Combining Part of Speech Induction and Morphological Induction

Charlotte Wilson

Supervisor: Steven Bird

Honours Thesis
Department of Computer Science and Software Engineering
University of Melbourne

November 2004
Abstract

Linguistic information is useful in natural language processing, information retrieval and a multitude of sub-tasks involving language analysis. Two types of linguistic information in all languages are part of speech and morphology. Part of speech information reflects syntactic structure and can assist in tasks such as speech recognition, machine translation and word sense disambiguation. Morphological information describes the structure of words and has application in automated spelling correction, natural language generation and information retrieval for morphologically complex languages.

Machine learning methods in natural language processing acquire linguistic information from corpora of natural language text. While supervised learning algorithms are trained on texts that have been annotated with linguistic features, induction algorithms learn linguistic information from unannotated corpora. Such algorithms avoid any requirement for linguistically annotated training data - a resource that is highly time-intensive to produce. However, in learning from unannotated corpora, only limited sources of information are available. In practice, part of speech induction methods usually learn from distributional evidence about the contexts in which words occur. In contrast, morphological induction methods tend to be based on the orthographic structure of the words in the corpus. However, a word’s morphological form and syntactic function often correlate: a word’s morphology may indicate its syntactic function and vice versa. Thus, both distributional and orthographic evidence may be useful for both tasks.

This thesis investigates the extent to which the information induced by one learner can be used to bootstrap the other: specifically, whether the incorporation of explicit annotations from one learner can improve the performance of the other.
# Contents

1 Introduction .......................... 1

2 Linguistic Background .................. 3
   2.1 Morphology .................................. 3
   2.2 Theories of Morphology .................. 4

3 Literature Review ....................... 5
   3.1 Part of Speech Induction .................. 5
   3.2 Morphological Induction .................. 6
       3.2.1 Using a Knowledge Source to Bootstrap .... 6
       3.2.2 Knowledge-free Discovery of Word Segments ... 7
       3.2.3 Knowledge-free Discovery of Morphologically Related Word Forms ... 9
       3.2.4 Discovering Allomorphy ................. 11
   3.3 Towards Combining Part of Speech Induction and Morphological Induction .... 12

4 Methodology .................. 13
   4.1 Part of Speech Induction .................. 13
       4.1.1 Class-Based Bigram Language Models .... 13
       4.1.2 Incorporating Morphological Annotation ... 16
       4.1.3 Evaluation .................................. 16
       4.1.4 Part of Speech Tagging .................. 17
   4.2 Morphological Induction .................. 17
       4.2.1 The Theory of Whole Word Morphology .... 17
       4.2.2 Whole Word Morphologizer ............... 18
       4.2.3 Incorporating Part of Speech Tags .......... 18
       4.2.4 Evaluation .................................. 19
       4.2.5 Morphological Annotation ............... 21
   4.3 Combining Part of Speech Induction and Morphological Induction .......... 22

5 Experiments .................. 23
   5.1 Incorporating Morphological Annotation into Part of Speech Induction .... 23
       5.1.1 Morphology as an Indication of Part of Speech .... 23
       5.1.2 Experiments .................................. 24
       5.1.3 Results .................................. 24
   5.2 Incorporating Part of Speech Tags into Morphological Induction .......... 27
       5.2.1 Experiments .................................. 27
       5.2.2 Results .................................. 28

6 Conclusions and Future Work .... 32
   6.1 Conclusion .................................. 32
   6.2 Future Work .................................. 33
Chapter 1

Introduction

Many tasks in natural language processing and information retrieval make use of linguistic information. As a result there is a need for automated learning of this information. For this task a sample corpus of natural language or training data is required from which a system can learn. Training data can take many forms: a speech signal, phonetically transcribed text, orthographic text etc.. Supervised learning techniques learn from training data that has been marked up with annotations indicating linguistic features or structures. Whilst a variety of annotated training corpora are quite widely available for many of the more commonly used languages and linguistic structures, they are time-intensive to produce. As a result such resources are not widely available for many of the world’s languages or for less commonly investigated linguistic structures.

In contrast to supervised learning methods, induction algorithms learn from unannotated corpora, avoiding the requirement for annotated training corpora. However, as a result such methods are provided with much more limited sources of information from which to learn. In general, this is limited to information about the structure of the words themselves (their form), the distribution of words and their frequency in the corpus.

This thesis investigates the induction of two kinds of linguistic information: part of speech (POS) and morphology. A word’s part of speech indicates its syntactic function within a sentence, e.g. noun, verb etc. Morphology describes how words are constructed from smaller units of meaning (morphemes). For example, the English word talked consists of two morphemes: talk- and -ed. Morphological analysis is the process of identifying the morphological structure of words and the systematic ways in which word forms are created in a language.

Part of speech induction systems learn an assignment of part of speech information to words. Part of speech tagging is the process of assigning an appropriate part of speech marker (or tag) to each word in a corpus of natural language. Part of speech tagging is often not an end in itself, but is an important preprocessing step for a variety of natural language processing tasks. For example, part of speech information is useful in determining how a word should be pronounced in a text-to-speech system: CONTENT (the noun) has different stress patterns to content (the adjective). In a speech recognition system part of speech information can help to predict which words are likely to occur next from a knowledge of syntactic structure (through language modelling). Part of speech information is also used in word sense disambiguation, machine translation systems and as an important part of general corpus-based linguistic research. Part of speech induction refers to the unsupervised learning of part of speech groupings (tags) from an untagged corpus.

Morphological analysis is another important component of a number of language processing tasks. At a basic level, a knowledge of morpheme boundaries assists with automatic text hyphenation. Rules governing how morphemes can be combined to form words (morphotactics) are a key component of spelling correction systems. Also, lemmatization relies on morphological analysis to reduce a set of lexical forms with the same stem to one common canonical
form. Lemmatization is used in tasks such as information retrieval in which both the search term and terms contained in documents are lemmatized to potentially increase the number of matches. Morphological analysers are also used in machine translation systems to identify morphosyntactic information contained in a source language word form, which can assist in generating appropriate target language structures.

In recent years, there has been a proliferation of research into computational methods for the unsupervised learning of morphology (morphological induction). Models that are able to learn the morphology of natural languages are “of great practical interest as tools for descriptive linguistic analysis and for minimizing the expert resources needed to develop morphological analyzers and stemmers.” (Baroni et al., 2002).

In general, part of speech induction methods learn from distributional evidence about the contexts in which words occur. In contrast, most morphological induction methods are based on orthographic evidence from the structure of the words in the training text. However, the syntactic function and morphological form of a word often correlate. For example, in English, verbs often end in -ing and a word ending -tion is usually a noun. This indicates that there may be some value in combining the two induction procedures and incorporating both distributional and orthographic information into both methods.

This thesis investigates the effect of combining the two sources of evidence by bootstrapping the two induction methods such that one provides annotated training data for the other. That is, a part of speech induction system is used to tag training data for the morphological induction system and vice versa. The training data used takes the form of orthographic text. By combining the two induction methods in this way, it is possible to incorporate the information that is learnt by one into the other. This means that the contextual information from which the part of speech induction system has learnt is incorporated into the morphological induction system. Similarly, the orthographic information used by the morphological induction system is incorporated into the part of speech induction system.

There are a number of areas that this study aims to address. Firstly, contextual evidence for rare words in the training text is limited. This is a problem for part of speech induction systems that are based on such evidence. However, additional evidence can be provided by a morphological induction system that should improve the accuracy of the assignment of such words to a particular part of speech. In contrast, the orthographic content of words provides no indication of grammatical function or meaning. Thus, many morphological induction systems discover morphemes, but are unable to associate meaning or syntactic function with these forms. As a result the accuracy of the induced morphological analyses is questionable and their application is limited. Part of speech information can clearly go some way towards alleviating this problem.

This thesis is structured as follows. Chapter 2 provides some linguistic background particularly focused on morphology. Chapter 3 presents a review of the relevant literature regarding part of speech induction, morphological induction and previous work that has combined the two. In Chapter 4 the methodology used in this thesis is presented. This is divided into two sections - the part of speech induction component and the morphological induction component. An overview of how the two components are combined is also given in this chapter. Chapter 5 discusses the two experiments conducted for this study and finally Chapter 6 will draw some conclusions and discuss future work.
Chapter 2

Linguistic Background

2.1 Morphology

Morphology has two distinct functions in a language: derivational morphology produces new words from existing words with a different meaning and grammatical function. For example, the noun government is formed from the verb govern and the verb help is the basis for the adverb unhelpfully. The second function of morphology is to indicate or agree with syntactic features, such as tense (walk, walked, walks), number (girl, girls), gender (actor, actress) and case (I, me, my). These syntactic features vary from language to language but do not change a word’s underlying meaning or part of speech. This kind of morphology is called inflectional morphology.

Across the world’s languages there is a great deal of typological variation between morphological systems. Some languages are isolating, in which every word is composed of a single morpheme. Others are polysynthetic languages in which a single word may be composed of many morphemes, and in fact a whole sentence may be a single morphologically complex word. The majority of languages lie somewhere between these two extremes, expressing some information morphologically and some syntactically through word order.

A further typological distinction can be made between agglutinative and fusional morphology. Agglutinative morphology involves morphemes, each with a distinct meaning or function, being added to a word one after the other (or glued on). In contrast in fusional morphology, a single morpheme can have multiple meanings or functions and cannot be divided into smaller components corresponding to the distinct components of meaning.

A common type of morphology found in many European languages is concatenative morphology. Concatenative morphology takes the form of affixes including prefixes, suffixes, circumfixes (a single morpheme realised by a prefix and a suffix) and infixes (affixes internal to a word). Suffixes are the most common form of concatenative morphology. Morphology can also take many other forms and often it is not possible to segment words into their component morphemes. For example morphology can be expressed by stress and tone, by vowel alternations, known as ablaut (e.g. sing, sang, sung) or have no form at all (e.g. a cut, to cut). An extreme example of this is suppletive morphology: in which morphologically related words show no similarity in their form. For example, go and went have the same relationship to one another as walk and walked, yet they show no obvious similarity in their form.

