Unfortunately, dependency parsing of non-projective structures using graph-based dependency parsing is found to be NP-Complete [McDonald & Nivre 2007]. Let us investigate this quickly by venturing into the two diagrams here. 2.4 represents a projective dependency structure. 2.5 represents a non-projective dependency structure.

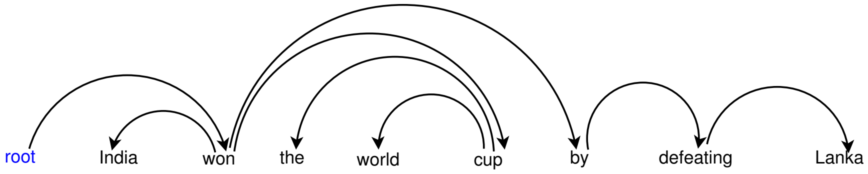

Figure 2.4 Projective Structure: none of the edges are crossing each other.

For English, projective trees are sufficient to analyze most sentence types, this is also stated in [Sleator & Temperley 1993]. However, there are certain examples in which a non-projective tree is preferable. We consider one of those examples here in the next sentence. But, in general, most of the free-word-order based languages have non-projective structures more often than for English.

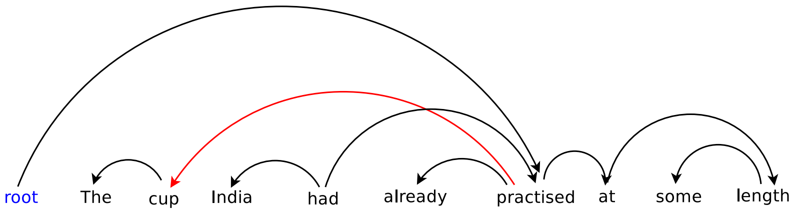

Figure 2.5 Non-Projective Structure: the red edge is crossing in the above dependency structure. In this specific case the establishment clause and the object it modifies are separated by an adverb. There is no way to draw the dependency tree for this sentence in the plane without crossing edges. Languages with more flexible wordorder non-projectivity is more pronounced.

In this thesis all our algorithms are following the above definition of projectivity. Also note, the parsing algorithms that we would discuss ahead in the thesis correspond to the the projective structures.

2.4 DISCRIMINATIVE GRAPH BASED DEPENDENCY PARSING

Let us now shift our focus to the type of dependency parsing upon which we will concentrate in this thesis. Graph based parsers are parsers which implement a discriminative learning technique. In this sub-section we will try to explain the theory of discriminative parsing.

2.5 NOTATIONAL CONVENTIONS

Let us consider that a dependency parser gets as an input a sentence **x** of n tokens and outputs a labeled dependency tree **y**. A labeled dependency would be the triplet $< h, m, l >$, where the index of the head token is represented as $h \in [0...n]$, the index of the modifier token is represented as $m \in [1...n]$ and the label for each dependency pair (h, m) is represented as $l \in [1...L]$ (here, **L** is the set of all possible dependency labels in the given dataset). Also note, the total number of sentences in the corpus is assumed to be χ and the number of possible trees for the sentences in the whole corpus is assumed to be γ.

The head of the sentence is assigned a value $h = 0$, in this thesis, we represent it by a special *root* symbol, which is as represented in the previous figures. $D(x)$ represents all possible dependencies and the set $(\vec{x})$ represents all possible dependency structures for a sentence **x**. $(\vec{x})$ for projective parsing algorithms and non-projective parsing algorithms differ.

Before we proceed ahead to define the parsing algorithms, let us first explore the discriminative modeling.

$$\mathbf{y}^*(\vec{x}; \vec{w}) =_{y \in \gamma(\mathbf{x})} \vec{w} \cdot \phi(\vec{x}, y) \tag{2.1}$$

In the above equation: please note, χ is the set of all the sentences in the corpus and γ. Here, $\phi(\vec{x}, y)$ produces a d-dimensional (where the d is dimensionality of $(\chi, \gamma)) \rightarrow \mathbf{R}^d$) vector representation of the event that dependency tree y is assigned to sentence x. Each dimension in $\phi(\vec{x}, y)$ is a feature that measures some quantifiable aspect of x and y. Hence we call $\phi(\vec{x}, y)$ as the feature vectors. The parameter vector $\vec{w}$ contains d weights corresponding to the d separate features; these parameters are learned on a training corpus of examples.

The maximization is performed over the set (**x**), the size of this set increases exponentially with the length of the sentence. This makes the enumeration intractable. To take care of this situation we factor the dependencies in the following way.

2.6 FACTORING STRUCTURES

Factorization constrains the feature representation so that each feature is only sensitive to a limited region of y. Essentially, the factorization breaks each structure into sets of parts, which are local substructures of y with well-defined interactions.

Consider, for a given sentence **x**, with parameter vector w, and parts p, we have a modified and reduced version of the equation mentioned in the last subsection:

$$\mathbf{y}^*(\vec{x}; \vec{w}) =_{y \in \gamma(\mathbf{x})} \sum_{p \in y} \vec{w} \cdot \phi(\vec{x}, p) \tag{2.2}$$

Let us consider an example to illustrate this, the simplest type of factorization is the generic factorization, that is, the first order factorization, which is also implemented by [McDonald et al. 2006]. In this case a tree y is broken into n component dependencies. If this is the case, the equation would then transform to:

$$\mathbf{y}^*(\vec{x};\vec{w}) =_{y\in\gamma(\mathbf{x})} \sum_{(h,m)\in y} \vec{w}\cdot\phi(\vec{x},h,m) \tag{2.3}$$

since - (h,m) would represent the head and modifier indices of a dependency in y. We can then apply dynamic programming algorithms efficiently to solve the parsing problem. This is well described in [Eisner 2000]. We will briefly explore this shortly.

The leftmost component i.e., the function is referred to as the factorization. This method is used to decompose the tree into parts. The $\phi()$ corresponds to the feature functions. In theory, every component can result in an opportunity for improvement, we would restrict ourselves to the feature section.

Now, if the above problem also considers non-projective structures, then a simple dynamic-programming would not suffice. But it could, still be efficiently solved by using directed maximum spanning tree algorithms as shown by [McDonald et al. 2005].

Consider a case where the dependency trees are factored into larger parts, i.e., scoring groups of two or more neighboring dependencies with a shared head. This is known as higher-order factorizations, and parsers which implement this are known as higher order dependency parsers. This forms one of the seminal parts of this thesis. We shall explore later the different approaches to the higher order dependency parsing and the advantages and disadvantages of the higher order dependency parsing.

Let us now switch our focus back to the previous mentioned equations, especially the concept of estimating the parameters .

2.6.1 *Parameter Estimation using Structured Perceptron*

A parameter estimation problem is usually formulated as an optimization problem. This is mostly because of different optimization criteria and also several possible parameterizations, a given problem can be solved in many ways. In this thesis we use one of the simplest parameter estimation methods - the structured perceptron algorithm, which is the generalization of the perceptron algorithm.

This algorithm was introduced by [Collins 2002] in his seminal work of discriminative training models for Hidden-Markov-Models (HMM). The perceptron is one of the easiest and the simplest parameter estimation algorithms.

The averaged perceptron begins with a parameter vector, which is initialized to 0. It is then proceeds in a series of T iterations, which are basically divided into a series of estimations. Each estimation step involves selecting a **random example** from the training set, parsing that example, and checking the parsers prediction against the standard structure. If the structures differ, then the parameters $\vec{w}$ are updated with the difference between the feature vectors of the gold standard and model prediction. The output of the algorithm is not the final parameter vector, but the average of all parameter vectors across every trial in the training run. Henceforth, it is also called the averaged perceptron for parameter estimation. The parameters are basically updated in the case of a mistake. We can see the pseudocode of the algorithm in the here:

Algorithm 1 Pseudocode for average perceptron algorithm: In this algorithm $\vec{w}$ is the normal parameters and $\vec{v}$ is the summed parameters. Also note that the resultant output as described here is $\vec{v}$.

```
Input: Training Data = $(\vec{x}, \vec{y})$ where $i \in [1...n]$
$\vec{w} = 0$
$\vec{v} = 0$
for $t = 1 \to T$ do
  for $j = 1 \to n$ do
    i = Random[1, n]
    $y' = y''(x_i; \vec{w})$
    if $y' \neq y_i$ then
      $\vec{w} = \vec{w} + \phi(x_i, y_i) - \phi(x_i, y')$
    end if
    $\vec{v} = \vec{v} + \vec{w}$
  end for
end for
$\vec{v} = \vec{v}/T_n$
```

The difference computed in the algorithm, i.e., $\vec{w} = \vec{w} + \phi(x_i, y_i) - \phi(x_i, y')$, has an additional property: if the model prediction y′ is mostly correct, then only a few of its parts will differ from the parts in the gold standard y_i. Thus, the update performed on the parameter vector will only modify features pertaining to incorrect or missing parts.

The averaging of parameter vectors is crucial for obtaining best results with the perceptron algorithm. As pointed out by [Carreras 2007] the actual perceptron parameters yield only mediocre parsing performance, while the averaged parameters v resulting from the same run are of much higher quality. This seems to generate desirable performance for our parsers. We are only sticking to this form of parameter estimation.

We now seem to have the necessary background to understand higher order discriminative parsing models.

What we have described above is just a formalization of the tree factored into parts, that is, we have described the score of the tree to be the sum of the edge scores. A detailed account on the factorization is explained in [McDonald et al. 2005]. In the process of factorization, the whole problem of finding the dependency tree of a particular sentence has been reduced to the problem of finding maximum spanning trees. In this section we restrict the definition to the more refined, projective dependency trees.

[Koo & Collins 2010] defines order of a part as *the number of dependencies a part contains*. In the following sub-sections we shall see explore the different projective algorithms and then we would concentrate on the existing feature structure and the proposed changes in the feature structure.

2.7 PARSING ALGORITHMS

In this section we briefly describe the current research on graph based dependency algorithms. We have briefly described the algorithms which are one of the most essential parts for the justification of our hypothesis.

Before we go further with the description of the dynamic programming structures, let us understand the terminology which is generally accepted:

- Simple Dependency: The simple dependency as we have been defining since the beginning, is made of a head (h) and a modifier (m) relationship.

Figure 2.6 The basic dependency structure.

- Sibling Structure: A sibling is defined as the relation where the modifiers share the same head. We consider here one as the main modifier (m), while the other as the sibling (s).

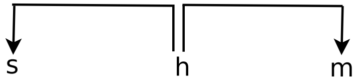

Figure 2.7 Sibling structure.

- Grandchild Structure: In case of a grandchild structure, the head has a parent, that is the grandparent (g). The structure is depicted in the figure here.

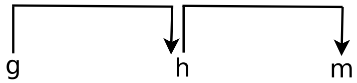

Figure 2.8 Grandchild Structure

- Grandsibling Structure: Here, the sibling structure defined above is headed by a grandparent. Hence, forth a grand-sibling structure.

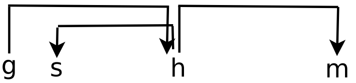

Figure 2.9 Grandsibling Structure

Let us now understand the basic dynamic programming structures. The algorithms in detail would not be mentioned, but the relevant reference would be made for each of the algorithmic formulation.

First-Order Factored Parsing Algorithm

One of the earliest implementation of discriminative dependency parsing was introduced in the seminal work by [Eisner 2000] who used dynamic programming for first order parsing. This laid the foundation to other parsing algorithms in the area of graph based dependency parsing.

The most important part of the algorithm is that it has two main components - *the complete span* and *the incomplete span*. The complete span on one hand consists of a headword and its modifiers, while the incomplete span consists of the region between the head and the modifier.

A slightly modified version of CKY [Eisner 2000] chart parsing algorithm would generate and represent the dependency trees in less than $O(n^5)$ (which is the standard time complexity of CKY algorithm) time to create. However, [Eisner & Smith 2010 (Chapter 8)] parses left and right dependents of a word independently and then combines them at a later stage. This reduces the time complexity from the standard - $O(n^5)$ to $O(n^3)$ as each derivation is defined by fixed boundaries of a 'span' (which is two) and 'a split point'.

We can think about a complete span as a **'half-constituent'** of a dependency tree part. This half-constituent is headed by a head '*h*' and is modified by a modifier '*m*'. Similarly, an incomplete span can be thought of as a **'partial half-constituent'**, because, this is extended by adding modifiers to m.