It is also quite common for the same morpheme to be realised in different ways (take different forms). This variation in form is known as allomorphy and the variants of the one morpheme are called allomorphs. Variation is conditioned by the environment in which the morpheme occurs: it can be phonologically, grammatically or lexically conditioned. That is, it can depend respectively on the surrounding phonological or grammatical environment or the lexeme (word) itself.
2.2 Theories of Morphology

Some of the fundamental concepts of traditional morphological theory were introduced in the section above. Under this view words are divisible into atomic units of meaning called morphemes. The surface variants of morphemes are called allomorphs and the rules that govern their combination are morphotactics. This view of morphology is known as *item and arrangement* morphology. This theory has some difficulty explaining certain aspects of morphological systems: in particular cases where there is not a one to one correspondence between the form of a morpheme and its meaning. For example, in non-concatenative morphology a word is not divisible into distinct units of meaning. This is a serious limitation of item and arrangement morphology.

An alternative theory of morphology is known as *word and paradigm* morphology, in which paradigms are groups of morphologically related words and morphological processes transform one word into another word in the same paradigm. As a simple example, English has a *+ed* morphological process that operates on a present tense verb such as *walk* to transform it into the corresponding past tense form *walked*. In this view, morphologically complex words are composed of a variable and a non-variable component. These components do not have to be contiguous, nor do they have to be associated with distinct meaning. In general, the word and paradigm view of morphology accounts for forms that cannot be decomposed into these component parts (such as suppletive forms) by analysing them as expressing morphological relationships that occur in very specific environments.
Chapter 3

Literature Review

This chapter presents an overview of previous work relevant to the current study. Firstly some approaches to part of speech induction are discussed and then a number of methods for morphological induction. Finally, previous work in combining the two induction systems is presented.

3.1 Part of Speech Induction

Part of speech induction systems usually model the distribution of words in a text and deduce the patterns of linguistic structure from these models. While distributional evidence for the syntactic function of frequent words is strong, for rare words it is limited. As a result many methods do not attempt to discover part of speech information for infrequent words and are limited in their coverage and applicability.

Most methods for the unsupervised learning of part of speech information are based on clustering algorithms that partition words into groups, according to some measure of similarity between the members of a group. Words are described in terms of features and are often formalised as a feature vectors, which define points in some multi-dimensional space. Cluster membership is then indicated by the distance between a feature vector and the centroid of a cluster. The clustering process is iterative and in its simplest form, involves starting with randomly assigned centroids of clusters, then repeatedly allocating feature vectors to a cluster and recalculating the centroid of each cluster. The iteration stops when the centroids become stable. The Expectation Maximisation algorithm discussed below is one such clustering algorithm.

Schutze (1993) describes a clustering algorithm for inducing part of speech categories for the 5,000 most frequent words in a corpus. He uses cooccurrence statistics between pairs of these words within a context window of two words to either side as features. The large resulting feature space is then reduced using singular value decomposition, a technique for obtaining an optimal lower-dimensional representation for the large sparse feature vectors. This dimension reduction dramatically decreases the time taken to cluster the feature vectors and Schutze also uses hypothesised word class information (rather than the word form itself) as an alternative feature set. This method presents a compact representation of high-dimensional vector space, increasing the applicability of clustering techniques in sparse high dimensional vector space to large vocabularies.

In their discussion of the application of n-gram models to natural language processing, Brown et al. (1992) present a method for inducing syntactic categories from large amounts of data. Their method also works iteratively, starting by assigning each word to its own cluster and then repeatedly merging the pairs of words which will result in the largest decrease in language model perplexity. The perplexity is calculated according to model of the data assigned by the current iteration and indicates the accuracy of the predictions of the language model. A smaller perplexity indicates greater certainty in the model’s predictions. They are able to adapt the method to be applicable to large corpora through a modification involving initially ordering
the words by frequency of occurrence in the corpus and assigning only the top $C$ words to their own clusters, where $C$ is the number of clusters required. The remaining words are then merged into these clusters one by one, minimising the perplexity of the language model each time. Thereafter the algorithm proceeds as for their basic method.

3.2 Morphological Induction

In recent years there has been a great deal of interest in methods for the unsupervised learning of the morphology of natural languages. Various approaches have been taken to this task: most of which have focused on simple models of concatenative morphology either addressing the task of identifying morpheme boundaries, obtaining affix inventories, identifying the morphemes that make up a word or identifying morphological relationships between words. More complex structures involving multiple affixes combined in specific ways are not generally discovered by morphology induction systems. Typologically, this concatenative model of morphology is quite limited and could not begin to include the wide variety of morphological systems that exist in the languages of the world. Nevertheless it has been quite successfully applied to some of the many languages that do exhibit this model of morphology.

While typological differences between morphological systems present a challenge for the development of universally applicable morphology learners, even within certain typological constraints, morphology presents a number of challenges: in particular, the learning of morphological grammars that model the different functions of morphemes and the identification of allomorphy. Certainly, it is desirable to be able to associate grammatical function with the morphological forms that have been discovered such that the model can be generalised to unseen data. Without this association, the discovery of morphological structures is of limited use.

A summary of morphological induction systems is presented below: some of these systems take advantage of a knowledge source in addition to a morphologically unannotated corpus, other knowledge-free systems segment words into candidate morphemes and a final class of knowledge-free systems discover sets of morphologically related words. Finally, some methods for the discovery of allomorphy and phonological rules are discussed.

3.2.1 Using a Knowledge Source to Bootstrap

A number of morphological induction systems have incorporated one or more additional knowledge sources. These knowledge sources have taken various forms: amongst them part of speech information.

Gaussier (1999) presents an algorithm for inducing derivational morphology from an inflectional lexicon containing part of speech information. He concentrates only on suffixation and locates stems that share at least the same first five letters, to identify potential pairs of related word forms. Part of speech information provides evidence that suffixes share a common function and thus that morphological relationships can be generalised to apply to multiple stems. He then uses hierarchical agglomerative clustering to group words into relational families: for example *deprecate, deprecation, deprecator, deprecative*.

As Goldsmith (2001) mentions in his discussion of Gaussier’s work, a model of morphology based solely on orthographic similarity presents challenges when attempting to cluster related word forms into paradigms. This is particularly difficult in an unsupervised method with no access to part of speech information. Goldsmith describes the problem of how to disambiguate the nominal suffix *-s* from the verbal *-s*, for example, and ensure that these suffixes are not assigned to single clusters. As he mentions, “it is very difficult to find a statistical and morphological basis for this knowledge” (Goldsmith, 2001).

Jacquemin (1997) describes a method that uses a corpus and a list of multiword terms to identify morphologically related word forms. He incorporates the idea that morphological variants of words that form multiword terms are likely to occur as collocates within a certain
window of one another in a corpus. He approaches the task in the context of improving the morphological analysis component of an information retrieval system, however, there are a number of issues centred in automated discovery of derivational morphology which he also aims to address with his method. The first is the ability to distinguish between those orthographically similar words that are morphologically related and those that are not (e.g. *imported* and *important*: words that share no synchronic meaning in common). Conversely, he also aims to be able to distinguish between morphologically related words whose orthographic similarity does not conclusively indicate the existence of this relationship (e.g. *give* and *gift*). His final aim is the discovery of rare or domain-related morphological relations.

Jacquemin’s algorithm uses a measure of orthographic similarity between words, extending this to the identification of potentially morphologically related collocates and grouping sets of terms that have the same morphology. He then clusters the groups together, such that different variants of the one derivational suffix are grouped together. The clustering procedure uses the orthographic similarity of suffixes as a measure of whether two suffixes express the same morphological relation, relying on the intuition that, “suffixes corresponding to similar linguistic operations tend to have common final strings” (Jacquemin, 1997). This algorithm presents a novel approach to using the idea that semantic relatedness is indicated by cooccurrence within a small window in a morphological learning system.

Yarowsky & Wicentowski (2000) present a minimally supervised method for the induction of English past tense forms. Their method requires a large unannotated text corpus, a list of the open class roots in the language, a table of the language’s inflectional parts of speech, the canonical suffixes for each part of speech and a list of the consonants and vowels of the language. They aim to induce a morphological analyser for both regular and irregular morphological processes in a language. Their approach is to treat morphological analysis as predominantly an alignment task and they combine various alignment procedures to identify morphologically related forms. These include the relative frequency of related forms in the corpus, contextual similarity, weighted edit distance and morphological transformation probabilities (as constructed from the other alignment procedures). Their method uses various knowledge sources and thus does not fall under the umbrella of unsupervised learning. However it is noteworthy for the variety of different sources of evidence (including part of speech) that it incorporates into identifying morphologically related word forms. They report very high accuracy in their results.

Neuvel and Fulop (2002) describe their system, Whole Word Morphologizer, that takes a part of speech tagged lexicon as input and learns the morphological relationships between words. The method is based on a string matching procedure that identifies common beginnings or endings of words and the orthographic and syntactic environments in which they occur. Through modifications to the string matching procedure, this approach has the potential to be generalised to apply to a wide variety of morphological systems, not just the concatenative model. Although they avoid calling the similarities between word forms *morphemes*, these similarities do effectively express the realisations of allomorphy and phonological conditioning within the morphological system of a language.

Semi-supervised methods for morphology learning have incorporated a variety of different sources of information to supplement unlabelled training data. Many of these leverage explicit part of speech information and others use different sources of contextual information. This information enables the discovery of morphology not only through orthographic similarity, but also through similarity in syntactic function.

3.2.2 Knowledge-free Discovery of Word Segments

The first method for locating the boundaries between morphemes was proposed by Harris (1955), who suggested that the predictability of the $n+1$th phoneme based on the previous $n$ phonemes, is an indication of likely morpheme boundaries. That is, the less likely the next phoneme (according to the model of the corpus created thus far), the more likely we are at a
morpheme boundary, or, formulated another way, the larger the distribution of possible next letters, the more likely we are at a morpheme boundary. Harris’ approach was later formalised in information theoretic terms and implemented using various measures by Hafer and Weiss (1974), such that local peaks of conditional entropy indicate morpheme breaks. This is a similar approach to the Prediction by Partial Matching (PPM) method commonly used in text compression systems, in which the probabilities in the model used to encode the text are adjusted as the data is processed letter by letter. There are clearly a number of ways in which this approach over-simplifies morphological complexity. Nevertheless, it has been shown to represent a reasonable heuristic for finding candidate morphemes in a corpus.

Indeed, Dejean (1998) uses Harris’ method as a first step for his method for the discovery of morphemes, assuming that a morpheme boundary exists where the count of the number of different letters following the current sequence of letters exceeds some threshold value. He then identifies further morphemes by searching for occurrences of the morphemes already identified and including any strings that occur in their place. From this he segments the words into morphemes, using the longest morpheme that matches the beginning or end of a word.

Many approaches to text segmentation are based on the Minimum Description Length (MDL) framework, in which the model of a corpus that results in the shortest description length is chosen to be the best fit (Rissanen, 1989). The description length of a corpus is calculated by summing the information content of each of the segments in the corpus, plus each of the segments in the lexicon (or codebook). Information theory states that the information content of a segment can be calculated by taking the inverse logarithm of the probability of a segment, where the probability of a segment is given by the number of occurrences of the segment divided by the total number of segments in the corpus. These approaches combine MDL with a search procedure which aims to optimise the segmentation of the corpus in respect to the description length.

De Marcken (1995) applied MDL to the problem of correctly segmenting a sequence of characters into words. This problem is particularly applicable for languages such as Chinese, in which word boundaries in text are unmarked, as well as to the segmentation of unbroken speech streams and to theories of language acquisition. De Marcken’s algorithm first assumes that each character forms a segment and iteratively improves the segmentation of the data, by gradually building up larger segments that reduce the MDL score for the model. This method is quite successful in locating word boundaries, however, it is important to note that the word internal segments that are located do not necessarily correspond to morphemes. Rather, this procedure locates commonly occurring strings that may not be associated with a distinct unit of meaning (for example in English this might include strings such as th or str).

Brent et al. (1995) also use MDL to discover the most common suffixes from a corpus, assuming all words are made up of a stem followed by a possibly empty suffix. They first use a heuristic, based on the ratio of the relative frequency of the sequence of letters in the suffix divided by the relative frequencies of its letters, to rank the candidate suffixes. Then they use a greedy search procedure amongst the possible suffixes to add or remove candidate suffixes from the suffix list. Most of the top ten suffixes that they identify are indeed English morphemic suffixes. However, they also describe a second model that is not completely knowledge-free, incorporating part of speech information, and this achieves slightly better results.

Kazakov (1997) uses MDL as a fitness metric in a genetic algorithm to find the globally optimal morphological analysis for a training corpus. This approach is also limited to a model of morphology in which words can be divided into exactly two parts: a stem and a suffix. Potential morpheme boundaries split the lexicon into a list of stems and a list of suffixes with the count of the number of letters in the two lexicons being minimised, to obtain the most compact representation for the data.

Snover (2002) also describes an algorithm for morphological induction based on the MDL framework, but interestingly, he adds a model of morphological paradigms into the basic stem
and suffix model. He defines paradigms as sets of suffixes that attach to the same stems and restricts each stem to be a member of exactly one paradigm. Although this limits the ability of his model to accurately reflect some aspects of morphological systems (e.g. the suffix -s in English occurs in both the nominal and verbal paradigms), the incorporation of paradigmatic structures into a model of morphology is an important feature of this system. In effect paradigms represent groups of suffixes that apply to words with a common syntactic function or functions. He reports improvement in the results when a model including paradigmatic information is used.

Most induction methods for the discovery of word segments use statistical models of the orthographic structure of words to discover and evaluate segmentations. These methods make no attempt to associate segments with their function or to identify the syntactic relationships that morphology may indicate. Snover’s work is a notable exception in its incorporation of paradigms (related suffixes) into a statistical model. He groups suffixes as applicable to certain groups of words: an approximation of words with a common syntactic function (part of speech classes). His results indicate that the incorporation of syntactic information can be advantageous for word segmentation. Brent’s experiments incorporating explicit part of speech information into a statistical model, also lead to improved performance. This is a further indication of the potential value that part of speech information can provide to morphological induction systems.

3.2.3 Knowledge-free Discovery of Morphologically Related Word Forms

Goldsmith (2001) presents a system known as Linguistica for learning of the morphology of stem + suffix languages which is commonly acknowledged as the state of the art in unsupervised morphological learning. Like many other approaches to morphological induction, his inductive procedure is based around word segmentation and the MDL hypothesis.

Goldsmith applies MDL as follows: given a model of the data, he uses information theoretic measures to calculate both the size of the model and the size of the data when compressed using the model. The optimal model for the data is then the one for which the sum of these two sizes is minimised. In effect, this hypothesis aims to find the simplest model that is the best fit for the data.

The algorithm can be broken down into three steps as follows;

1. Create an initial grammar (morphological analysis), by segmenting words.

2. Compute the description length of the grammar and the corpus when compressed using this grammar as a model. If the grammar has improved the resulting description length, replace the old grammar with the modified grammar.

3. Modify the grammar and return to step 2 to evaluate the modification.

An initial grammar is constructed using the Expectation Maximisation (EM) algorithm and a number of bootstrapping heuristics. The EM algorithm aims to maximise the probability of the segmented data, by iteratively assigning possible splits to the data. The first heuristic is incorporated with the aim of ensuring that the candidate segments are relatively long (rather than suffixes consisting of a single letter) and involves weighting the probability distribution to favour longer segments. A second heuristic locates candidate suffixes by maximizing the likelihood of the word-final n-grams across the corpus. This segmentation procedure models the data into stems, suffixes and signatures, where a suffix must occur more than once in the corpus and a signature is a structure that associates each stem with exactly one set of suffixes.

The description length of this grammar is then given by the compressed length of the model plus the length of the corpus when compressed using this model. The length of the model is given by:
• length of stem list = length of each stem + length of the information specifying the length of the list

• length of the suffix list = length of each suffix + length of the information specifying the length of the list

• length of a signature = sum of pointers to stems + sum of pointers to suffixes + length of the information specifying the number of stems and suffixes

The description length of the corpus when compressed using this model is calculated from the probabilities that this model assigns to each word in the corpus, which are used to calculate the information content of each word and then summed together. The probability of a word \( w \) is the product of the probability of \( w \)'s signature, the probability of \( w \)'s stem given its signature and the probability of \( w \)'s suffix given its signature. That is, words are assigned a probability based on the model, rather than their empirical frequency in the corpus. This measure results in the compressed corpus being longer than it would be if it were modelled according to word frequency alone, however the size of the model has been significantly decreased by the inclusion of signatures, relating stems to suffixes. Goldsmith’s model also allows a word form to be decomposed into a recursive structure consisting of \(((stem + suffix) + suffix)\), which effectively allows for a word form to contain multiple suffixes. This addition increases the complexity of the model (as stems have a flag indicating whether they point to a simple or complex stem structure), while increasing its expressive power.

Goldsmith employs a number of heuristics to decrease the description length with the aim of improving morphological analyses. These include checking suffixes to see if they can be broken down into the concatenation of two (or more) suffixes (e.g. -ments is reanalysed as -ment + -s), reanalysing a set of suffixes that form a signature if they all begin with the same letter (e.g. te.ing.ts becomes e.ing.s) and a number of methods to explore the value of signatures containing a small number of stems or a single suffix.

Goldsmith’s system is unique in that it attempts to establish categories of stems that accept the same suffixes. The creation of these categories goes some way towards generalising the function of certain suffixes and associating suffixes with the forms to which they can be applied. As the orthographic form of a suffix indicates a syntactic function or meaning, this is a first step towards associating morphology with its syntactic function. It is interesting to note that Goldsmith does not incorporate contextual information from the training text to assist either in the discovery of morphology or its function.

Schone and Jurafsky (2001a) describe a knowledge-free induction procedure for discovering sets of morphologically related words from a large corpus. These conflation sets contain words that are inflected or derived from the one root form: for example, “the conflation set of the word abuse would contain abuse, abused, abuses, abusively, and so forth” (Schone and Jurafsky, 2001a). Their method combines evidence of various kinds, including orthographic, semantic and syntactic. Orthographic evidence is tailored to an affix-based model of morphology and includes prefixes, suffixes and circumfixes. It is also worth noting that they incorporate affix frequency into their measure for identifying candidate affixes, potentially limiting its ability to identify rare or irregular forms. Unlike many of the methods discussed above orthographic evidence is taken only as a loose indication of a morphological relation and the other kinds of information are used to refine the list of possibly related words. Similar local contexts indicate syntactic similarity and provide evidence of morphological relationships.

Baroni et. al. (2002) present a method for unsupervised discovery of pairs of morphologically related words from a corpus of text, which produces a ranked list of morphologically related pairs of words. Like Schone and Jurafsky they combine measures of orthographic and semantic similarity to score pairs of words as morphologically related. However, they measure these characteristics differently: using minimum edit distance to measure orthographic
similarity and a mutual information-based measure for semantic similarity. Their algorithm is not limited to a model of concatenative morphology, nor do they attempt to segment words into distinct morphemes based on the distributional properties of word substrings, rather they aim to “discover pairs that are related by rare and/or non-concatenative morphological processes” (Baroni et al., 2002). Their orthographic similarity score does not attempt to model the frequency of the pattern of morphological relation and thus a pair of morphologically related word forms that occur once in the training corpus can have the same orthographic similarity score as a much more commonly occurring morphologically related pair. The mutual information score of semantic similarity is based on the assumption that morphologically related words occur near one another. This contrasts with the approach to measuring semantic similarity taken by Schone and Jurafsky (2001a), who consider words that occur in similar contexts to be semantically related. It is unclear which provides a better measure and whether, “the two approaches produce complementary or redundant results” (Baroni et al., 2002).

Baroni et al. do not attempt to generalise the relationships between pairs of related words by extracting the patterns that relate the words. However, the authors do suggest that the output of this algorithm might be appropriate as a preprocessing step to provide input for such a rule induction procedure. There is also mention of future work to extend the pairs by transitivity (much like Schone and Jurafsky) and create larger sets of morphologically related words.

Unsupervised methods for the discovery of morphologically related words have incorporated various types of information as evidence for their hypotheses. Where it is used, syntactic evidence has taken the form of implicit contextual evidence from the training data. While the success of Goldsmith’s Linguistica indicates that the sole use of orthographic evidence can provide a strong foundation for morphological discovery, Schone and Jurafsky have shown that further evidence (including syntactic evidence) can also be helpful. Certainly, in order to generalise the morphological relationships that are discovered by these systems and associate morphology with its function, syntactic information would seem of great use.

### 3.2.4 Discovering Allomorphy

Attempts to automate the discovery of allomorphy have centred on learning the regular rules of phonological conditioning of allomorphs. As mentioned above, Goldsmith (2001) attempts to identify related signature sets and modify the suffixes in these sets so that multiple signatures are collapsed into one where appropriate. However, this is only a first step in identifying allomorphy and is unable to be generalised to include phonological rules specifying conditioning environments and differences in the realisation of the one suffix within a signature.