Let us consider a complete span as $C_{h,e}$ where h and e are the indices of the span's headword and endpoint. An incomplete span may then be written as $I_{h,m}$, where m is the modifier of h. As we have seen above, with the Eisner's independent combination strategy, again, each span is created by recursive combination of smaller spans. An incomplete span is constructed from a pair of complete spans, while a complete span is created by combining the incomplete span with the other half of the constituent.

The above process is a recursive process. Also in the figure, the span combines at a point m (in (1)) and at a point r (in (2)) is the free index that must be enumerated to find the optimal construction. This is also split point. The point of concatenation is found to infer the optimal construction. This is exactly accomplished by modifying the CKY parsing techniques as mentioned in [Cocke & Schwartz 1970] and [Kasami 1965].

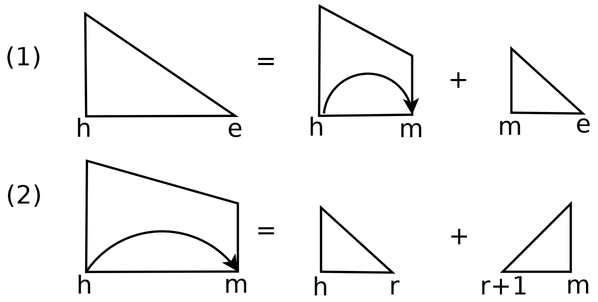

Figure 2.10 First-Order dynamic programming structures
Kindly note here that complete spans are triangles and incomplete spans are trapezoids.

2.7.1 *Second-Order Factored Parsing Algorithm*

In a second order factored algorithm a part contains 2 dependencies. The implementation of second order parsing algorithms have been majorly done in the below mentioned ways:

- ***Sibling factorization*** - This was introduced by [McDonald et al. 2006] where the dynamic programming structures were modified to explore the possibility of extending the parts to include the sibling information. That is two words with a shared head word. In this case, a sibling information is a triplet $< h, m, s >$. Extending from the previous algorithm, (h, m) and (h, s) are dependencies and s and m are successive modifiers to the same side of h. For this case, the dynamic programming structure has been augmented to include an extra structure: sibling spans. Sibling span represents the region between successive modifiers and of a head. Let us consider a sibling span as $S_{s,m}$, here s and m are the successive modifiers involved in the relationship.

  In this case, the incomplete spans are constructed in a completely different way as opposed the earlier case. Here, the parser combines incomplete span, that represents the innermost dependency with a sibling span. Even in this case, each derivation is still defined by a span and split point only. Hence, even here the parser requires $O(n^3)$ time.

- ***Grandchild factorization*** - The grandchild factorization was introduced by [Carreras 2007] in which parser tries to exploit the grandchild parts, that is, the part includes the children of head and modifier. So, in this case, a grandchild information is a triplet $< h, m, c >$. Extending from the same first order factorization, (h, m) and (m, c) are now the dependencies. Again, for this case, the dynamic programming structure is modified to include the identity of the *outermost* modifier of the head of the complete span. These also make use of the aforementioned sib-

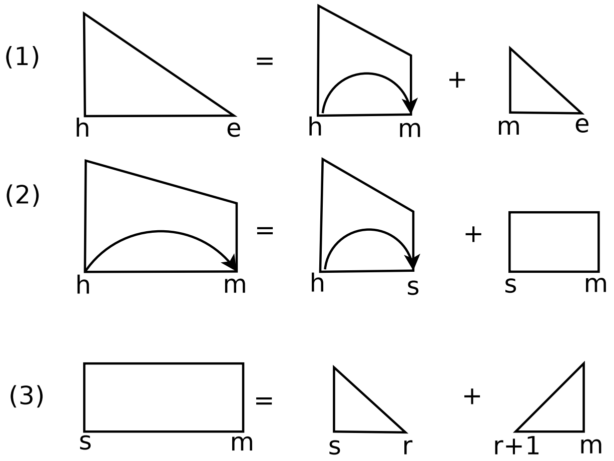

Figure 2.11 Second order dynamic programming structure
This shows the sibling spans

ling parts - (h, m, s) but then s and c are effectively independent of each other and hence the algorithm is optimized to deal with each index separately. Here, please note that the grand child relation changes the parsing algorithm now, in terms of the computational complexity. That is the complexity increases from $O(n^3)$ to $O(n^4)$.

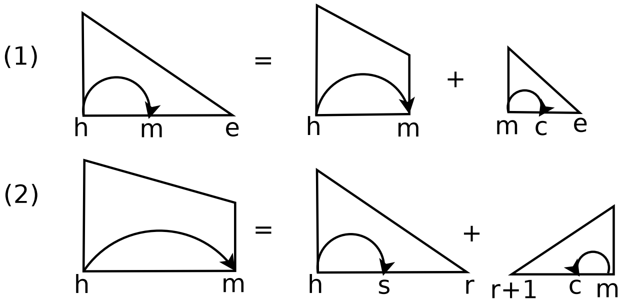

Figure 2.12 Second order dynamic programming structures
Grandchild spans

2.7.2 *Third-Order Factored Parsing Algorithm*

Third order parsing algorithms was introduced by [Koo & Collins 2010] which basically extends the above approaches. This is mostly by augmenting the grand-parent index. [Koo & Collins 2010] remarks that the efficiency of the third order algorithms is due to a fundamental asymmetry in the structure of a directed tree, i.e., a head can have any number of modifiers but a modifier always has a single head. So in a way, this exploits the structural asymmetry. [Koo & Collins 2010] specifically divides the parsing algorithms into three different models, the important two are mentioned below, which we have tried to experiment on:

- *Include all grandchildren* In [Koo & Collins 2010]'s structures a grandchild is a part contains the information of the triplet $< g, h, m >$, where (g, h) and (h, m) are dependencies. For this both complete and incomplete spans are augmented with g-spans. Hence, in other words, it basically represents the same first-order algorithm, but now it includes the indices of the grandparent. Please do note that here each derivation copies the grandparent index g into smaller g-spans. This actually causes each g-span to have non-contiguous structure. This is basically an extension of the second order grand-child factorization

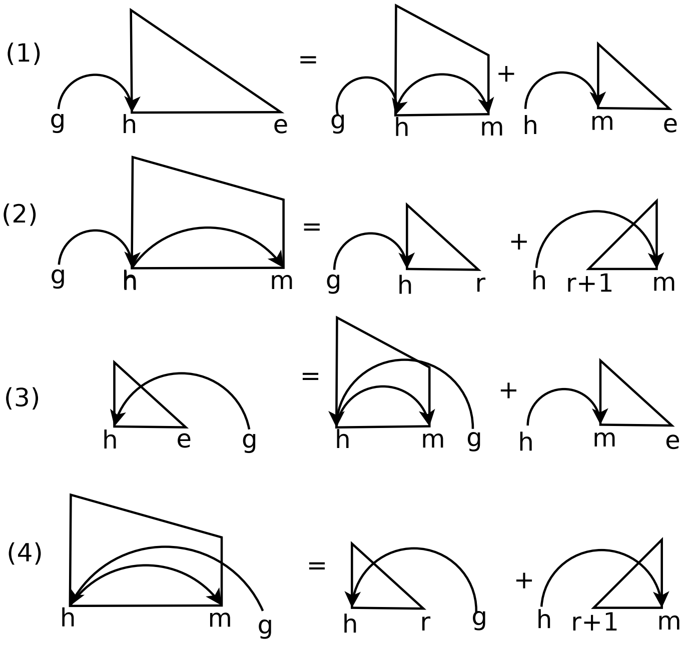

Figure 2.13 Third order dynamic programming structures
Grandchild

Even with this algorithm the time complexity is $O(n^3)$ as each derivation is defined by three fixed indices and one split point.

- *Include all grand-siblings* In this case, we decomposed each tree into a set of grand-sibling parts which consist of the sibling parts and the grand-child parts. i.e., a grand-sibling is a quadruple $< g, h, m, s >$ where

(h, m, s) is basically the sibling part from above and (g, h, m) is the grandchild parts. Its almost like a hybrid of the aforementioned approaches.

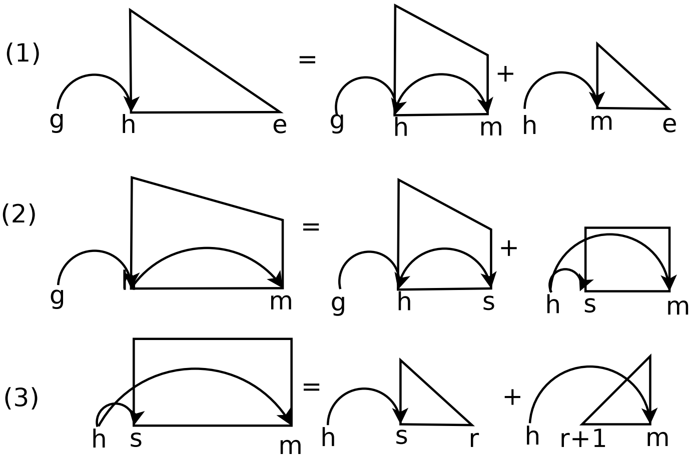

Figure 2.14 third order dynamic programming structures grand-sibling.

2.8 FEATURE SPACE

All through the previous sections we defined the score of an edge, but we intrinsically made an assumption about the feature space. We supposed that we have a high dimensional feature representation for each edge. The feature set description in the above implementations maintain the successful previous work in first order dependency parsing [McDonald et al. 2005], [McDonald et al. 2006], [Carreras 2007].

Let us understand feature space with an example from [McDonald et al. 2005]. Let us consider one of combination of fetures from the form of the words, the lemma of the words, the pos tags of the words. The features are basically the indicatory functions (most often binary), each of the functions evaluate the presence of a certain pattern in a dependency. Consider a feature pattern which takes into account part-of-speech tag and the form of the word into consideration. There is an implicit direction on each dependency part - the left or the right. Let us assume that if $|h|$ ¡$|m|$ then the direction is right and it is left otherwise. If these are the given conditions, then the part $\phi(x, h, m, c)$ can be defined as:

- dir.pos (h).pos (m)
- dir.form (h).pos (m)
- dir.form (m).pos (h)

- dir.form (h).pos (m)
- dir.form (h).form (m)

The most basic feature patterns consider the surface form, part-of-speech, lemma and other morphosyntactic attributes of the head or the modifier of a dependency. The representation also considers complex features that use a variety of part-of-speech tags of the following items: the head and modifier; the head, modifier, and any token in between them; the head, modifier, and the two tokens following or preceding them. Most of the above implementations of the parsing algorithms involve using the syntactic features.

In our experiments we are working with a deprived tree, by depriving it of the dependency-labels. This is because the addition of labels actually increases feature space. For simplicity and computational reasons, we have restricted the experiments to unlabeled parsing. Just like the concept of direction defined above, the label is actually dependent on both head and modifier.

2.9 EFFECT OF FEATURES IN DEPENDENCY PARSING

[Bikel 2004b] provided a detailed analysis of the contribution of each class of distribution to the generative power of the model for generative parsing models. But, unfortunately the non-probabilistic nature of our models prevents that detailed analysis. One of the reasons for the non-probablistic nature of the models is because the models do not perform an exhaustive computation for inference. Thereby, these models are highly generalized in comparison to the generative models as done in [Bikel 2004b]. In this thesis we investigate the effective improvement of certain features and also explain their importance.

2.9.1 *Effect of Semantic Features*

Semantic information basically focuses on the relation between signifiers, like words, phrases, signs and symbols, and what they stand for. Use of semantic information to improve parsing accuracy has been an interesting but difficult goal since the early days of NLP [Ratnaparkhi et al. 1994], [Hektoen 1997], [Xiong 2005]. There have been some good results as shown by [Ratnaparkhi et al. 1994] but the overall integration has been a tough challenge. Recently [Agirre et al. 2011] made an attempt to integrate semantic word classes with transition based data driven dependency parser - the Maltparser [Nivre & Hall 2005] using basic semantic representations using WordNet [Fellbaum 1998a]. Recently, [Agirre et al. 2011] concludes that semantic information gives an improvement on a transition-based deterministic dependency parsing. Also, he mentions that feature combinations give an improvement over using a single feature. Semantic information can be further classified according to its exactness to its intended meaning.