In contrast, Goldwater and Johnson (2004) discuss a Bayesian framework for learning phonological rules. Their method runs on the output of Goldsmith’s Linguistica system and is also unsupervised. It learns a series of insertion, deletion and replacement rules to identify similar signatures, then hypothesises the contexts in which these rules apply and finally collapses signatures together by applying the most widely attested of these rules in the data. They experiment with using both a standard MDL based measure of the utility of the resulting morphological grammar and other cost functions, concluding that a measure that is tailored to morphological grammars is most appropriate to the discovery of phonological rules. The phonological rules learnt by their system are quite limited and are unable to model many complex phonological regularities. However, work on a supervised learning algorithm by Albright and Hayes (2002) models a more complex system, allowing multiple character changes in one rule and incorporating rule ordering information.

The discovery of allomorphy is one area in which the incorporation of part of speech information could be of great help, in particular in cases where allomorphs do not display a great deal of orthographic or phonetic similarity. For example, if two words have the same syntactic function but this is realised by different morphology this is a clear indication of allomorphy. As a result, the identification of conditioning environments is also simplified.
3.3 Towards Combining Part of Speech Induction and Morphological Induction

As described above, a number of morphological induction systems have incorporated syntactic information, either by using a part of speech tags as an explicit knowledge source, or by incorporating syntactic evidence into a knowledge-free framework in the form of the distributional properties of words. This is in contrast with methods that are based solely on orthographic information, treating training corpora as a bag of words and throwing away the structural information that they contain. Not only is this extra information (whether explicit in its incorporation or not) a potentially valuable source of evidence for discovering morphemes and morphological relationships, it also has a potential role in discovering further complexities of morphological systems such as allomorphy and its conditioning, paradigmatic relations and the functions of morphology.

Belkin and Goldsmith (2002) use syntactic (part of speech) clusters induced from an unannotated corpus to investigate whether suffixes derived from the unsupervised learning of morphology have a consistent syntactic function. Syntactic clusters are represented in two-dimensional space and the occurrences of a morphological suffix are then located in this space. If the suffix occurrences form a cluster they hypothesise that the suffix has an unambiguous syntactic function. This research is a first step in linking the form of morphemes with their syntactic function, however, it provides no means for generalising this link between form and function to enable the generation of new word forms.

The main source of evidence in the unsupervised learning of part of speech categories is usually distributional. These methods commonly rely on the cooccurrence properties either of lexical items or hypothesised syntactic categories, to cluster words into syntactic categories. However, Clark (2003) incorporates morphological evidence into a part of speech induction algorithm with the aim of improving the performance on infrequent words. He includes a model of word string probabilities in a context-based part of speech clustering algorithm to increase the probability of partitions in which morphologically similar strings occur in the same cluster. He reports an improvement in performance of the part of speech induction method, especially on infrequent words.
Chapter 4

Methodology

This study investigates the correlation of morphology and syntax within the framework of unsupervised machine learning: specifically, whether the incorporation of information induced by a part of speech induction system can improve the performance of a morphological induction system and vice versa. Some previous work reported in the literature has involved the incorporation of syntactic information into morphological induction and the incorporation of orthographic information into part of speech induction. However, information induced by one learner in the form of explicit annotation is not known to have been used to bootstrap the other.

There are two induction components in this study. The first section below describes the part of speech induction component, including the incorporation of morphological annotation, evaluation techniques and the approach to tagging text with part of speech information induced by this component. The second section describes the morphological induction component, how part of speech tags are incorporated into it, its evaluation and how it can provide morphological annotation for the first component.

4.1 Part of Speech Induction

Many part of speech induction algorithms are unable to appropriately assign infrequent words to clusters due to the lack of evidence involving these words in the training data. However, these words often make up a large proportion of the vocabulary and are therefore of great importance. The part of speech induction component used in this study is based on class-based bigram language models and does not move infrequent words between clusters. The incorporation of morphological annotation into this component provides stronger evidence for the assignment of infrequent words to clusters. This thesis investigates the effect of incorporating this information on overall part of speech induction performance.

4.1.1 Class-Based Bigram Language Models

Language models have been applied to various natural language processing tasks including speech recognition, machine translation and automatic spelling correction. In general, given some sequence of observations, a language model can assist in choosing between a number of hypotheses about the linguistic structures that produced these observations. For example a spelling correction system may find that pouring rain and pouring pain are equally likely to have been mistyped as pouring tain. However, a language model can distinguish which of these phrases has a higher a priori likelihood in English. In this example it would indicate that pouring rain is much more likely than pouring pain, leading to the conclusion that pouring rain was what the user intended to type.

The a priori likelihood of a sequence of words $w_i^k$ can be calculated by assuming that each word $w_i$ in $w_i^{k-1}$ depends on prior context. Thus the probability of this sequence can be
expressed as:

\[ Pr(w_k^k) = Pr(w_1) Pr(w_2|w_1) \cdots Pr(w_k|w_{k-1}) \]

In the conditional probability \( Pr(w_k|w_{k-1}^{k-1}) \), \( w_k \) is called the prediction and \( w_{k-1}^{k-1} \) the history.

An n-gram language model makes the Markov assumption that a word can be predicted by local prior context: specifically the \( n-1 \) previous words. Under this assumption \( Pr(w_k|w_{k-1}^{k-1}) \) is equal to \( Pr(w_k|w_{k-n+1}^{k-1}) \). Thus, in a bigram model two histories are considered equivalent if they end in the same word, in a trigram model they are considered equivalent if they end in the same two words, and so on. This assumption enables the construction of equivalence classes (groups of equivalent histories) from which probability estimates can be made. So, in a bigram model,

\[ Pr(w_k^k) \approx \prod_{i=2}^{k} Pr(w_i|w_{i-1}) \]

The process of training enables the estimation of the parameters of an n-gram language model. Training is performed on sample linguistic data called a training text. Using a simple estimation technique called maximum likelihood estimation, the probability of a word is calculated by:

\[ Pr(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})} \]

where \( C(w) \) is the number of occurrences of \( w \) in the training data.

In n-gram language modelling there is a trade-off between accuracy and coverage. Clearly, when \( n \) is small, the limited contextual information provides less accurate probability estimates. However, when \( n \) is large, the number of occurrences of a certain history within a limited amount of training data are likely to decrease, in turn reducing the reliability of the model’s probability estimates. Even using low-order n-gram language models, there are a large number of rare observations for which there is not enough training data to make accurate probability estimates. However, if a generalisation is made by assigning words to classes and then using these classes for probability estimation, it is possible to increase the amount of evidence in the training data, resulting in more accurate probability estimates for rare events. This approach was first taken by Brown et al. (1992) and is called class-based language modelling.

A class-based bigram language model is constructed from training data of length \( k \) as follows. If each word in the training data has been assigned to one class, then the occurrence of a word \( w_i \) from cluster \( c_i \) depends on the cluster \( c_{i-1} \) of the preceding word. Thus bigram probabilities are given by:

\[ Pr(w_i|w_{i-1}) = Pr(c_i|c_{i-1}) \cdot Pr(w_i|c_i) \]

Given an assignment of words to clusters, the perplexity of the resulting language model with respect to the training text can be calculated. As perplexity is an information-theoretic measure indicating the uncertainty of its predictions of the text, a lower perplexity indicates a more accurate language model. Perplexity, \( B \), is given by the reciprocal of the geometric average of the per-word probabilities in the text.

\[ B = Pr(w_1^k)^{-\frac{1}{k}} \]

This can be simplified as detailed in Appendix A to:

\[ \sum_{c_{i-1}, c_i} C(c_{i-1}, c_i) \cdot \log \frac{C(c_{i-1}, c_i)}{C(c_{i-1}) \cdot C(c_i)} + \sum_w C(w) \cdot \log C(w) \tag{4.1} \]

A suitable assignment of words to equivalence classes is obtained by a clustering algorithm. A few different clustering algorithms are described in the literature and the approach taken in this study is described by Martin et al. (1998). The algorithm aims to find the assignment of
words to classes that minimizes the perplexity of the resulting language model. However, no algorithm, this one included, is known to guarantee an optimal clustering result. This is an exchange algorithm similar to the k-means algorithm for clustering and proceeds as follows:

Initialize the clusters:

- If \( N \) classes are required, assign the \( N - 1 \) most frequent words to their own class and the remaining words to the class \( N \).

Iterate for a predefined number of iterations:

1. for each word \( w \) in the vocabulary that has a frequency count above a certain threshold
2. remove \( w \) from its current class \( N(w) \)
3. for all other classes \( n \) compute the perplexity as if \( w \) were moved to \( n \)
4. assign \( w \) to the class with the lowest perplexity.

Previous research has indicated that the initial assignment of words to clusters has little effect on the quality of the clusters, but can have a significant effect on the speed of convergence of the algorithm. Frequency-based initialization, as implemented in this study, is reported as resulting in fast convergence. To improve the speed of the algorithm, infrequent words whose occurrence counts in the training text are under a prespecified threshold are never moved. The algorithm also stops if no words were moved between clusters in the previous iteration.

Computation of the perplexity of a language model has a high complexity, however Martin et al. describe a method that improves the efficiency of the calculation of improvements in perplexity. For each iteration, only the differences in perplexity are calculated. Thus, only those terms in equation 3.1 above which are affected by moving words \( w \) from class \( n_w \) to class \( m_w \) are considered. When removing a word \( w \) from class \( n_w \), the following update formulae are used:

\[
\forall n \neq n_w : C(n, n_w) := C(n, n_w) - C(n, w)
\]
\[
\forall n \neq n_w : C(n_w, n) := C(n_w, n) - C(w, n)
\]
\[
C(n_w, n_w) := C(n_w, n_w) - C(n_w, w) - C(w, n_w) + C(w, w)
\]

Similarly, when adding \( w \) into class \( m_w \), the update formulae are:

\[
\forall m \neq m_w : C(m, m_w) := C(m, m_w) + C(m, w)
\]
\[
\forall m \neq m_w : C(m_w, m) := C(m_w, m) + C(w, m)
\]
\[
C(m_w, m_w) := C(m_w, m_w) + C(m_w, w) + C(w, m_w) + C(w, w)
\]

If we have \( N \) classes, \( N \cdot V \) move attempts are made per iteration, where \( V \) is the size of the vocabulary. The updates to compute the change in perplexity can be calculated in \( O(N) \) time and are calculated for each attempt to move a word to a new class. For each iteration of the algorithm, a scan through the training text is performed to calculate the counts \( C(w, n) \) and \( C(n, w) \). Thus, the time complexity of the algorithm is:

\[
O(I \cdot (T + V \cdot N^2))
\]

where \( I \) is the number of iterations and \( T \) is the length of the training text.