2.9.2 *Morphosyntactic and Morphosemantics with Dependency Parsing*

Morphosyntactic feature is a feature which is involved in either syntactic agreement or government. A typical example is the gender, number and person are involved in agreement. In languages like English and many other languages, syntax is not sensitive to the tense value of the verb. But in languages, especially, highly inflected languages, the tense plays a major. This is a very important attribute for languages with rich morphology and inflections. The relationship of the concept of 'gender' i.e., the concepts 'masculine', 'feminine', 'neuter'; or between the concept 'case' i.e., the concepts 'nominative', 'accusative', 'genitive', etc., with many languages also play a major role in deciding the sentence structure. There are some interesting results about integrating Tense, Aspect, Modality and Minimal Semantics (with constrained and selective semantic relations) with a dependency parser to obtain better results for parsing morphologically rich free word order language [Ambati et al. 2010] [Ambati et al. 2009]. They claim that with the introduction of semantic features there is a significant improvement in the performance of both the parsers. They further state that adding semantic features for nouns helps with label identification more than head identification.

Most of the current research on dependency parsing is focussed on the algorithms employed by the parsers, in this thesis we would concentrate on the issue of the relevant information to the parsing algorithm. That is experimenting on extending the feature structure, the problem of relevant and important feature extraction is as important as the research on parsing algorithms. Feature structures provide information for the inference sub-part of the algorithm.

3 State-of-the-Art and Current Research

The research in the field of dependency parsing was boosted by the successful results in the open shared tasks of the Conference on Computational Natural Language Learning (CoNLL) which concentrated on the task of dependency parsing. The relevant shared tasks which concerns this thesis and the field of dependency parsing directly are shared tasks in CoNLL-2006 [Buchholz & Marsi 2006a], CoNLL-2007 [Nivre et al. 2007a], CoNLL-2008 [Surdeanu et al. 2008] and CoNLL-2009 [Hajič et al. 2009]. Most of the participating teams had novel and highly competitive methods. Though, CoNLL-2008 and CoNLL-2009 shared tasks were joint assignment of syntactic and semantic dependencies.

This thesis focuses on the aspect of pure dependency parsing rather than the task of joint assignment of syntactic and semantic dependencies, hence we will be more concerned with the systems with exclusive dependency parsing results. The results of the CoNLL 2007 shared task is presented in 3.1.

We can see that shift-reduce (or transition based parsers) and [Eisner 2000] based techniques are the most used as parsing approaches and have a very good performance. Eisner-based parsers use edge-factorizations, which are usually simple to approach. Very recently, the new parsing algorithms have been trying to extract maximum number of context by considering more number of words in a part - that is the concept of higher order [Koo & Collins 2010]. Let us now see some of the most important developments in the field of dependency parsing in the recent past which has inspired this thesis directly and indirectly.

In spite of these workshops and shared tasks, there is still a question of whether, even for English, a dependency parser could compete with the other well established phrase-structure parsers, which have access to a richer structure to perform disambiguation. This is still a question for potential research. [McDonald et al. 2005] extracted dependencies from the output of phrase

Parsing Algorithm	English	Czech	Type of Parser
Carreras	90.63	85.16	Discriminative Graph Based
Nakagawa	90.13	84.19	Probabilistic based on Gibbs Sampling
Sagae	89.87	81.27	Transition Based
Nilsson	88.93	83.59	Inductive Transition
Titov	89.73	81.20	Probabilistic

Table 3.1 CoNLL 2007 Shared Task Results. Unlabeled Attachment Score. The parsers are - [Carreras 2007], [Shimizu 2007], [Sagae & Tsujii 2007], [Nivre et al. 2007b], [Titov & Henderson 2007].

structured parser - Collins parser [1] using head finding rules and found that Collins parser was relatively more accurate (0.5 percent) at correctly assigning the heads even with the fact that Collins parser was not trained for the domain.

[Carreras et al. 2008] introduced a dependency parsing algorithm inspired by TAG formalism [Joshi 1969]. The main advantage of this algorithm is the ability to use features from dependency trigrams. This approach uses the splittable grammar formalism - TAG into maximum usage and also produces accurate results.

[Koo et al. 2008] used an interesting approach by using semi-supervised learning. They use Brown clustering algorithm as features and achieve some very interesting results. This is specifically useful method when small amount of training data is available. This is one such research where the use of feature has shown a relatively big improvement in the parsing accuracy. This thesis is totally exploiting the work done by [Koo & Collins 2010] where he introduced the parsers with third order. These algorithms have been previously explained.

Noticeable research in the field of dependency parsing, especially in the area of exploring the feature set has been rather less and not as exhaustive as with other constituent parsers. Some of the recent work that concentrates with exploiting the features are mentioned here.

[Agirre et al. 2011] introduces semantic classes using WordNet, but for transition based parser. The work shows an improvement in the retrieval of labeled accuracies. But, the work does not provide an exhaustive analysis of the semantics on parsing. However, this strongly motivates us to explore the importance of including the semantics into the parser.

[Kitagawa & Tanaka-Ishii 2010] have augmented the selection of parsing actions by using a tree based approach. They build a model that considers all words necessary for selection of parsing actions by including words in the form of trees. It chooses the most probable head candidate from among the trees and uses this candidate to select a parsing action. This is a very new and interesting approach, but, it is restricted to transition based parsers.

[Song et al. 2011] demonstrates a method in which a classifier is used to determine whether a pair of words forms a dependency edge. The classifier trained on the projected classification instances significantly outperforms previous projected dependency parsers when augmented with graph based dependency parsers.

[Novák & Žabokrtský 2007] have showed that optimizing feature templates, there is a chance of getting a better parse structure in state-of-the art graph based dependency parsing algorithms. This work also makes a comparison to the resources spent v/s the improvement in the obtained result.

In general we can make a dichotomy of the features -

1. Lexical Features and
2. Semantic Features.

[1] http://www.cs.columbia.edu/mcollins/code.html

Also, history-based models are simpler to design, as they do not require problem-specific independence assumptions to be crafted, nor independence-assumption-specific inference algorithms to be invented. Transition-based parsers have the opportunity to extract non-local features by examining the stack, but they are intractable when they exploit such non-local features (hence the greedy search). All of these ″history based″ models have a very restricted kind of history, though in a way we could condition on anything but, in practice we only work with local knowledge. Also, the search is structured in a way that it is very difficult for later decision to override earlier ones. These ″history-based″ models are not really stronger than a forth-order parsing model.

The experiments done by [Koo et al. 2008] is mostly lexical but the semantic information is extracted using the lexical semantics. While, in the work by [Agirre et al. 2011] is more into the real semantic features. In any of the above cases, the number of features has to be limited. Else, there is always a chance of overfitting due to the potential loss of information and introduction of unwanted random noise in the learning and parsing process.

Experimentation Details

4

In order to evaluate the impact of the semantic level features, we conducted several dependency parsing experiments in English and Czech. In this chapter we will describe the system settings, for the experimentations that we will be doing to obtain the results. We have tested with several sets of feature based experiments, while maintaining the algorithmic constraints. We test it on first order, second order and third order parsing models with several configurations of features. In the following sections we shall explore the the idea and the configurations of the parsing experiments.

4.1 IN FOCUS

Let us recollect the algorithm that we have explained before. The parsing algorithms used in the parsers are fine tuned on the framework of averaged perceptron [Koo & Collins 2010]. The parser basically scores the part as shown here:

$$Part(\vec{x}, p) = \vec{w} \cdot \phi(\vec{x}, p) \tag{4.1}$$

In the above equation, ϕ is a feature vector mapping and $\vec{w}$ is a vector of related parameters. As described before, following the standard approaches here from [Koo & Collins 2010], [Carreras 2007] and [McDonald et al. 2006], the model scores for all of the parsers follow a generic pattern.

Please note, it is not only the higher order parts that the parser are mapped, but even the lower order parts are evaluated. For example, let us consider the third order grandsibling parser. This parser evaluates the mappings for dependencies, siblings, grandchildren and all the grandsiblings as well.

The score is calculated as:

$$\begin{aligned} Score(\vec{x}, y) = \sum_{(h,m) \in y} \vec{w} \cdot \phi(\vec{x}, h, m) + \sum_{(h,m,s) \in y} \vec{w} \cdot \phi(\vec{x}, h, m, s) + \\ \sum_{(g,h,m) \in y} \vec{w} \cdot \phi(\vec{x}, g, h, m) + \sum_{(g,h,m,s) \in y} \vec{w} \cdot \phi(\vec{x}, g, h, m, s) \end{aligned} \tag{4.2}$$

In this equation, we have defined the dependency- $(\vec{x}, h, m)$, the siblings - $(\vec{x}, h, m, s)$, the grandchildren - $(\vec{x}, g, h, m)$ and the grandsiblings - $(\vec{x}, g, h, m, s)$ as the different parts which are actually the decomposed part structures.

What is special in this parsing approach is that it is not just the considering the lower order parse structures, but the way the feature combinations take place. In this parsing substructure, we can have 4-gram context features - consisting of 4-gram POS augmented with adjacent POS tags (when using

POS as the feature). Consider an example of $\phi(\vec{x}, g, h, m, s)$. This basically includes the POS features at these positions - $(g, h, m, s, g+1, h+1, m+1)$, this means a POS 7-gram feature structure.

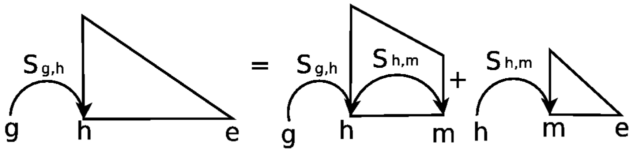

Figure 4.1 Example Structure where Sense information is included

Consider a simple example as shown in figure 4.1. Here, 'S' - wordsense information with the parser it increases the amount of information required for the parser to parse the structure of the sentence correctly. This is, in general, more pronounced when we talk about higher order structures. [Koo et al. 2008] has mostly tried to augment structures which are either syntactic or approaches like the brown clustering. In this thesis, we try to experiment with a linguistic analysis of feature addition.

4.2 INHERENT FEATURE COMBINATION IN THE PARSER

The first order feature set in [Koo & Collins 2010]'s parser works exactly similar to the feature set of [McDonald et al. 2005]. It consists of indicator functions for combination of words and parts of speech for the head and modifier of each dependency, as well as certain contextual tokens. [Koo & Collins 2010] has further augments the above feature set with backed off versions of"Surroundng Word POS Features" which includes only neighboring POS tag.

The higher-order baseline features are implemented in the similar manner as implemented by [Carreras 2007]. Essentially, here the parsing algorithm considers part of speech tags for sibling interactions and grandparent interactions. This also includes additional bigram and trigram based features on pairs/triplets of words involved in these higher-order interactions.

4.3 SYSTEM INFORMATION

A large number of experiments were performed through the development process. As we stated at the start of this work, our main point of interest is to compare the scores with different features and see if there is any significant effect on the score. We re-ran some experiments with the latest system configuration to facilitate a comparison across experiments. Some experiments are really expensive and hence were done using a reduced corpus, which will be explained below.

4.3.1 *Input Format*

The input format for our experiments is using the CoNLL-X format [1] [Buchholz & Marsi 2006b].

The data format is as follows:

- The data files contain sentences separated by a blank line.
- A sentence consists of one or more tokens and the information for each token is represented on a separate line.
- A token consists of at least 8 fields, described in the table below. The fields are separated by one or more whitespace characters. Whitespace characters are not allowed within fields.

The CoNLL format which we used for our experimentation Czech:

1. ID - Token counter, starting at 1 for each new sentence.
2. FORM - Word form or punctuation symbol.
3. LEMMA - Lemma or stem (depending on particular data set) of word form, or an underscore if not available.
4. CPOSTAG - Coarse-grained part-of-speech tag, where tagset depends on the language.
5. POSTAG - Fine-grained part-of-speech tag, where the tagset depends on the language, or identical to the coarse-grained part-of-speech tag if not available.
6. FEATS - Unordered set of syntactic and/or morphological features (depending on the particular language), separated by a vertical bar (—), or an underscore if not available.
7. HEAD - Head of the current token, which is either a value of ID or zero ('0').
8. DEPREL - Type of dependency relation. The set of dependency relations depends on the particular language.

4.3.2 *Corpus Used*

We are experimenting with two languages - **English and Czech** which are widely used in shared tasks and are also widely experimented. This is also largely because of the ready availability of finely tagged standard corpora. In this section we will describe the corpora which were used in the process of experimentation. We used a "reduced corpus" as these experiments are heavily computationally intensive and need a lot of resources for different experiments. The parsing algorithms, as shown by [Koo & Collins 2010], has a time-complexity of $O(n^4)$. In practice, other computations which concern the data conversion and other sub processes also consume a lot of computational time. For easy replication of the results, the exact details of the corpus are also provided.