By storing the bigram counts involving each word it is possible to avoid scanning the whole training text in order to calculate the counts involving a word \( w \). Rather these bigram counts can be calculated from an initial pass over the training text and then accessed directly where necessary. This reduces the time complexity to:

\[
O(I \cdot (B + V \cdot N^2))
\]
where $B$ is the number of unique word bigrams in the training text.

The part of speech induction method adopted for this study uses Martin et al.’s optimized algorithm assign words to clusters. The exchange procedure aims to minimize the perplexity of the class-based bigram language model constructed from these clusters. The cluster assignment can be thought of as modelling the syntactic structure of a text and in this framework, the resulting clusters correspond to part of speech groupings.

### 4.1.2 Incorporating Morphological Annotation

Morphological annotation in the form of a tag on each word indicating its suffix is incorporated into the part of speech induction algorithm described above, with the aim of improving the assignment of infrequent words to clusters. Recall that words with a frequency count below some predefined threshold in the training data are never moved from the cluster to which they were initially assigned. Whilst this strategy greatly reduces the computational complexity of the clustering algorithm, it precludes any meaningful assignment of infrequent words to clusters. Where infrequent words form a large proportion of the vocabulary, this will have a great effect on the overall performance of the part of speech induction system.

To address this issue, morphological annotation is incorporated into the part of speech induction algorithm as follows. After the final iteration infrequent words are moved to the cluster that contains the highest percentage of their suffix. This approach assumes that a word’s morphological annotation is likely to provide some indication of its part of speech.

### 4.1.3 Evaluation

A number of approaches have been taken to the evaluation of part of speech clusters in the literature. Some early work relied on intuitive judgements of the correspondence between clusters and syntactic categories. Clearly this approach relies on subjective judgements of a competent speaker of the language. As a result, comparison across different systems could not be relied upon to be consistent and experimentation would be limited to languages in which a researcher is competent.

A better method of evaluation compares the induced part of speech tags with the tags of some manually or semi-automatically tagged text. Each word in the gold standard part of speech-tagged text is also tagged with its induced part of speech cluster. Of course this requires some form of gold standard tagged text or a reliable automated part of speech tagger, and such resources are widely available. An information theoretic measure called conditional entropy is used to assess the correlation of two tagsets. Conditional entropy measures the average amount of information per induced tag, given the gold standard tag. Thus, low average information indicates that there is high agreement between the tags. In contrast, if there is little correlation between the tags, this is indicated by low mutual information and the conditional entropy will be close to the entropy of the gold standard. If $G$ refers to the gold standard tags and $T$ the induced tags, conditional entropy $H(G|T)$, is given by:

$$H(G|T) = H(G) - I(G; T)$$

where $H(G)$ is the entropy of the tagset, given by:

$$H(G) = \sum_{t \in G} -Pr(t) \cdot \log Pr(t)$$

and $I(G; T)$ is the mutual information between the tags, given by:

$$I(G; T) = \sum_{g \in G, t \in T} Pr(g, t) \cdot \log \frac{Pr(g, t)}{Pr(g) \cdot Pr(t)}$$

Although this measure evaluates against a gold standard, it is important to note that there are many different possible part of speech tag sets that differ in the linguistic information that they
add to text. Thus, evaluation against one gold standard tagset will give different results to another. As Clark (2003) notes, this measure also favours accurate tagging of frequent words and therefore a simple baseline system that assigns the most frequent words to different classes can score quite well.

An alternative form of evaluation is to measure the perplexity of the induced language model. However, whilst perplexity is a common evaluation measure for statistical language models, it is not clear that it evaluates the linguistic accuracy of such models. In contrast to conditional entropy, perplexity does not measure accuracy with respect to a linguistically gold-standard. In particular, n-gram language models do not take into account long distance dependencies in text, yet some parts of speech (such as relative pronouns) cannot be predicted by local context.

### 4.1.4 Part of Speech Tagging

Induced part of speech information is provided to the morphological induction component in the form of part of speech tagged text. Part of speech tagging is probably the most common annotation added to natural language texts, however, even for one language there may be multiple part of speech tagsets that differ in the granularity and range of syntactic information that they indicate. The part of speech clusters that are learnt by this part of speech clusterer are not given meaningful linguistic labels: that is, no attempt is made to decide which cluster corresponds to nouns, which to prepositions etc. This task presents a separate challenge for unsupervised learners (Schone and Jurafsky, 2001b). A word’s cluster number becomes its part of speech tag, providing an indication of the similarities between words that belong to the same cluster.

Each occurrence of a word in a text is tagged with its part of speech in the current context. In general each word form may have multiple parts of speech. However, the part of speech induction algorithm used for this study is a hard clustering algorithm, assigning each word form to exactly one cluster. As a result, ambiguity in syntactic function of a word form is ignored and one part of speech is assigned to each word form.

### 4.2 Morphological Induction

This thesis aims to investigate the effect of incorporating part of speech information, in the form of part of speech tagged training data into a morphological induction system. This part of speech information not only assists in the identification of morphologically related words, but also in the identification of allomorphy through the association of meaning or syntactic function with morphologically complex forms. Part of speech information is also useful in the generation of new words, as it indicates how a new word form can be used within the structure of a sentence or utterance (although in the case of induced part of speech information, part of speech tags can only indicate syntactic function within the context of the model from which they are induced). The association of meaning with form is a challenge that most morphology induction systems do not attempt to address, yet it is vital in order to construct accurate models of the morphology of languages.

The Whole Word Morphologizer system developed by Neuvel and Fulop (2002) incorporates part of speech tags as a knowledge source to identify morphologically related word forms and generate new words. In this system part of speech information is used to improve the linguistic accuracy of the induced model of morphology and to disambiguate morphology where necessary. Although outside the scope of the current study, its basis on string similarity means it also has the potential to be extended to identify non-concatenative morphology through the incorporation of more complex string alignment procedures.

#### 4.2.1 The Theory of Whole Word Morphology

The theory of Whole Word Morphology is based on word and paradigm morphology, defining morphology as transformations that relate whole word forms with other whole word forms (Ford
and Singh, 1991)(Ford et al., 1997). However, it differs from the word and paradigm approach in its treatment of suppletion. According to the theory of Whole Word Morphology, suppletive morphology is not really morphology at all, but merely word forms related by convention. Thus, for example go and went are not considered to be morphologically related, despite the fact that they express the same grammatical relationship as a regular morphologically related pair such as walk and walked. Defining suppletion as outside the scope of morphology removes the requirement for a morphological learner to recognise such forms. However, an inability to acknowledge or account for suppletive forms limits the practical use of such a learner. Indeed, suppletive forms tend to include some of the most frequently occurring words in a language. For example, in many languages the verb to be (probably the most commonly used verb) is suppletive. Suppletive morphology presents a challenge for most morphological learners and in particular those that are based on similarity of orthographic form.

4.2.2 Whole Word Morphologizer

The morphological learner adopted in this study is the Whole Word Morphologizer (henceforth WWM) system developed by Neuvel and Fulop. This system is based on the theory of Whole Word Morphology and aims to identify morphological relationships in a language. The algorithm discovers the morphology of a language from the differences and similarities between the strings of letters that form words in the language. Morphologically related word pairs are identified by a comparison of word forms, using the intuition that two words could be morphologically related if they have similar forms. An approximation of similarity is made by comparing the segments (letters in this case) in word A with those in word B. If the two words share a certain number of letters they are listed as a potential morphologically related pair. Words are compared letter by letter, starting from the beginning of both words and working forwards. For example, evaluate and evaluation share a common prefix evaluat, but differ in their suffixes e and ion. It is important to note that this preliminary comparison process locates words with common prefixes and thus the morphological differences between words that it identifies are suffixes.

Once all pairs of potentially morphologically related words have been identified, duplicate differences between pairs are located. If a number of pairs with the same differences are identified, this is considered sufficient evidence of a morphological relationship between the words. Neuvel and Fulop term this morphological relationship a morphological strategy. To continue the above example if, as well as evaluate and evaluation, complete and completion have also been identified as a pair of potentially related words, then it is clear that the formal differences between the word in the pairs are identical. Thus we can say that the stems evaluat and complet both participate in the same morphological strategy, and this strategy is realised by e and ion.

The identified strategies are also refined to apply only in as restricted an environment as possible. This is accomplished by incorporating any similarities in the form of the stems to which this strategy applies into the strategy. The differences between morphologically related word forms are modified to include any common suffixes of the word forms to which they apply. Thus, as evaluat and complet both end in a t, this t is added into the strategy as part of the differences between related words. The refined strategy is now realised by te and tion. In effect, this ensures that each strategy incorporates maximal similarities between the word forms from which it was discovered.

4.2.3 Incorporating Part of Speech Tags

Syntactic information in the form of part of speech tags is incorporated into this morphological learning algorithm to restrict the morphological strategies further, thus increasing their accuracy and descriptive power. For example, WWM may discover two pairs of morphologically related words that exhibit the same formal differences; walk and walks and dog and dogs.
Without part of speech information these pairs would be considered instances of the same morphological strategy and both would be considered evidence for the existence of such a strategy. However, with part of speech information, it is clear that the relationship between the words is different: \textit{walk} (verb) to \textit{walks} (3rd person singular verb) and \textit{dog} (singular noun) to \textit{dogs} (plural noun). Thus, WWM would consider the two related pairs as evidence for two separate morphological strategies and the resulting model of morphology better indicates the linguistic structure of the language being learnt.

As well as this, part of speech information provides a basis for the disambiguation of paradigmatic variants and the association of meaning or syntactic function with morphological form. Such an association provides important linguistic information that is necessary in applications requiring not only morphological form, but a knowledge of the syntactic contexts in which it occurs. For example, a language generation system might construct a sentence with a 3rd person singular subject and therefore require the form of the verb \textit{walk} (for example) that agrees with the subject and is appropriate in this syntactic context. Assuming an appropriate morphological strategy has been learnt, a morphological analyser with part of speech information is able to connect the 3rd person singular verb with the form +\textit{ed}, to generate the appropriate form \textit{walked}. In addition, this association of form with function enables the easy identification of allomorphic variants and the contexts in which they occur. As mentioned above, systems such as Goldsmith’s Linguistica attempt to make this association after morphological forms have been discovered. However, where one morphological form has multiple syntactic functions however, the association can be unclear. This is clearly an area where part of speech information is of great use.