[1] This can be accessed here http://ilk.uvt.nl/conll/

English

The corpus consisted of Penn Treebank (PTB) [Marcus et al. 1993] corpus and it was converted to the required format by using Penn2Malt [Johansson & Nugues 2007] constituency-to-dependency converter. We used a subset of this corpus that consisted of:

1. 15,000 Sentences - Training
2. 1000 Sentences - Validation
3. 2000 Sentences - testing

For English, the interest in developing exclusive dependency parsed corpus has been a bit weaker than for other languages. This is probably because of the strong tradition of constituent analysis in Anglo-American linguistics, this is reinforced by the creation of a big treebank for American English, the Penn Treebank [Marcus et al. 1993], that is annotated with constituent analyses. At the same time, there has been increasing interest in using dependency parses for a range of NLP tasks, from machine translation to question answering. This is one of the reasons, why most of the other languages have shifted to building treebanks natively with a dependency grammar formalism. Even in this work, we use Penn2Malt [Johansson & Nugues 2007] for the process of conversion from the constituency tagged corpus format to a dependency tagged corpus.

The exact details about the corpus divisions are provided here. Please note that the corpus was extracted from Penn Treebank. The specifics are:

1. Training Set: This was built from section 2 to section 10. A set of fifteen thousand sentences were extracted from this set. These sections were combined n a serial order and 15,000 sentences were extracted. In a similar manner every other sections were extracted.
2. Validation or Development Set: Random selection of sentences from sections 15, 17, 19 and 25. This set contained a set of one thousand sentences.
3. Testing Set: This was chosen from sections 0, 1, 21, 23. This is contained two thousand sentences in all.

Czech

We used the Prague Dependency Treebank (PDT) [Böhmová et al. 2001] for Czech, which was converted by using some scripts provided in the TectoMT. Though the data consisted of a big set of sentences, unfortunately generating models for all of the them is a very expensive task. Since the result was computed with several models with similar dataset, we believe that the results still holds merit as these were comparative results.

The Exact division of the sets were.

1. 15,000 Sentences - Training
2. 1000 Sentences - Validation
3. 2000 Sentences - testing

The Prague Dependency Treebank [Böhmová et al. 2001] consists of Czech texts annotated with syntactical information consisting mainly of dependency relationships. Unlike English, Prague Dependency Treebank is natively a corpus following the dependency grammar formalism. One of the most popular benchmarks for evaluating parser quality is by evaluating against the surface-syntactic trees provided by the Prague Dependency Treebank.

The Czech side of the experimentation was extracted from the "pdt-full-automorph" dataset of the Prague Dependency Treebank. It constituted of:

1. Training Set: This was built from train1 to train5 splits of the dataset. It had in total a set of fifteen thousand sentences.
2. Validation or the Development Set: This section was taken from the train6 and train7 sets. It had a set of one thousand sentences.
3. Testing Set: The test set was made up of dtest and etest parts of the dataset. It contained two thousand sentences in total.

The Czech wordnet is unfortunately a closed work, this was tackled by exploring the Prague Dependency Treebank for the near-semantic annotations as explained in the later half od the thesis.

4.3.3 *Tools*

Dependency Parser

We use [Koo & Collins 2010]'s higher order dependency parser dpo3[2] which is freely available with a GPL license. This parser already builds algorithms for all the higher order parsers and provides the ideal basis to experiment. This lets us use the ideal base to experiment with different dynamic algorithms and test our hypotheses.

The training is done using the standard average perceptron training algorithm. The learning method was retained as it is a very fast converging algorithm, typically converging in around ten iterations. For our experimentations, we have trained each parsing experimentation for ten iterations and then we have selected the best score on the development set. In practice, the convergence varies for different corpora and with the input feature set.

4.4 WORD SENSE EXTRACTION

- **Fine-grained Word Sense Extraction** We use the standard word sense disambiguation [Pedersen & Kolhatkar 2009] algorithm to do the basic word sense disambiguation. What it does is it, finds the sense of each word that is most related to the senses of the surrounding words based on a similarity measure. It proceeds word by word from left to right, centering each content word in a balanced window of context, whose size is determined by the user, of surrounding words. At each stage, the token being disambiguated is called "the target", and the surrounding

[2] This can be accessed here http://groups.csail.mit.edu/nlp/dpo3/

tokens "the context window". The size of the context is determined by the user and will be referred to as window size. A balanced context is chosen according to the size of the window. The goal of the algorithm is to select one of the senses from the set of possible senses. This is done by measuring the semantic relatedness between the possible senses of the target and the possible senses of each of the tokens in the context window. Consider a sense pair (s_k, s'_{il}), where s_k is the k^{th} sense of the word being considered and s'_{il} represents the l^{th} sense of the i^{th} word in the context window. A 'relatedness' function takes as input two senses, and outputs a real number. It is assumed that this real number is indicative of the degree of semantic similarity between the two input senses. A larger number denotes high relatedness between the two senses and a smaller number denotes low relatedness between the senses. In simple terms the equation that is used to calculate the sense is given below:

$$S_k = \sum_{1ton} (\max(s_k, s'_{il})) \tag{4.3}$$

Here, S_k is the final-chosen sense for the given word in the context.

- **Coarse-grained Word Sense Extraction** [Fellbaum 1998b] describes that the organization of words in the WordNet are done as sets of synonyms, called synsets. Each synset in turn belongs to a unique semantic le (SF). There are total of 45 SFs (1 for adverbs, 3 for adjectives, 15 for verbs, and 26 for nouns), based on syntactic and semantic categories. A script is written in order to fetch coarse-sense disambiguation. This script also uses [Pedersen & Kolhatkar 2009]'s word sense disambiguation tool, but then makes it more coarse grained for our experiments. The tags are $SF00...SF44$.

This is not exactly the case for Czech. The difficulty is due to the problem of having a reliable sense-tagged corpus for Czech [1]. It would have been much more easier to get the senses disambiguated if there was a good sense-tagged corpus. In lieu of the sense tagged corpus for Czech, we extracted an almost similar feature from the Prague Dependency Treebank - "SEMPOS" or the semantic part of speech. We will explain in detail about the extraction in the next chapter.

4.5 MORPHOSYNTACTIC FEATURE EXTRACTION

These were properly done for the Czech language since we had the tagged corpus with a set of morphosyntactic features, also these were of Gold standard. The morphosyntactic features are the basic morphological features in the Prague Dependency Treebank which are extracted from the m-layer.

There are 13 categories in the Czech morphological tagset with 4452 plausible combinations. These are briefly mentioned in the table 4.1. We extract the morphosyntactic features out of this table - both the whole feature set

[1] We did not have the Czech WordNet available for our experiments.

Category	Number of Values
POS	10
SUBPOS	75
GENDER	8
NUMBER	4
CASE	9
POSSGENDER	4
POSSNUMBER	3
PERSON	5
TENSE	4
GRADE	5
NEGATION	3
VOICE	3
VAR	3

Table 4.1 Morphological Tagset for Czech

and take specific morphosyntactic features into consideration and try to experiment with different possible combinations. Some of these tags are very important and are mostly specific for an inflectional language as these contain a lot of information which is essential to build a good sentence. This inspires us to embed these features into the various dependency parsing algorithms and gain an insight into these feature structures.

Again for the English language, we used the fine-grained part-of-speech tags. These tags also contain some morphological information, but not as extensive as the Prague Dependency Treebank has. This is again directly extracted from the Penn Treebank.

5 Experiments

This chapter presents a set of results from the experiments carried out on the data introduced in the previous chapter. At first, we present the hypothesis and discuss the theoretical plausibility of the hypothesis, then we provide a brief description of the experimental setup. Finally, we present the results and a discussion on the obtained results and the experimental outcome.

The major part of the experimentation was with features. To have a holistic idea of the performance of the features, we considered all possible combination of the feature space. This included from grandchild and grand sibling to the head and the modifier with the possible feature combination. In all the experiments we took the basic coarse grained part-of-speech as one of the constant features. This was also the baseline for all the experimentation.

The performance of the feature enhanced parsing algorithms are compared with the parsing algorithms augmented with basic coarse-grained part-of-speech feature and form as the features. We calculated the unlabeled attachment score, that is, the score which calculates the number of times a decision - where the head has been correctly predicted. This is one of the standard metrics used in the evaluation of the dependency parsers both in the shared tasks as well as in the research publications. The result hence, could easily be verified. As explained before, the test set comprised of two thousand sentences.

The experiment is made to run at least ten times to get the maximum convergence with the score and get the optimum result. We have provided a list of run-by-run variations for each parsing algorithm across all the experiments in the "Appendix". In most of the experiments we have got the convergence, but on some of them there was a necessity of more iterations for the convergence.

5.1 FEATURE SPACES

As explained in the previous chapters, our parsing algorithms have a high-dimensional feature representation for each edge (i, j). Again, we have described the feature space in Section 2.8 where we discussed about the structure of combining the feature patterns. For our experiments, we have experimented with all possible feature structure combinations. One of the feature structure combinations is just similar to the feature structure combination as mentioned in [McDonald & Nivre 2007]. The second is basically the one with joins the structure with the distance between the two siblings as well as the direction of attachment (from the left or right). These features were modified on a development set. We tried additional features, such as the POS of words in-between the two siblings and POS of words in between parent and children.

5.2 EXPERIMENT 1

5.2.1 *Research Question*

"What is the effect of using the fine-grained semantic word-sense as a feature with different graph based dependency parsing algorithms?"

5.2.2 *Theoretical Plausibility*

English

As explained in the previous chapters, we can see that dependency parsing, in general, is very useful for semantic analyses. This draws us closer to the question of whether there is any effect on the parsing accuracy if any semantic information is available to the parser. We investigate the possibility by including specific word-senses. This becomes especially interesting with the higher order parsing algorithms, since, these can explore better context information.

To explain the possibility of wordsense making an impact, let us consider an example as shown in 5.1 and 5.1. Now, given the word form and the part of speech tag, "cricket" has exactly the same POS tag in both 'cricket as an insect' and 'cricket as a sport'. We can see that the possibilities, given several parses with the same form and POS information, would make it ambiguous. This would change if we have more information. In this example, the sense can be useful to distinguish between the two different forms of the word.

A sense tagged corpus has the ability to give a better set of information to the parsing algorithm. Also, if we consider, second and higher order parsing algorithms, then the amount of information would be significant and hence theoretically the accuracy in predicting the structure would be considerably higher. This experiment was inspired by [Agirre et al. 2011]'s recent work on improving dependency parsing accuracy by using semantic classes as features. In [Agirre et al. 2011]'s experiments, they use the most generic semantic tags as features, we use the same logic in the next experiment.

Czech

Following the same logic as for English, for Czech, there is a well defined Semantic POS (the tag name is SEMPOS) tag in the t-layer of the Prague Dependency Treebank. We have symbolically treated this as the fine grained semantic tagset. But one interesting issue is that these are Gold standard tagsets and hence we can have a clearer look at the results. The tagset for Czech is explained in the following section.

The reason for experimentation with Gold standard tags is because the relative use of the semantic features is still new and is an untested area. This experiment might prompt a tool which could generate a sense tagged corpus based on the relative usefulness in this experiment. The guiding factor again

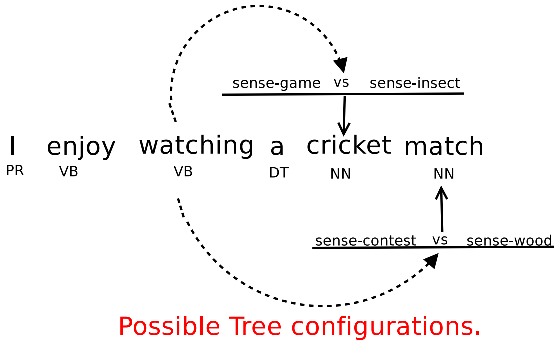

Figure 5.1 Effect of Wordsense

1	Ms.	-	NN	n2	-	2	NMOD
2	Haag	-	NN	-	-	3	SUB
3	plays	-	VB	v3	-	0	ROOT
4	Elianti	-	NN	-	-	3	OBJ
5	.	-	.	-	-	3	P

Table 5.1 A sample set from the Penn Treebank tagged with fine-grained word-sense tags for English

is based on the experiments done by [Agirre et al. 2011]. Also, in [Agirre et al. 2011]'s experiments, they have worked with the semantically Gold tagged corpus to explore the effect of the semantic tags.

5.2.3 *Experimentation*

English

As explained before, we use [Pedersen & Kolhatkar 2009]'s algorithm to disambiguate and associate senses for each word in a given sentence. Now, given a target word and its part of speech, the algorithm chooses the best possible sense. Consider an example sentence "I enjoy watching a cricket match", in this case cricket as a game would in the WordNet belong to the sense number 1. Hence, the output would be n1 when tagged. A sample sentence from the training dataset is mentioned in 5.1:

Czech

The t-layer of the PDT [Böhmová et al. 2001] is tagged with the grammateme semantic postags or sempos. We have extracted the sempos from the PDT and we are have experimented with the sempos. The tagset of the sempos is given in 5.2.

Type	Explanation
n.denot	denominating semantic noun
n.denot.neg	denominating semantic noun with separately represented negation
n.pron.def.demon	definite pronominal semantic noun: demonstrative
n.pron.def.pers	definite pronominal semantic noun: personal
n.pron.indef	indefinite pronominal semantic noun
n.quant.def	definite quantificational semantic noun
adj.denot	denominating semantic adjective
adj.pron.def.demon	definite pronominal semantic adjective: demonstrative
adj.pron.indef	indefinite pronominal semantic adjective
adj.quant.def	definite quantificational semantic adjective
adj.quant.indef	indefinite quantificational semantic adjective
adj.quant.grad	gradable quantificational semantic adjective
adv.denot.ngrad.nneg	nongradable denominating semantic adverb, impossible to negate
adv.denot.ngrad.neg	nongradable denominating semantic adverb, possible to negate
adv.denot.grad.nneg	gradable denominating semantic adverb, impossible to negate
adv.denot.grad.neg	gradable denominating semantic adverb, possible to negate
adv.pron.def	definite pronominal semantic adverb
adv.pron.indef	indefinite pronominal semantic adverb
v	semantic verb

Table 5.2 Semantic POS-Tags for Czech from the Prague Dependency Treebank

5.2.4 *Results*

English

As we have stated above, for this experiment, we evaluated the performance of our system using 15,000 sentences from the Penn Treebank corpus which are converted by the "pennconverter" into the CoNLL format. This is a tool to automatically convert the constituent format used in the Penn Treebank into dependency trees. This was also used in the previous versions of the CoNLL shared task. We tagged the corpus with the fine-grained semantic tags, i.e., these tags actually mention the possible sense number from the list of senses retrieved from the WordNet. Seeing table **??** we notice a gradual improvement in the parsers accuracy for unlabeled attachment score.

Czech

The table 5.4 shows the result for the Prague Dependency Treebank. In this case, the experiment was done with Gold standard tags. Hence, this just gives us an indication of the relative importance of the semantic tags for Czech.

5.2.5 *Discussion*

One basic difference between our approach and the other previous approaches on augmenting wordsense is that here, we use a relatively 'more specific' wordsense of the word. This provides more specific information to the parsing algorithm. The grand-sibling based parsing algorithm shows better per-

Parsing Algorithm	Semantic Tags	Baseline	Difference	Mean Score (refer Appendix 8)
Third order grand-sibling	86.87	85.67	+1.20	86.15
Third order GrandChild	86.82	86.03	+0.79	86.44
Second order grand-sibling	85.19	84.32	+0.87	85.45
Second order GrandChild	85.23	84.79	+0.44	84.26

Table 5.4 Results for Prague Dependency Treebank

formance than the grandchild based parsing algorithm. A close analysis reveals that the sibling based interactions that are local, are easily retrieved. While, the farther sibling interactions don't necessarily give better results.

If we closely look at the results we find that for both the languages, there is a generic uplift in performance for all the different kinds of algorithms. But, significant results can be seen especially in the sibling-based parsing algorithms. The standard test for statistical significance is heavily expensive and hence we take a different route. We average all the learning stage models score with respect to the test data to compute the average significance. Both in English and in Czech the grand-sibling based algorithm show a good improvement. In this particular case you see the difference, there is a big difference for the Czech language, where the third order sibling parser differs from the original score by a difference of +1.20. The other differences are of similar magnitude, but more pronounced with sibling parsers.

In comparison to [Agirre et al. 2011]'s work, we have got a significant improvement on both Czech and English datasets. An automatically tagged Czech corpus with the semantic tags could be a good experiment, since the Gold tagged one has given us a positive result.

5.3 EXPERIMENT 2

5.3.1 *Research Question*

What is the effect of using the coarse-grained semantic word-sense as a feature with different graph based dependency parsing algorithms?

5.3.2 *Theoretical Plausibility*

Before we justify the cause of the hypothesis, let us try to understand what coarse-grained semantic word-sense tags are. For English, tthese tags are basically a very high level representation of the possible semantic orientation of the word. For example, an animate noun which denotes action or acts would be classified in a particular set containing all of the animated nouns.

For English, a similar approach has been tried for transition based dependency parsing by [Agirre et al. 2011]. We approach in a similar manner and work on the level of semantic files to extract the details of the coarse-grained wordsenses. Though, in our case, due to the lack of gold tagged data that has been explored in [Agirre et al. 2011]'s work, we use just WordNet to tag our corpus.There has been some improvement with the transition based parsing, we experiment if at all there is an effect with graph based dependency parsing algorithms, especially the ones with higher order context information being made available.

Unfortunately, we couldn't extend the same results to Czech. This was primarily due to a problem in fetching a similar structural source like the WordNet due to several other practical constraints for the coarse grained semantic tags.

The difference between the coarse and fine grained semantic tags lies in the amount of information made available to the parsing algorithms. The coarse grained information is a more generic form of information. There can be a big set of words which belong to a particular semantic file. This presents the parser witch generic set of information. Consider an example of the word 'path' and the word 'way'. These two belong to the same family henceforth, they will be numbered with the same semantic file number.

Now, consider a hypothetical case where, the training data has a sentence, say, *"Path of least resistance."*. Assume that this sentence is semantically tagged and then the parser trains on this dataset. Again, consider a sentence in the test data *"Way of least Resistance"*. In this sentence, it can be seen that all the other words are the same set except for the word "Way". In this case, both "Path" and "Way" have the same POS and the same semantic tags. This would make the sentence to be parsed correctly, due to the presence of semantic information. There are cases when the POS tags are not similar, but semantic-file would be same and the sentence would have a similar parse structure. This would help the parser decide a better option of selecting the best possible head by using the semantic-file as one of the features.

Parsing Algorithm	Semantic Tags	Baseline	Difference	Mean Score (refer Appendix 8)
Third order grand-sibling	91.10	90.29	+0.81	91.22
Third order GrandChild	90.72	90.57	+0.15	90.36
Second order grand-sibling	88.54	87.45	+1.09	88.35
Second order GrandChild	88.62	88.34	+0.28	88.90

Table 5.5 Results for Penn Treebank

5.3.3 *Experimentation*

The algorithmic implementation is described here. The algorithm uses [Pedersen & Kolhatkar 2009]'s algorithm to extract the best possible sense of the word for the given context. This then is used to extract the synset of the word, through which we extract the final semantic file number for the target word.

Algorithm 2 Pseudo-code for the semantic file tagging algorithm.

```
Input: Tokenized sentence and the window size n
function: disambiguate-all-tokens (input[ ], n): disambiguated-input[ ]
for tokens t in input[ ] do
    best-sense = disambiguate-single-token (input[ ], t, n)
    disambiguated-input[t] = $w_t$ with best-sense assigned
end for
end function
function: tag-symantic-file (disambiguated-input[ ]): semantic-file-tagged-input[ ]
for tokens and senses t in disambiguated-input[ ] do
    synset-information = synset-extraction(disambiguated-input[ ], y)
    semantic-file-tagged-input[y] = t (tagged with semantic-file information for the synsets)
end for
return semantic-file-tagged-input[ ]
end function
```

This algorithm assigns the semantic file information for the given target word given the sentence. As we have explained before, these were basically the semantic files and out tag representations were the closed set $[SF00...SF45]$.

5.3.4 *Results*

The results for the experiment are mentioned in table 5.5. We can note that there is a significant change due to the addition of the semantic tags. Also note that the lower order parsing algorithm has a better relative improvement than the higher order parsing algorithms.

5.3.5 *Discussion*

Again, in this case, we see that the sibling based parsers perform better than the other forms of parsing algorithms. Also, please note that the second order parser shows a better relative improvement than the third order sibling parser with an improvement of +1.09 units.

The improvement in the performance of the parser is a very encouraging result. Given the nature of semantic classes and word sense disambiguation algorithms, there seems to be room for a lot of improvement. This gives us the possibility of exploring information like WordNet concepts, wikipedia concepts and other related concepts, which could be essentially important for various Natural Language Processing tasks like Semantic Role Labeling, etc., also, these results are very interesting for fields of Machine Translation and other related fields.

This experiment, to a certain extent, indicates the effect of having high level word classes. An experiment with word classes was performed by [Koo et al. 2008], which showed a significant effect on the lower order parsing algorithms. This experiment provides the motivation for a similar experiment for the higher order dependency structures. Though, this thesis has not explored any of the effects of using the approaches based on lexical word-classes, which are basically semi-supervised structures. In a generic word-clustering approach, lexical information is considered as a crucial step to resolving ambiguous relationships. While in our case, we have tried to go beyond the lexical relationship. We are trying to scale the features upto a semantic level.

5.4 EXPERIMENT 3

5.4.1 *Research Question*

"What is the effect of using the morphosyntactic tags as a feature with different graph based dependency parsing algorithms?"

5.4.2 *Theoretical Plausibility*

Czech

Czech is a morphologically rich language [Horák et al. 2007]. The morphological tags are known to contain a lot of significant lexical information. This information seemingly play an important role in the parsing of morphologically rich free order languages. With a big window of contexts, the amount of useful information could be extended.

English

For English, though morphological information is important, but the amount of contribution of the morphology might not be as great as the contribution from Czech. We used the fine-grained POS tags directly from the Penn Treebank's tagged corpus. We found out that the fine-grained POS contains a significant amount of morphological information. The tags comprised of the standard tag set from the Penn Treebank.

5.4.3 *Experimentation*

Czech

We initially tried with the 15-letter tags as individual features and exploited the whole tagset with the parsing algorithm. But unfortunately, it couldn't give better results mostly due to the problem of over-fitting of the feature space. This made us make experimentation on linguistically coherent choices with respect to the parsing decision. Later, we chose a subset of these morphological features, the subset was chosen on the basis of relative importance of the particular morphological tag in providing the relevant information which might be helpful. The tags that we considered were specific tags like:

- **Gender** - Gender is an inherent feature of nouns and is also a contextual feature. Gender in basically determined through agreement. Gender is lexically produced and its value is fixed for the noun.
- **Number** - It is a morphosyntactic feature if it participates in agreement.
- **Case** - Case is a feature that expresses a syntactic or semantic function of the element that carries the particular case value.
- **Person** - Person as a morphosyntactic feature is typically a feature of agreement.
- **Tense** - It denotes the semantic feature of location in time.

Parsing Algorithm	Morph. Tags	Baseline	Difference	Mean Score (refer Appendix 8)
Third order grand-sibling	86.12	85.67	+0.45	85.97
Third order GrandChild	87.75	86.03	+1.72	87.54
Second order grand-sibling	84.88	84.32	+0.56	84.73
Second order GrandChild	85.51	84.79	+0.72	85.32

Table 5.6 Results for Prague Dependency Treebank

Parsing Algorithm	Morph. Tags	Baseline	Difference	Mean Score (refer Appendix 8)
Third order grand-sibling	90.50	90.29	+0.21	90.55
Third order GrandChild	91.78	90.57	+1.21	91.04
Second order grand-sibling	87.67	87.45	+0.22	87.45
Second order GrandChild	89.32	88.34	+0.98	88.91

Table 5.7 Results for Penn Treebank

- **Voice** - The temporal feature between the subject and the verb.

English

As explained before, for English we use the standard Penn Treebank's fine grained tagset. It contained a set of thirty six tags. These tags are specifically mentioned in [Marcus et al. 1993] and these are used as the standard tagsets for tagging English in most of the part-of-speech taggers. Also, in most of the dependency parsing shared tasks, one of the columns is mentioned for the fine-grained part-of-speech.