Part of speech tags are easily incorporated into the WWM algorithm by extending the notion of a word as not just a sequence of letters, but rather as the combination of both the word’s orthographic form and its part of speech. When collecting duplicate differences between morphologically related words, it is therefore necessary that the related words exhibit not only the same orthographic differences, but also the same syntactic differences as indicated by their part of speech. So, in the earlier example, \textit{walk} (verb) to \textit{walks} (3rd person singular verb) and \textit{dog} (singular noun) to \textit{dogs} (plural noun) would not be considered duplicate examples of the same morphological relationship.

The version of this algorithm described by Neuvel and Fulop takes a small lexicon (up to 5,000 words) as input, whereas the version implemented for this research has been adapted to operate on continuous text. However, it is important to note that the structure of text and the contexts in which words occur are not used in any way in learning morphological information. The text is processed as a word list (where a word is the combination of a word from and its part of speech tag) and any duplicates are removed. Thus, the only indication of a word’s syntactic role is given explicitly in the form of the word’s part of speech tag.

### 4.2.4 Evaluation

Despite the recent proliferation of research into morphological induction, there is no standard approach to the evaluation of such systems (Maxwell, 2002). A number of evaluation metrics have been employed for morphology learning systems and these evaluation metrics vary as much as the concept of morphology that these systems aim to learn. However, some of the commonly employed metrics in the literature have aimed to evaluate:

1. Word segmentation into morphemes
2. Morpheme boundary identification
3. Discovery of affixes (possibly including allomorphic variation)
4. Discovery of morphologically related word forms
5. Ability to generalise morphological relationships to generate new words

6. Performance improvement of another system that includes the morphological learner as a component

Of these learning aims, all except the first two are relevant to the model of morphology learnt by WWM. The third aim, to discover affixes, is not overtly encompassed by the model of Whole Word Morphology, however, the variable components between related words (e.g. \textit{ed} and \textit{ing} in \textit{walked} and \textit{walking}) that WWM discovers can be viewed as loosely corresponding to affixes that incorporate allomorphic variation. Indeed allomorphy can be discovered quite easily by this system, as the syntactic classes of related words clearly indicate common morphological relationships that are realised by multiple morphological strategies.

In their discussion of the system, Neuvel and Fulop argue that the value of a morphological learner (WWM in particular) lies in its ability to do two things: learn the relations between words and generate new words from this information (a word being the combination of a word form and its part of speech tag). However, their evaluation metric measures only the second. To evaluate the system they measure the precision of generated new words: the number of new words (generated words that are not already in the lexicon) that are valid words in the language. Words are generated by unifying each word in the training data with one side of a morphological strategy that has been learnt by the system. If a match is found, then a new word form is generated by removing the variable component of this side of the strategy and replacing it with the variable component of the other side of the strategy. For example, if the lexicon contains the past tense verb \textit{walked} and from other examples in the corpus the system has learnt a strategy relating past tense verbs ending in \textit{ed} to gerunds ending in \textit{ing}, then the system will remove the \textit{ed} from the sequence \textit{walked} and replace it with \textit{ing} to generate a new gerund of the form \textit{walking}.

Neuvel and Fulop do not specify how they make a judgement as to whether a newly generated word is in the language: perhaps the judgement of a native speaker, presence of the word on the internet (as suggested in previous discussions of this system (Neuvel, 2002)) or membership of a lexicon. However, the first two options seem untenable. Consultation of a native speaker leaves open the possibility of subjective judgements of word validity and is a manual form of evaluation: unsuitable for an automated system that generates thousands of new forms. The internet seems attractive in its availability and the potential for wide coverage, yet it covers multiple languages even within one site or page and is an unreliable source of ‘correct’ language usage. In contrast, a machine-readable lexicon containing both inflectional and derivational forms or a spellchecker would provide a reference against which to assess the validity of generated words. Whilst such resources could not include all word forms (e.g. proper nouns), the coverage could be expected to be reasonable and such resources are quite commonly available for a variety of languages.

In this study a metric taken from Schone and Jurafsky (2001a) was used to evaluate WWM’s ability to discover conflation sets. A conflation set is a group of morphologically related words, for example \{\textit{appoint}, \textit{appointing}, \textit{appointed}, \textit{appointment}, \textit{appointments}\}. Through the construction of morphological strategies, WWM learns which words are related to one another. This enables the construction of conflation sets, which can be compared with a gold-standard. The gold standard takes the form of a machine-readable morphologically analysed lexicon, from which sets of related words can be extracted. To evaluate, for each word that occurs in both data sets, the proportion of correct\((C)\) words in WWM’s conflation set, the proportion WWM has inserted\((I)\) and the proportion it has deleted\((D)\) are calculated. For a word \(w\), if \(X_w\) is its WWM-induced conflation set and \(Y_w\) its gold standard conflation set, the proportions are calculated as follows:

\[
C = \sum_w \frac{|X_w \cap Y_w|}{|Y_w|} 
\]

20
Table 4.1: Example Scoring of Conflation Sets

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3/4</td>
<td>3/4</td>
<td>3/4</td>
<td>1/4</td>
<td>1/2</td>
<td>1/2</td>
<td>3.5</td>
</tr>
<tr>
<td>I</td>
<td>0/4</td>
<td>0/4</td>
<td>0/4</td>
<td>1/4</td>
<td>1/2</td>
<td>0/2</td>
<td>0.75</td>
</tr>
<tr>
<td>D</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>3/4</td>
<td>1/2</td>
<td>1/2</td>
<td>3.5</td>
</tr>
</tbody>
</table>

For example, suppose that WWM induced the conflation sets: \{a,b,c\}, \{d,e\}, \{f\} and the gold-standard conflation sets were \{a,b,c,d\}, \{e,f\}. Table 4.1 indicates how WWM would be scored. So, for word \(a\) the WWM-induced conflation set correctly includes 3 of the 4 members of the gold-standard conflation set, therefore it has a \(C\) score of \(\frac{3}{4}\). The other values in the table are calculated in a similar way.

The scores are totalled over all words and the precision and recall of the induced conflation sets \((c)\) are calculated from these totals, where:

\[
Precision = \sum_c \frac{C_c}{C_c + I_c}
\]

\[
Recall = \sum_c \frac{C_c}{C_c + D_c}
\]

Precision measures the proportion of answers given by the system that were correct, where recall measures the proportion of possible correct answers that were given by the system. These two measures are in opposition to one another, as a system that strives for accurate answers will often provide poor coverage as a result and similarly, a system that generates a high number of correct answers is likely to also generate a large number of incorrect ones. Thus, to balance the precision and recall scores, the F-score is calculated, as follows:

\[
F\text{-score} = \frac{Precision \cdot Recall}{(Precision + Recall)/2}
\]

4.2.5 Morphological Annotation

The morphological information discovered by WWM can be used to annotate the text from which it is discovered. This morphologically annotated text can be used as training data for another natural language learning system, providing additional linguistic information from which to learn. In this study, morphological annotation supplements the unlabelled data used to train the part of speech induction component. The incorporation of this morphological information into the part of speech induction system also enables the evaluation of the performance of the morphological learner as a component of another system by comparing the performance of the part of speech induction system on unlabelled text with its performance on morphologically annotated text.

As discussed above, WWM analyses words as being composed of a variable and a non-variable component, loosely corresponding to suffix and root respectively in the traditional theory of concatenative morphology. Each word is annotated with its variable component. So, for example a morphologically complex word like \(dogs\) would be annotated with \(s\) and a morphologically simple word would be annotated with an empty suffix. Where a word participates in multiple morphological strategies the shorter suffix is used to annotate the
word, under the intuition that the final suffix is most likely to indicate a word’s part of speech. For example, it is possible that carefully is analysed as care + fully in its relation to care, but careful + ly in its relation to careful. In this case it would be annotated with the suffix ly.

In order to assess the value of generalising morphological strategies to new words, two approaches to annotating the text are investigated in this study. The first is to restrict annotation to the model of the training text and only annotate those words that have actually been identified as member of the relevant morphological strategy. This means that both halves of the morphological strategy must be instantiated in the training data. This approach ignores the system’s ability to generalise morphological strategies to apply to new words and will be referred to as the under-generation tagging strategy. The second approach, the over-generation tagging strategy, involves generalising the morphological strategies learnt by WWM and annotating a word according to any morphological strategy that it matches. In contrast to the first strategy, this means that any word that matches one side of a morphological strategy will be annotated accordingly.

4.3 Combining Part of Speech Induction and Morphological Induction

The study investigates the effect of combining the two induction systems such that one provides information to the other in the form of explicit annotations.

To incorporate morphological information into part of speech induction, the morphological strategies discovered by WWM are used to annotate training text for the part of speech induction algorithm. The morphological annotations are used to move infrequent words between clusters, where these words would otherwise remain in their initial cluster.

In the other direction, induced part of speech information is incorporated into WWM. Each word in the training text is tagged with its cluster number as assigned by the part of speech induction algorithm. This tagged text is used to train WWM and restrict the resulting morphological strategies, increasing their linguistic accuracy.
Chapter 5

Experiments

Two sets of experiments were conducted to investigate the effect of the incorporation of explicit morphological annotations into a part of speech induction system and vice versa. These experiments were limited to the English language and were performed using items 1 to 22 (50,084 tokens) of the Brown Corpus (Francis and Kučera, 1982). The first 25,000 words were used for training and the remainder for testing where necessary. The Brown Corpus is a corpus of modern American English consisting of excerpts from English prose documents and the sections used in this study comes under the genre of press-reportage. The corpus is hand-tagged with part of speech tags and can therefore be used as a gold-standard for part of speech tagging. Its tagset consists of 87 distinct tags.

To increase the chance of locating orthographically similar words, the WWM system converts its training text to lower case. In contrast the part of speech induction algorithm operates on the training text in its original form. This explains the different vocabulary sizes quoted for the two experiments.

5.1 Incorporating Morphological Annotation into Part of Speech Induction

5.1.1 Morphology as an Indication of Part of Speech

Preliminary experiments were conducted to investigate the extent that suffixed morphology indicates a word’s part of speech. Using suffixed annotation as an approximation of part of speech tags, the conditional entropy of suffixed-based clusters with respect to the gold-standard part of speech tags was calculated. The results as measured over the training text and the vocabulary are given in Table 5.2. Three methods for annotation morphological suffixes were investigated. The first is a naive approach that tags every word with its final two letters. The other two approaches, used the knowledge-free WWM induction system trained on the same training text to annotate each word with suffixed information. One of these used the over-generation, the other the under-generation tagging strategy as described in section 3.2.5. Both of these annotation approaches can annotate words with empty suffixes.