5.4.4 *Results*

Czech

As explained in the previous section, preliminary tests on a portion of train data showed that the complete morphological tagset feature templates decrease the accuracy. Hence we concentrated on experimenting with the smaller morphological tagset. Table 5.6 shows the result.

English

As explained before, we are using the fine-grained tagset here. The results are provided in 5.7. What is interesting is that the present set of morphological features are minimal while the improvement is statistically significant.

5.4.5 *Discussion*

Third order grandchild shows a relative improvement of about 1.72% for Czech. This is an important result, since grandchild based parsing algorithms seem to be better than their counterparts when included with the morphological tags. Also, please note, this corroborates the linguistic assumption about morphological information, an important factor for the morphologically rich languages.

In case of English, the fine-grained POS tags, basically these tags are enriched with morphological information, also show a very interesting improvement. The third-order grandchild based parser shows +1.21 relative improvement.

6 Analysis

The goal of the thesis was to explore the higher order dependency parsing algorithms. We have tried to experiment with the parser using some of the semantic and syntactic feature sets. While the practical implementation of some of the features is not directly available, for example, the SEMPOS from the Prague Dependency Treebank was manually tagged as there is no published material to show an automated way of tagging the semantic tags without the availability of dependency trees. But, the experimentation gives an indication of usefulness of the features. For languages with a good WordNet and a semantically tagged corpus, extracting semantic information should not be a big challenge, since they would have almost the same disambiguation algorithm as has been described by [Pedersen & Kolhatkar 2009].

Ideally it is desirable to use many features collectively and perform the process of parsing, eventually arriving at an optimal solution. But, each feature increases the search space quite remarkably and hence, there are two important problems here -

1. **Problem of Over-fitting** Over-fitting generally occurs when a model is excessively complex, such as having too many parameters. This works in a contra-productive way most of the time.
2. **Parsing Time** The amount of parsing time increases with new features too. This is again because of the increase in the search space which might act in a negative way to reduce the parsing accuracy.

Also, please note that the current tagging of the English wordsense was done using a simple algorithm [Pedersen & Kolhatkar 2009]. There are better algorithms which could give better tagging accuracies. This might have a direct effect on the parsing performance. Another important thing to note is that the morphological taggers have better accuracies in practical applications. Hence the proposed approach would be useful, if we were able to carefully combine the accurate features.

POS tags provide very basic linguistic information in the form of broad grained categories. Among all the parsing structures, we saw that an augmentation of the specific wordsenses has improved the unlabeled accuracy scores. It is evident that one of the parsing algorithms - performs much better than other parsing algorithms. It is more important to note, it is the type of the parsing algorithm along with the features that makes a noticeable change in the score.

Also, after a close investigation when we introduced morphological tags, we found that we have a very strong problem with agreement. The agreement problem is bad whenever there is a coordinating conjunction or when there is a complex verb. There is still a lot of work to be done with the parsing structures and on the observe the correct set of features to take full advantage of dependency grammar in practical application of NLP.

6.1 COMPARISONS

In this section we try to compare our experiments with other similar experiments. We try to analyze the experiments that we have done in the thesis and their relative significance.

6.1.1 *Work done by Agirre et al. 2011*

This work is technically directed at introducing generic semantic features as semantic information into the parsing algorithm. This was primarily done on the Penn Treebank. Maltparser was used as the main parser for the parsing experiment. The parser produces a dependency tree in a single pass over the input using a stack of partially analyzed items and the rest of the items in the input. It determines the best "action" at each step by using models which maintain history and SVM classifiers. Maltparser allows the introduction of semantic and other related features in the training model as the parsing action is dependent on the feature set. The only drawback is that the decision of the action is local, that is, the history is restricted, and in most cases this might turn out to be a bad action that might affect in the result. [Agirre et al. 2011] uses the semantically tagged Semcor and full Penn Treebank intersection as the dataset. The experiment was done by using three different ways, these are:

- Gold standard tags: Manually tagged corpus used as training set.
- 1^{st} sense extraction: Extracting only the 1^{st} and the most relevant sense.
- Automatically Sense tagged corpus: This uses a similar algorithm to the one we have used for getting the wordsenses for English and tries to predict the correct wordsense of the word.

[Agirre et al. 2011]'s work makes use of the semantic-file type of tags, which in our thesis we have referred to as the coarse-grained semantic tags. The results are evaluated using the Labelled Attachment score, that is, the score is the proportion of tokens that are assigned the correct head word as well as dependency type. In the experiments, the resultant improvement in the score is mildly significant, but then it shows that semantic features have a positive effect on the parsing algorithm.

The experiments in the thesis, in relation to semantic tags, had a similar setup. Though, for English, we experimented with almost the same approach when using coarse grained semantic sense as feature set. The experiment which we had done showed a significant improvement. It showed that the second order parsing algorithm performed much better with the semantic tags than others. In the first experiment we try to evaluate the parser with specific semantic tags, in a way this is closer to the 1^{st} sense tag extraction evaluation. In the case of English, the data was tagged by an unsupervised algorithm, while, for Czech the experiment was done with tags extracted from the Prague Dependency Treebank. Again, in both the cases we got a significant improvement in the results.

An important observation is that the semantic tags - both specific and generic, actually do significantly improve the parsing accuracy. And this is more pronounced in higher order dependency parsing algorithms.

6.1.2 *Work done by Koo et al. 2008*

The work done in [Koo et al. 2008] is using a semi-supervised method for training the dependency parsers. This is basically done by using features that incorporate word clusters that are derived from a huge unannotated corpus. The basic algorithm used is the Brown Clustering algorithm, where the input to the algorithm is a vocabulary of words to be clustered and a corpus of text that contains the given vocabulary. Then a standard clustering procedure is followed and it results in a hierarchical clustering of words. The clusters can be represented by a binary tree, here, each word is uniquely identified by its path from the root. This path can further be represented by a bit string. These were then encoded in the training data, hence, augmenting the lexical information. In a way, there is a mismatch between the kind of lexical information that is captured by the Brown clusters and the kind of lexical information that is modeled in dependency parsing.

The experiments were performed on both English and Czech, again, using Penn Dependency Treebank and Prague Dependency Treebank respectively. The parser used the first order and second order graph based algorithms that are explained before. The results are presented with a comparison of the respective Unlabeled Accuracy Score. The cluster based features achieve improvement at all training set sizes. There was an important result which came out of this work and that was data-reduction factors, that is, the result for a particular set of data by a normal algorithm (without the addition of the features), was equivalent to the result obtained by using half the data with the parsing algorithm which used word cluster features.

The work done in our thesis has a similar strategy of breaking into groups. In this case, we are breaking the vocabulary into a cluster of senses instead of the lexical clusters. This is prevalent in the second experimentation. Though, the second experiment was not extended to the Czech language, we received a significant improvement with the English data set. The results that we have achieved a encouraging for a combination of lexical and semantic feature experimentation. Again, the same methods as we have used for English can be again used for Czech.

6.1.3 *Work done by Øvrelid 2008*

[Øvrelid 2008]'s work is mostly augmenting linguistically motivated features on Maltparser. The research work tries to investigate if including linguistic features would improve the parsing accuracy for relations like subject and object, argument etc.. The basic feature extraction was done from the Swedish Talbanken corpus. Also, experiments were done with acquired features, that is the features like Animacy, Morpho-syntactic features etc. being acquired by the automatically tagged features for the given corpus. The author explains the theoretical motivation behind the acquisition of these features. Several taggers were used to extract the relevant feature information for the training dataset.

As explained before the experimentation was done with the Swedish Talbanken corpus. The experiment with Gold standard tags resulted in a significant improvement of the parsing accuracy. Our work in the thesis also targets a similar strategy for the third experiment. In case of English, the morphological set of features available was restricted, we used the fine-grained part of speech tags, this contains to an extent some specific linguistic features. This gave a significant improvement in the higher order parsing structures. We see here also, that the grandchild based parsing algorithms were always outperforming other forms of algorithms.

While, in the case of Czech, we had a big list of morphological features. But, when we included all the morphological features the results were not as good as the baseline scores. We then restricted this set of features to a smaller set (by looking at the relative linguistic importance), which improved the parsing accuracy. This is possibly due to the case of overfitting of data. That is the parser probably throws a lot of noise instead of the basic relationship of dependency links. This is mostly due to the fact that whenever a statistical model learns random error or noise instead of the underlying relationship of dependencies as the number of parameters now that is considered is comparatively large relative to the number of valid observations of the pair/triplet.

Discussion and Future Work

7

The thesis starts by explaining the graph based dependency parsing algorithms and their extensions that are also called the higher order structures. The objective of the thesis was to explore higher order parsing algorithms by augmenting several syntacto-semantic features. The above mentioned three papers acted as the inspiration for the experiments that were done. We have successfully established here that the higher order parsing structures are more responsive to the linguistic features. Also, we can see in the Appendix that there is a linear increase in the parsing accuracy after adding the features. The inclusion of Gold standard data was made to actually test the possibilities of the effect of the features on the parsing algorithms.

One of the most important things that we would be working next would be to experiment with the labelled accuracy score and to augment parser with wordsenses. The current parser doesn't score the labels. We would like to extend the parser to score both the labels and also integrate wordsenses in the parsing structure.

The overall approach is to augment each part and each dynamic-programming structure with senses and labels. Let us assume that the word senses can be represented as indices in the set $1, ..., S^*$ while dependency labels can be represented as indices in $1, ..., L$; here, S^* and L denote the total number of senses and labels. For every part now, we will have (h, m, l, s_h, s_m) that is s_h and $s_m \in S^*$ and $l \in L$ that is we need to embed both labels and senses with the parse structures.

A natural idea for future work is to evaluate the effects of combining several semantic features with the different parsing algorithms and building the specific set of features which increase the parsing accuracy for each language.

We can also investigate if increasing the order of the parsers - that is 4th order and higher would be interesting. Primarily because these would give us a range of orders for optimal parses. Though, this might be a difficult task due to the inherent computational complexity of both space and time problems.

Another interesting dimension of approach concerns the application of these parsers in several fields. The field of Machine Translation and automated summarization would especially need high end results. Also, since semantics plays an important role in these parsing frameworks, it might help to experiment with higher order parsers. Based on the interesting work by [Popel et al. 2011], it would be interesting to use these higher order parsers for Czech in TectoMT framework to extract better parse information.

As the higher order parsers are better in extracting the structural information, we could use these for the preliminary stages of various kinds of treebank tagging. Also, even if we have observed a positive change in the performance of dependency parsers, the best parsing techniques still fall short of almost accurate performance obtained by part of speech tagging etc., and hence present a promising area for future research.

8 Conclusion

In this work we provide insight into extension of feature set by augmenting some lesser known features. With the introduction of semantic and morphological features there is a significant improvement in the performance of parsing algorithms. We have seen the importance of some of these features individually.

We have shown in this thesis that the feature set in the dependency parsing algorithm need not be restricted to syntactic features only, the semantic features also add a lot of information to the parsing structure. This provides evidence that there is a relative improvement in unlabeled accuracy score whenever there is an inclusion of semantic features. The effects of semantic features are visible with higher order parsing structures.

Also, we see from the results that order plays an important part with languages which have rich morphology or inflectional languages, especially the grandchild parsing algorithms. Although we worked and presented our results only on two languages, our approach can be generalized to other languages and frameworks.

Finally, we hope that the work done in this thesis inspires the use of dependency parsers, especially the higher order dependency parsers in several tasks in the field of NLP. We also hope, this further increases the interest in research for better features.

Appendix

The graphs below show the performance of various parsers with the 10 runs on the validation dataset.

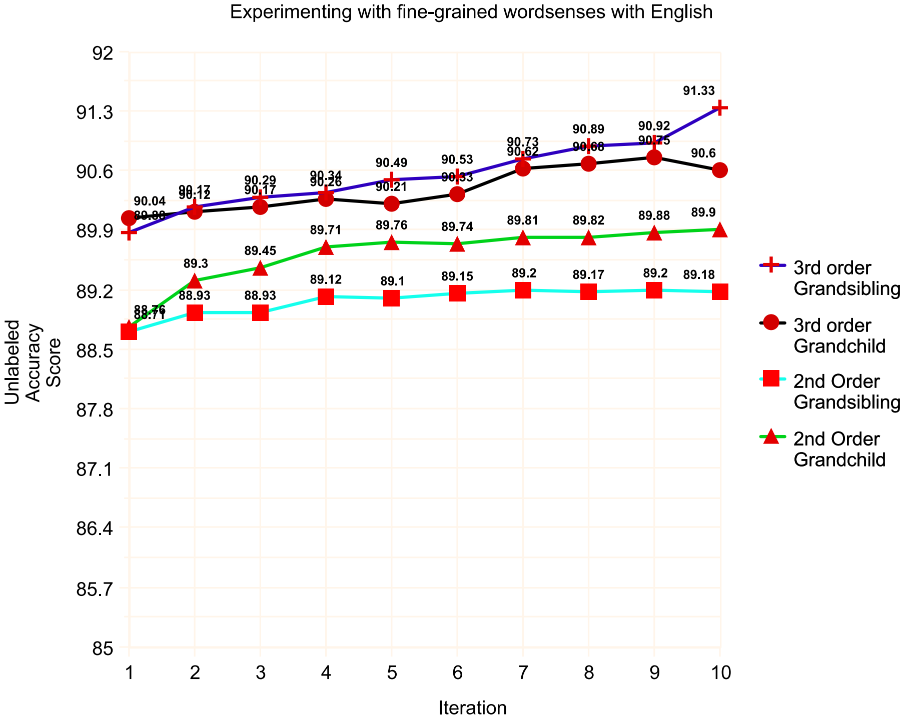

Figure 8.1 Result for experimentation with fine-grained wordsenses with English

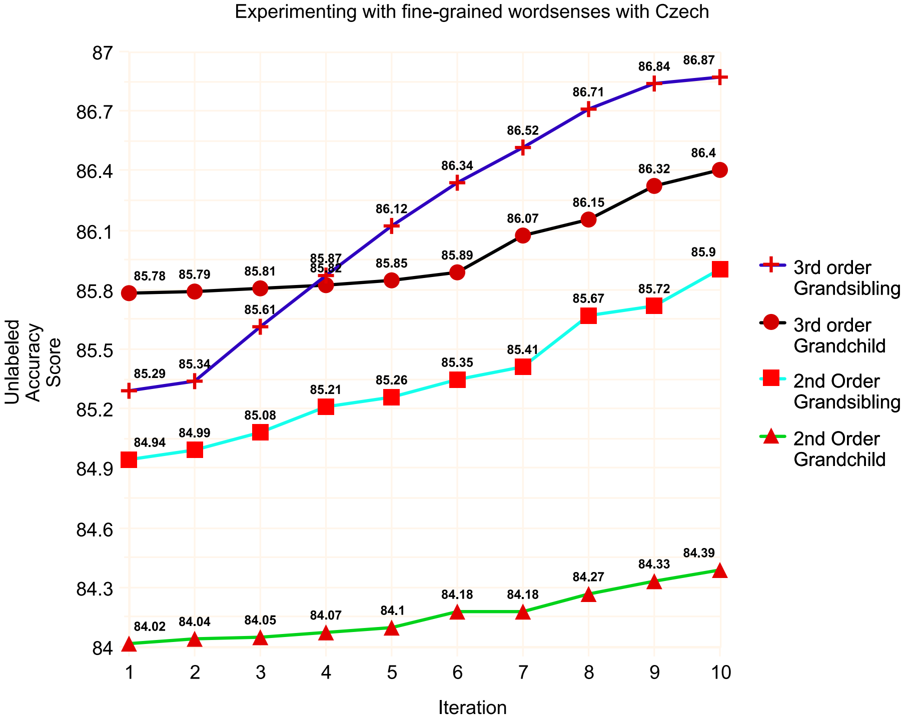

Figure 8.2 Result for experimentation with fine-grained wordsenses with Czech

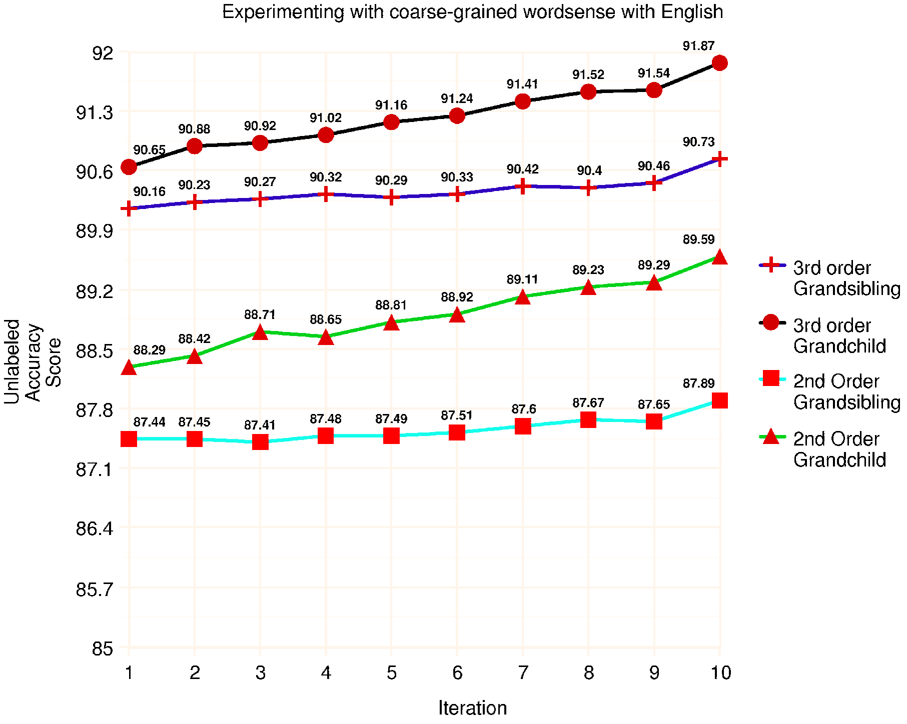

Figure 8.3 Result for experimentation with coarse-grained wordsense with English

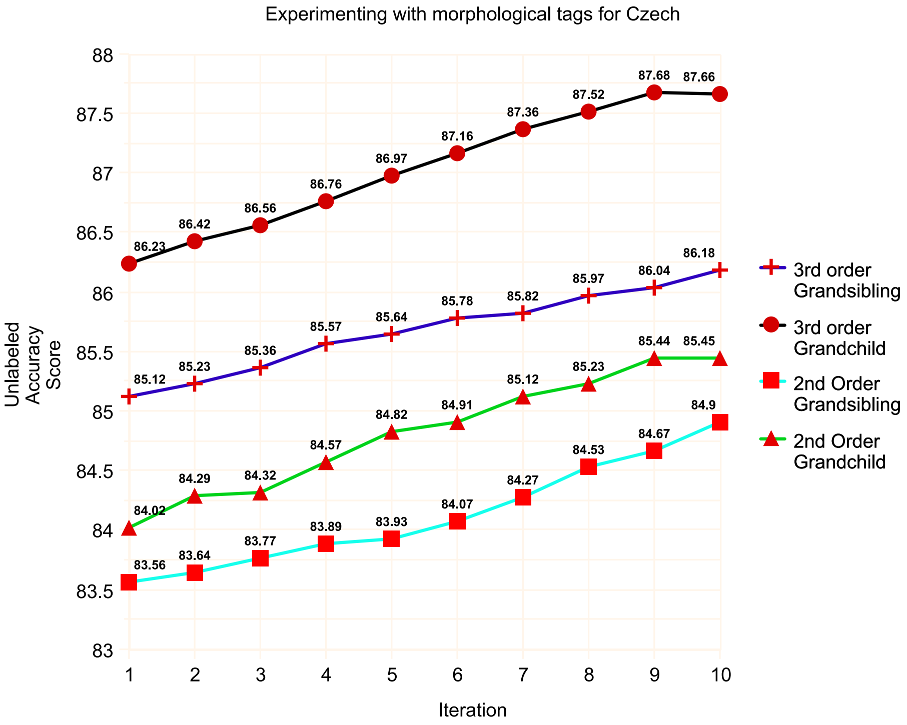

Figure 8.4 Result for experimentation with morphological tags for Czech

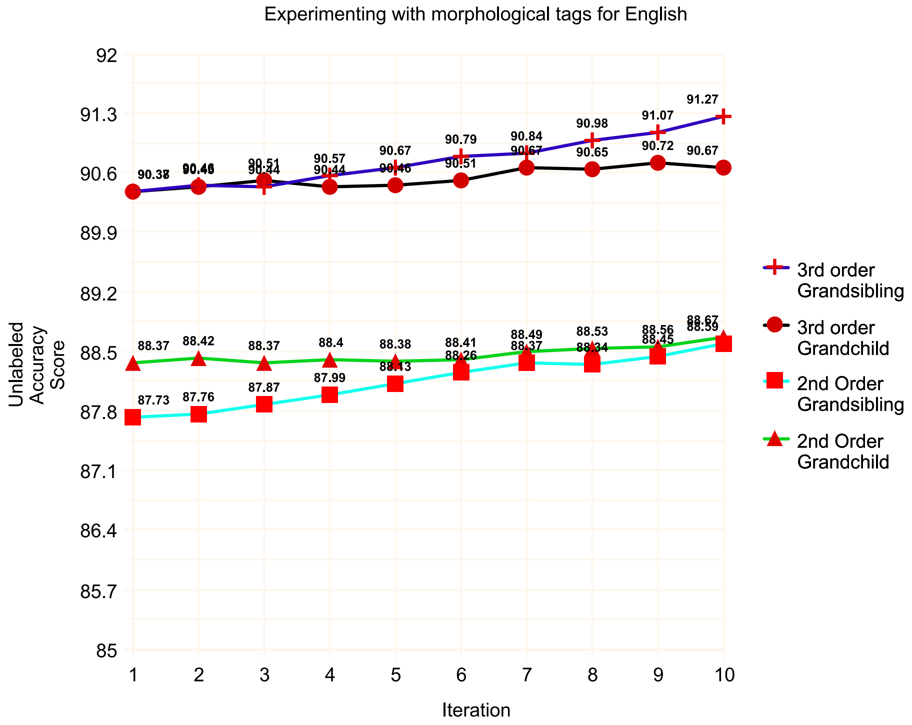

Figure 8.5 Result for experimentation with morphological tags for English

Bibliography

[Agirre et al. 2011] Agirre, E., Bengoetxea, K., Gojenola, K., & Nivre, J. (2011). Improving dependency parsing with semantic classes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, (pp. 699–703)., Portland, Oregon, USA. Association for Computational Linguistics. (Cited on pages 18, 22, 23, 34, 35, 37, 38, and 46.)

[Ambati et al. 2009] Ambati, B. R., Gade, P., Gsk, C., & Husain, S. (2009). Effect of minimal semantics on dependency parsing. In *Proceedings of the Student Research Workshop*, (pp. 1–5)., Borovets, Bulgaria. Association for Computational Linguistics. (Cited on page 19.)

[Ambati et al. 2010] Ambati, B. R., Husain, S., Nivre, J., & Sangal, R. (2010). On the role of morphosyntactic features in hindi dependency parsing. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, (pp. 94–102)., Los Angeles, CA, USA. Association for Computational Linguistics. (Cited on page 19.)

[Bangalore et al. 2009] Bangalore, S., Boulllier, P., Nasr, A., Rambow, O., & Sagot, B. (2009). Mica: a probabilistic dependency parser based on tree insertion grammars application note. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, (pp. 185–188)., Stroudsburg, PA, USA. Association for Computational Linguistics. (Cited on page 2.)

[Bikel 2004a] Bikel, D. M. (2004a). *On the parameter space of generative lexicalized statistical parsing models*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA. AAI3152016. (Cited on page 4.)

[Bikel 2004b] Bikel, D. M. (2004b). *On the parameter space of generative lexicalized statistical parsing models*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA. AAI3152016. (Cited on page 18.)

[Böhmová et al. 2001] Böhmová, A., Hajič, J., Hajičová, E., & Hladká, B. (2001). The prague dependency treebank: Three-level annotation scenario. In A. Abeillé (Ed.), *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers. (Cited on pages 28, 29, and 35.)

[Bourdon et al. 1998] Bourdon, M., Sylva, L. D., Gagnon, M., KHARRAT, A., Knoll, S., & MACLACHLAN, A. (1998). A case study in implementing dependency-based grammars. (Cited on page 2.)

[Buchholz & Marsi 2006a] Buchholz, S. & Marsi, E. (2006a). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, (pp. 149–164)., New York City. Association for Computational Linguistics. (Cited on page 21.)