All three approaches to morphological annotation improve on the entropy of the gold standard tags (as given in Table 5.1), indicating that there is some correspondence between these annotations and part of speech tags. Of the three, the naive morphological annotation strategy provides the best indication of part of speech. This annotation strategy results in 279 different suffixed tags, with the most frequent occurring on 7% of the words in the training text. In

<table>
<thead>
<tr>
<th></th>
<th>Entropy (training text)</th>
<th>Entropy (all words)</th>
<th>Entropy (infrequent words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold standard tags</td>
<td>3.4463</td>
<td>3.0488</td>
<td>2.8413</td>
</tr>
</tbody>
</table>

Table 5.1: Entropy of the Gold-Standard Tags
contrast the annotation strategies based on WWM’s morphological induction, contain 20 – 40 different suffixal tags with the most frequent (an empty suffix) occurring on more than 79% of the words in the training text. Intuitively a larger number of small clusters are more likely to result in a tagging that correlate with the gold standard tags, and this probably explains the performance of the naive annotation strategy.

### 5.1.2 Experiments

Various part of speech induction methods were implemented in order to investigate the effect of incorporating morphological information in the form of suffixal annotation into a part of speech induction system. All methods were run with 40 clusters. This value was chosen after analysis of the resulting clusters upon variation of the number of clusters. As a result of some initial runs, it was decided that methods with fewer than 40 clusters did not allow adequate distinction between the part of speech classes. The smallest number of clusters quoted in experimental results in the literature is 32 (Clark, 2003) and this is in comparison to 64 and 128 clusters. As a result of limited computation time and resources, larger numbers of clusters could not be investigated in this study. Despite optimizations as described in section 4.1.1, the part of speech clustering algorithm implemented in Python in order to take advantage of tools available as part of the Natural Language Toolkit (Loper and Bird, 2002), took between 6 and 10 hours to run to completion on a 2.8GHz Intel Xeon processor.

The number of clustering iterations was set to 30. However, in practice the runs of the iterative part of speech induction algorithm for this study stopped after 18 iterations: by this stage no words could be moved to obtain a lower perplexity of the language model. The word frequency threshold was set to 1. That is, words which only occurred once in the training text were not moved between clusters in the iterations of the part of speech clustering algorithm. These words are the infrequent words, whose clustering may be improved by the incorporation of morphological annotation. In the 25,000 word training text used, 2,973 words occurred only once: close to 12% of the text. The total size of the vocabulary (without converting words to lower case) was 5,225. Thus, over half of the vocabulary occurred only once.

Experiments were run using a number of part of speech induction methods. Two of these methods do not incorporate morphological annotation: one is the part of speech cluster induction method described in the previous chapter, that iteratively minimizes the perplexity of a class-based bigram language model. The other assigns words to clusters according to their frequency: performing only the initialisation step of the former method. Thus for 40 clusters, the 39 most frequent words are assigned to clusters 1–39 and the remaining words to the 40th cluster. This method provides a baseline for part of speech induction performance. Three further part of speech induction methods incorporating suffixal morphology were investigated. In these methods, infrequent words were moved after the final iteration into the cluster containing the highest percentage of their suffix. The three approaches to morphological suffix annotation are those described in section 5.1.1 above.

### 5.1.3 Results

In general the clusters constructed by the part of speech induction algorithm contain words of many different parts of speech. They are quite noisy groupings, some of which contain

<table>
<thead>
<tr>
<th>Morphological Annotation (suffixes)</th>
<th>Conditional Entropy (training text)</th>
<th>Conditional Entropy (all words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nave suffix tagger</td>
<td>3.2006</td>
<td>1.4821</td>
</tr>
<tr>
<td>WWM: over-generation tagging strategy</td>
<td>3.4156</td>
<td>1.8848</td>
</tr>
<tr>
<td>WWM under-generation tagging strategy</td>
<td>3.4179</td>
<td>2.4568</td>
</tr>
</tbody>
</table>

Table 5.2: Conditional Entropy of Suffixal Tags with respect to the Gold-Standard
on, around, at, during, into, under, against, off, getting, aside, through, investigating, retained, injured

will, can, cannot, can’t, could, should, ever, Superior, safety, didn’t, P.M., impressive, really, places, actually

been, brought, forced, met, attracted, drive, served, provided, gone, caused, held, set, consulted, issued, positions, viewed, somewhat, active, denied, pulled, errors, particularly, lost, operated, dropped, studied, turned, considered, appointed, completely, received, noticeable, placed, constituted, gave, gets, anonymous, raised, informed, given, gotten, ranged, fulfilled, insist, sought, decided

who, whose, we, I, there, also, well, Pfaff, details, they, dinner, you’ll, Throneberry, I’d, Khrushchev, Shiflett, Allen, he, Worth, reporters, she, technicians, Ratcliff, India


Figure 5.1: Example Part of Speech Clusters

predominantly words that have the same syntactic function, others do not. Some example clusters are given in Figure 5.1. These have been chosen as exhibiting some tendency towards part of speech groupings: prepositions, modal verbs, past tense verbs, pronouns, proper nouns. Many other clusters are not composed of words with one clear syntactic function.

In comparison with the sample clusters given by Brown et. al.(1992), these clusters are very noisy. This is probably as a result of the limited amount of training data used in this study: Brown et. al. cluster a 260,741 word vocabulary taken from 365 million words of training text into 1,000 clusters. Clearly the current study is quite limited in comparison. Nevertheless, it is interesting to investigate the extent to which part of speech induction methods trained on limited amounts of training text can learn useful information.

Perplexity

To evaluate the various part of speech induction methods, the perplexity of the resulting language models with respect to both the training and testing texts was calculated. The results are presented in Table 5.3. In all cases the perplexity with respect to the training text is better (i.e. lower) than that with respect to the testing text. This is to be expected as the model is more likely to fit the training text than an unseen text.

The method incorporating morphological annotation from WWM using the over-generation tagging strategy stands out as having worse perplexity even than the baseline clustering method. This is surprising but plausible, as the assignment of over half the vocabulary to clusters according to morphology will effect the predictions of a language model constructed from these clusters. It is important to note that perplexity does not necessarily reflect the linguistic accuracy of a language model, but is a purely statistical measure of the accuracy of the model with respect to some data.
<table>
<thead>
<tr>
<th>Clustering Method</th>
<th>Perplexity (training text)</th>
<th>Perplexity (testing text)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold-standard tags</td>
<td>433.19</td>
<td>471.60</td>
</tr>
<tr>
<td>Baseline clusterer</td>
<td>722.29</td>
<td>960.45</td>
</tr>
<tr>
<td>POS clusterer</td>
<td>411.33</td>
<td>422.80</td>
</tr>
<tr>
<td>POS clusterer with infrequent words clustered by naive morphology</td>
<td>437.17</td>
<td>451.21</td>
</tr>
<tr>
<td>POS clusterer with infrequent words clustered by morphology: over-generation tagging strategy</td>
<td>1267.46</td>
<td>1377.14</td>
</tr>
<tr>
<td>POS clusterer with infrequent words clustered by morphology: under-generation tagging strategy</td>
<td>412.57</td>
<td>424.10</td>
</tr>
</tbody>
</table>

Table 5.3: Perplexity of Induced Language Models

For the other forms of morphological annotation, the incorporation of this information into part of speech induction results in a slight increase in perplexity. This increase is not large and the resulting perplexities are not significantly worse than that of a language model constructed from the gold-standard tags. This is despite the fact that the larger number of tags in the gold-standard tag set (87 compared with 40 induced clusters) would be expected to increase the perplexity.

### Conditional Entropy

A second approach to evaluating the part of speech induction methods is to calculate the conditional entropy of the induced tags with respect to a gold-standard. These results are summarised in Table 5.4. The incorporation of morphological information of all three kinds degrades performance over the training text. However, conditional entropy as measured against the training text results in a score that is weighted according to word frequency in this text. Thus, if a word occurs very frequently in the training text and it is tagged accurately, then this will have a larger effect on the conditional entropy over the training text than the accurate tagging of an infrequent word. This evaluation technique has an inherent bias towards accurate tagging of frequent words.

This bias can be viewed both as a positive and a negative. If the training text is considered to be representative of the frequency patterns of a the words in a language, it may be valid to bias a scoring mechanism towards frequent words and reflect actual usage of the language. This approach is usually taken when evaluating supervised tagging methods: a tagger is tested on a section of text (the testing text) and the accuracy of its tags is measured over this text. However, as the current study focuses on improving the performance of part of speech induction on infrequent words, it seems appropriate to avoid this bias towards frequent words. Clearly, an evaluation technique that is weighted towards accuracy on frequent words will not provide a good indication of performance on infrequent words. Thus, Table 5.4 also provides the results of conditional entropy evaluation over the vocabulary of the training text.

All of the methods that incorporate morphological information improve the conditional entropy both over all words and over infrequent words. Cluster assignment by naive morphological annotation results in the largest performance improvement. It is difficult to explain why this approach to morphological annotation gives better results than annotations from WWM. It is to be hoped that the annotations produced by WWM are more linguistically accurate than the naive annotations, however this is not reflected in the results of this study.

The incorporation of morphological annotation from the under-generation tagging strategy from WWM results in better performance than that from the over-generation tagging strategy.
<table>
<thead>
<tr>
<th>Clustering Method</th>
<th>Conditional Entropy (training text)</th>
<th>Conditional Entropy (all words)</th>
<th>Conditional Entropy (infrequent words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline clusterer</td>
<td>1.9085</td>
<td>2.9668</td>
<td>2.8413</td>
</tr>
<tr>
<td>POS clusterer</td>
<td>1.5185</td>
<td>2.5126</td>
<td>2.8413</td>
</tr>
<tr>
<td>POS clusterer with infrequent words clustered by naive morphology</td>
<td>1.6047</td>
<td>2.2394</td>
<td>2.0637</td>
</tr>
<tr>
<td>POS clusterer with infrequent words clustered by morphology: over-generation tagging strategy</td>
<td>1.6268</td>
<td>2.4993</td>
<td>2.7997</td>
</tr>
<tr>
<td>POS clusterer with infrequent words clustered by morphology: under-generation tagging strategy</td>
<td>1.5900</td>
<td>2.3192</td>
<td>2.3267</td>
</tr>
</tbody>
</table>

Table 5.4: Conditional Entropy of Induced Tags with respect to the Gold-Standard

This indicates that generalising the morphological strategies learnt by WWM may degrade the overall accuracy of resulting annotation.