[Buchholz & Marsi 2006b] Buchholz, S. & Marsi, E. (2006b). Conll-x shared task on multilingual dependency parsing. In *In Proc. of CoNLL*, (pp. 149–164). (Cited on page 27.)

[Carreras 2007] Carreras, X. (2007). Experiments with a higher-order projective dependency parser. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, (pp. 957–961)., Prague, Czech Republic. Association for Computational Linguistics. (Cited on pages 3, 11, 14, 17, 21, 25, and 26.)

[Carreras et al. 2008] Carreras, X., Collins, M., & Koo, T. (2008). Tag, dynamic programming, and the perceptron for efficient, feature-rich parsing. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, (pp. 9–16)., Manchester, England. Coling 2008 Organizing Committee. (Cited on page 22.)

[Carreras et al. 2006] Carreras, X., Surdeanu, M., & Màrquez, L. (2006). Projective dependency parsing with perceptron. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, (pp. 181–185)., New York City. Association for Computational Linguistics. (Cited on page 2.)

[Chomsky 1956] Chomsky, N. (1956). Three models for the description of language. *IEEE Trans. Information Theory*, *2*(3), 113– 124. (Cited on pages 2 and 5.)

[Cocke & Schwartz 1970] Cocke, J. & Schwartz, J. T. (1970). Programming languages and their compilers. Technical report, Courant Institute, NYU. Preliminary notes. (Cited on page 13.)

[Collins 2002] Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, (pp. 1–8). Association for Computational Linguistics. (Cited on page 10.)

[Covington 2001] Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In *In Proceedings of the 39th Annual ACM Southeast Conference*, (pp. 95–102). (Cited on page 2.)

[Eisner 2000] Eisner, J. (2000). Bilexical grammars and their cubic-time parsing algorithms. In H. Bunt & A. Nijholt (Eds.), *Advances in Probabilistic and Other Parsing Technologies* (pp. 29–62). Kluwer Academic Publishers. (Cited on pages 10, 13, and 21.)

[Eisner & Satta 1999] Eisner, J. & Satta, G. (1999). Efficient parsing for bilexical context-free grammars and head-automaton grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, (pp. 457–464)., University of Maryland. (Cited on page 7.)

[Eisner & Smith 2010 (Chapter 8)] Eisner, J. & Smith, N. A. (2010). Favor short dependencies: Parsing with soft and hard constraints on dependency length. In H. Bunt, P. Merlo, & J. Nivre (Eds.), *Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing* chapter 8, (pp. 121–150). Springer. (Cited on page 13.)

[Fellbaum 1998a] Fellbaum, C. (1998a). A semantic network of english: The mother of all wordnets. *Computers and the Humanities*, *32*(2-3), 209–220. (Cited on page 18.)

[Fellbaum 1998b] Fellbaum, C. (1998b). Towards a representation of idioms in wordnet. In *In Proceedings of the workshop on the Use of WordNet in Natural Language Processing Systems (Coling-ACL*, (pp. 52–57). (Cited on page 30.)

[Hajič et al. 2009] Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., & Zhang, Y. (2009). The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, (pp. 1–18)., Boulder, Colorado. Association for Computational Linguistics. (Cited on page 21.)

[Hektoen 1997] Hektoen, E. (1997). Probabilistic parse selection based on semantic cooccurrences. In *5th International workshop on parsing technologies (IWPT-97)*, (pp. 113–122). (Cited on page 18.)

[Horák et al. 2007] Horák, A., Pala, K., Duží, M., & Materna, P. (2007). Verb valency semantic representation for deep linguistic processing. In *ACL 2007 Workshop on Deep Linguistic Processing*, (pp. 97–104)., Prague, Czech Republic. Association for Computational Linguistics. (Cited on page 41.)

[Johansson & Nugues 2007] Johansson, R. & Nugues, P. (2007). Extended constituent-to-dependency conversion for English. In *Proceedings of NODALIDA 2007*, (pp. 105–112)., Tartu, Estonia. (Cited on page 28.)

[Joshi 1969] Joshi, A. K. (1969). Properties of formal grammars with mixed types of rules and their linguistic relevance. In *Proceedings of the 1969 conference on Computational linguistics*, COLING '69, (pp. 1–18)., Stroudsburg, PA, USA. Association for Computational Linguistics. (Cited on page 22.)

[Kasami 1965] Kasami, T. (1965). An efficient recognition and syntax algorithm for context-free languages. Technical Report AFCLR-65-758, Air Force Cambridge Research Laboratory, Bedford, MA. (Cited on page 13.)

[Kitagawa & Tanaka-Ishii 2010] Kitagawa, K. & Tanaka-Ishii, K. (2010). Tree-based deterministic dependency parsing — an application to nivre's method —. In *Proceedings of the ACL 2010 Conference Short Papers*, (pp. 189–193)., Uppsala, Sweden. Association for Computational Linguistics. (Cited on page 22.)

[Koo et al. 2008] Koo, T., Carreras, X., & Collins, M. (2008). Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, (pp. 595–603)., Columbus, Ohio. Association for Computational Linguistics. (Cited on pages 22, 23, 26, 40, and 47.)

[Koo & Collins 2010] Koo, T. & Collins, M. (2010). Efficient third-order dependency parsers. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (pp. 1–11)., Uppsala, Sweden. Association for Computational Linguistics. (Cited on pages v, 2, 3, 11, 15, 16, 21, 22, 25, 26, 27, and 29.)

[Kuboň et al. 1998] Kuboň, V., Holan, T., Oliva, K., & Plátek, M. (1998). Two useful measures of word order complexity. In *Proceedings of the Dependency-Based Grammars Workshop, the COLING - ACL Conference*. (Cited on page 7.)

[Marcus et al. 1993] Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, *19*, 313–330. (Cited on pages 1, 28, and 42.)

[Marneffe et al. 2007] Marneffe, M. C. D., Grenager, T., Maccartney, B., Cer, D., Ramage, D., Kiddon, C., & Manning, C. D. (2007). Aligning semantic graphs for textual inference and machine reading. In *In Proc. of the AAAI Spring Symposium at*. (Cited on page 2.)

[McDonald et al. 2005] McDonald, R., Crammer, K., & Pereira, F. (2005). Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, (pp. 91–98)., Ann Arbor, Michigan. Association for Computational Linguistics. (Cited on page 21.)

[McDonald et al. 2006] McDonald, R., Lerman, K., & Pereira, F. (2006). Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, (pp. 216–220)., New York City. Association for Computational Linguistics. (Cited on pages v, 3, 10, 14, 17, and 25.)

[McDonald & Nivre 2007] McDonald, R. & Nivre, J. (2007). Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, (pp. 122–131)., Prague, Czech Republic. Association for Computational Linguistics. (Cited on pages 3, 7, and 33.)

[McDonald et al. 2005] McDonald, R., Pereira, F., Ribarov, K., & Hajic, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, (pp. 523–530)., Vancouver, British Columbia, Canada. Association for Computational Linguistics. (Cited on pages 2, 10, 11, 17, and 26.)

[Mel'čuk et al. 1987] Mel'čuk, I., Pertsov, N., & Kittredge, R. (1987). *Surface syntax of English: a formal model within the meaning-text framework*. Linguistic & literary studies in Eastern Europe. John Benjamins Pub. Co. (Cited on page 2.)

[Merlo et al. 2011] Merlo, P., Bunt, H., & Nivre, J. (2011). Current trends in parsing technology. In N. Ide, J. Vronis, H. Bunt, P. Merlo, & J. Nivre (Eds.), *Trends in Parsing Technology*, volume 43 of *Text, Speech and Language Technology* (pp. 1–17). Springer Netherlands. 10.1007/978-90-481-9352-31$.(Citedonpages$ 1and 2.)

[Nilsson et al. 2006] Nilsson, J., Nivre, J., & Hall, J. (2006). Graph transformations in data-driven dependency parsing. In *Proceedings of the 21st International*

Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, (pp. 257–264)., Sydney, Australia. Association for Computational Linguistics. (Cited on pages 2, 3, and 5.)

[Nivre & Hall 2005] Nivre, J. & Hall, J. (2005). Maltparser: A language-independent system for data-driven dependency parsing. In *In Proc. of the Fourth Workshop on Treebanks and Linguistic Theories*, (pp. 13–95). (Cited on page 18.)

[Nivre et al. 2007a] Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., & Yuret, D. (2007a). The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, (pp. 915–932)., Prague, Czech Republic. Association for Computational Linguistics. (Cited on page 21.)

[Nivre et al. 2007b] Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., & Yuret, D. (2007b). The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, (pp. 915–932)., Prague, Czech Republic. Association for Computational Linguistics. (Cited on page 21.)

[Nivre & McDonald 2008] Nivre, J. & McDonald, R. (2008). Integrating graph-based and transition-based dependency parsers. In *Proceedings of ACL-08: HLT*, (pp. 950–958)., Columbus, Ohio. Association for Computational Linguistics. (Cited on page 3.)

[Novák & Žabokrtský 2007] Novák, V. & Žabokrtský, Z. (2007). Feature engineering in maximum spanning tree dependency parser. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, (pp. 92–98). (Cited on page 22.)

[Øvrelid 2008] Øvrelid, L. (2008). Linguistic features in data-driven dependency parsing. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL 2008)*. (Cited on page 47.)

[Pedersen & Kolhatkar 2009] Pedersen, T. & Kolhatkar, V. (2009). WordNet::SenseRelate::AllWords - a broad coverage word sense tagger that maximizes semantic relatedness. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session*, (pp. 17–20)., Boulder, Colorado. Association for Computational Linguistics. (Cited on pages 29, 30, 35, 39, and 45.)

[Popel et al. 2011] Popel, M., Mareček, D., Green, N., & Žabokrtský, Z. (2011). Influence of parser choice on dependency-based mt. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, (pp. 433–439)., Edinburgh, Scotland. Association for Computational Linguistics. (Cited on page 49.)

[Ratnaparkhi et al. 1994] Ratnaparkhi, A., Reynar, J., & Roukos, S. (1994). A maximum entropy model for prepositional phrase attachment. In *Proceedings of the workshop on Human Language Technology*, HLT '94, (pp. 250–255)., Stroudsburg, PA, USA. Association for Computational Linguistics. (Cited on page 18.)

[Sagae & Lavie 2006] Sagae, K. & Lavie, A. (2006). Parser combination by reparsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, (pp. 129–132)., New York City, USA. Association for Computational Linguistics. (Cited on page 3.)

[Sagae & Tsujii 2007] Sagae, K. & Tsujii, J. (2007). Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, (pp. 1044–1050)., Prague, Czech Republic. Association for Computational Linguistics. (Cited on page 21.)

[Sgall 1984] Sgall, P. (1984). *Contributions to Functional Syntax, Semantics and Language Comprehension*. Amsterdam, Netherlands/Academia, Czech Republic: Benjamins/Academia. (Cited on page 2.)

[Shimizu 2007] Shimizu, N. (2007). Structural correspondence learning for dependency parsing. In *In Proc.* (Cited on page 21.)

[Sleator & Temperley 1993] Sleator, D. D. & Temperley, D. (1993). Parsing english with a link grammar. In *Third International Workshop on Parsing Technologies*. (Cited on page 8.)

[Song et al. 2011] Song, Y., Wang, H., & Jiang, J. (2011). Link type based pre-cluster pair model for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (pp. 131–135)., Portland, Oregon, USA. Association for Computational Linguistics. (Cited on page 22.)

[Surdeanu et al. 2008] Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., & Nivre, J. (2008). The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, (pp. 159–177)., Manchester, England. Coling 2008 Organizing Committee. (Cited on page 21.)

[Tesnière 1959] Tesnière, L. (1959). *Éleménts de syntaxe structurale*. Paris: Klincksieck. (Cited on pages 1 and 6.)

[Titov & Henderson 2007] Titov, I. & Henderson, J. (2007). A latent variable model for generative dependency parsing. In *Proceedings of the Tenth International Conference on Parsing Technologies*, (pp. 144–155)., Prague, Czech Republic. Association for Computational Linguistics. (Cited on page 21.)

[Xiong 2005] Xiong, Z. (2005). Downstep effect on disyllabic words of citation forms in standard chinese. In *INTERSPEECH*, (pp. 1393–1396). (Cited on page 18.)

[Yamada & Matsumoto 2003] Yamada, H. & Matsumoto, Y. (2003). Statistical dependency analysis with support vector machines. In *In Proceedings of IWPT*, (pp. 195–206). (Cited on page 2.)