In general, morphological information does improve the conditional entropy of infrequent words and of the vocabulary. However, in a text these effects can be masked by overall performance biased resulting from the frequency patterns found in actual usage.

### 5.2 Incorporating Part of Speech Tags into Morphological Induction

#### 5.2.1 Experiments

Experiments were conducted to investigate the effect of incorporating part of speech tags into WWM. As described above, the part of speech clusters induced by the part of speech induction system can be used to tag the training text. Of course, the resulting tags are noisy and not highly accurate, but do provide an approximation for part of speech information. Thus, these experiments also investigate the effect of reduced accuracy part of speech information on the performance of WWM and whether induced part of speech tags can improve performance or coverage despite their inaccuracy.

Five different methods for tagging the training text were implemented and the results of training WWM on these texts are presented in the next section. The first two taggings provide baseline and ceiling performance indicators for the system: the baseline training text is untagged (effectively the whole text is tagged with the one tag) and the ceiling is gold-standard tagged training text. Three experimental tagging methods were also explored. The first is a baseline part of speech clusterer that performs only the initialisation step of the part of speech induction algorithm described in the previous chapter. This method produces baseline part of speech cluster tags. The second tagging method uses the clusters induced by the iterative perplexity minimisation part of speech induction algorithm described in the previous chapter and its tags are the induced part of speech cluster tags. The final tagging method is based on the induced part of speech clusters, with morphological information incorporated to assign infrequent words to clusters. The morphological annotations used to make these assignments are those resulting from running WWM on untagged text and using the over-generation annotation strategy described in the previous chapter. These tags are the induced part of speech cluster tags with morphology.

For all experiments the following parameter values were used: a common prefix length of four letters was taken as indicating potentially morphologically related words and each morphological strategy required a minimum of two distinct instantiations to be considered valid. The common prefix length value was chosen after an initial assessment of the quality of the
morphological strategies resulting from the use of both smaller and larger prefix lengths. With a shorter prefix length, more words were considered potentially morphologically related, many of which were incorrect. For example, with a common prefix length of three, *want* could be related to *wander* or *pin* to *pint*. Using a prefix length this short often results over-analysis of words that are not morphologically complex. However, a longer prefix prevents analysis of some words that are morphologically complex.

5.2.2 Results

As discussed in the previous chapter, two metrics have been used to evaluate the performance of the morphological induction system. The first focuses on evaluating the system’s ability to discover conflation sets (sets of morphologically related words) and the second assesses the system’s ability to generate new words by unifying words in the lexicon with the morphological strategies it has discovered. Both evaluation techniques give an indication of the coverage and accuracy of the morphological strategies discovered by WWM.

In general, training WWM on untagged text results in the discovery of morphological strategies which are only restricted by the orthographic form of the words that they can apply to. This can lead to incorrect morphological analyses: a problem which can be avoided if part of speech information is available. As an example, a strategy discovered by WWM trained on untagged text includes the stems *unit-, case-, read-, part-, need-*, and *factor-* as all taking the suffixes *-s* and *-y*. Clearly there is more than one morphological relationship between the pairs of words that are members of this strategy. For example, *reads* v. *ready* is a relationship between a 3rd person singular verb and an adjective, where *units, factors, parts* are all plural nouns respectively potentially related to the singular nouns *unity, party, factory*. The WWM system trained on gold standard part of speech tagged text is able to make this distinction and discovers only the one strategy relating plural nouns ending in *s* to singular nouns ending in *y*. Only the stems *unit-, part-, and factor-* participate in this strategy. This is evidence that the inclusion of part of speech information assists in the identification of morphological relationships and increases the accuracy of the morphological strategies that are discovered.

**Discovery of Morphologically Related Words**

Using the method described in section 4.2.4 above, the conflation sets discovered by WWM were evaluated against the CELEX database, a hand-tagged morphologically-annotated lexicon (Baayen et al., 1995). CELEX is a machine-readable lexicon containing 160,594 word forms, compiled from the Oxford Advanced Learner’s Dictionary (1974) and the Longman Dictionary of Contemporary English (1978). It is important to note that the CELEX’s coverage with respect the the Brown corpus may be limited due to the differences between British and American English and limited coverage of inflected forms. As a result, when comparing conflation sets produced by WWM with those in the CELEX database, any word that is not in both data sets is ignored. Conflation sets are constructed from CELEX by grouping together words that contain the same stem.

The results of evaluating the conflation sets produced by training WWM on text that has been tagged using various methods are given in Table 5.5. In addition, the results of two naive approaches to constructing conflation sets are shown: the first assigns each word to its own conflation set and the other assigns all words to the one conflation set. Comparison with these approaches gives an indication of the results when the composition of the conflation sets tends towards either of these extremes.

The method trained on gold-standard tagged text results in the best F-score, indicating that the incorporation of accurate part of speech information into WWM is advantageous when discovering morphologically related words. It is important to note that the incorporation of part of speech information introduces distinctions when constructing morphological strategies which could lead to a lack of evidence for certain morphological strategies. This could cause
The methods trained on untagged text and that tagged on base-line cluster tagged text perform almost identically. This is to be expected as the baseline clusterer assigns the vast majority if the vocabulary to one cluster and the effect of tagging all words with one tag is the same as tagging no words at all. Additionally, the most frequent words which the baseline clusterer does tag with distinct cluster tags are the least likely to be morphologically complex: they are mostly closed class words such as the, in, a etc.

Therefore, leaving aside the method trained on baseline part of speech cluster tags as almost identical to the untagged approach, all other methods that incorporate part of speech information result in an increase in precision and a decrease in recall in comparison to the method trained on untagged text. This effect can largely be attributed to the high number of single item conflation sets produced by these methods and is particularly exaggerated in the methods trained on induced part of speech cluster tagged text. The single item conflation sets skew the results by increasing the precision score (no extra members are inserted into the gold standard conflation sets by WWM), whilst decreasing recall (as a result of an increased number of words deleted from gold-standard conflation sets). This tendency is illustrated in its extreme form by the results when each word is assigned to its own conflation set.

Table 5.6 shows the results when only those conflation sets containing more than one item are considered, removing the effect of any skew in the scores as a result of single item conflation sets. The best F-score performance is achieved by the method trained on induced part of speech cluster tagged text with morphological information used to assign infrequent words to clusters. This method has a better F-score than the method trained on gold standard tagged text.

By comparing the methods trained on induced part of speech cluster tagged text with and without morphological information, it is possible to analyse the effect of the this information on part of speech clustering when incorporated into morphological induction. The inclusion of morphological information does result in a small improvement in the accuracy of the resulting morphological strategies. This is probably because moving words between clusters according to morphological similarity increases the evidence for the existence of certain morphological strategies.

It is interesting to note that, even without the inclusion of single item conflation sets, the incorporation of part of speech information increases the precision of the induced conflation sets. This indicates that the restrictions imposed on morphological strategies by part of speech information decrease the number of incorrect words inserted into conflation sets and improve accuracy of the morphological relationships discovered. Additionally, the strong performance
Table 5.6: Comparison of WWM conflation sets with >1 item with CELEX conflation sets

<table>
<thead>
<tr>
<th>Training Text</th>
<th>No. conflation sets with &gt;1 item</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>No tags</td>
<td>804</td>
<td>80.21</td>
<td>96.41</td>
<td>87.57</td>
</tr>
<tr>
<td>Gold-standard tags</td>
<td>719</td>
<td>85.95</td>
<td>96.46</td>
<td>90.90</td>
</tr>
<tr>
<td>Baseline POS cluster tags</td>
<td>803</td>
<td>80.23</td>
<td><strong>96.43</strong></td>
<td>87.58</td>
</tr>
<tr>
<td>Induced POS cluster tags</td>
<td>565</td>
<td>86.43</td>
<td>95.25</td>
<td>90.58</td>
</tr>
<tr>
<td>Induced POS cluster tags with morphology</td>
<td>559</td>
<td><strong>87.80</strong></td>
<td>92.25</td>
<td><strong>91.37</strong></td>
</tr>
</tbody>
</table>

Table 5.7: Accuracy of word generation compared with CELEX

<table>
<thead>
<tr>
<th>Training text</th>
<th>No. strategies</th>
<th>No. new words generated</th>
<th>No. correct new words</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No tags</td>
<td>135</td>
<td>136,441</td>
<td>3,588</td>
<td>2.63</td>
</tr>
<tr>
<td>Gold-standard tags</td>
<td>205</td>
<td>37,870</td>
<td>6,180</td>
<td>16.32</td>
</tr>
<tr>
<td>Baseline POS cluster tags</td>
<td>135</td>
<td>136,305</td>
<td>3,598</td>
<td>2.64</td>
</tr>
<tr>
<td>Induced POS cluster tags</td>
<td>122</td>
<td>42,746</td>
<td>4,386</td>
<td>10.26</td>
</tr>
<tr>
<td>Induced POS cluster tags with morphology</td>
<td>125</td>
<td>38,933</td>
<td>4,357</td>
<td><strong>11.19</strong></td>
</tr>
</tbody>
</table>

of the methods trained on induced part of speech cluster tagged text indicates that induced clusters do provide part of speech information that is useful. However, these methods produce a significantly smaller number of conflation sets with more than one item than all other methods, indicating that the induced clusters provide worse coverage when discovering morphological relationships than the gold-standard tags.

Whilst analysis of those conflation sets with more than one item does provide some insights into the value of part of speech information in morphological induction, it is important to note that this is not a true indication of the performance of the systems. There are clearly a large number of conflation sets with more than one item that the methods trained on noisy part of speech information have failed to discover.

Generating Words

A second evaluation method for the morphological induction methods is based on their ability to generate valid words of the language that were not seen in training. One approach to evaluation is to compare generated words outside the training lexicon with a machine-readable lexicon: the CELEX database. In evaluating against this database, part of speech information is ignored and only the accuracy of the orthographic froms of generated words is evaluated. In terms of the precision-recall tradeoff, this measures precision. Neveu and Fulop (2002) mention the difficulty of defining a satisfactory recall metric for new word generation: there is no finite set of words in a language, nor is it possible to define the set of words that WWM might be expected to learn, given some training text. Whilst recall, like precision, could be measured against the CELEX lexicon, such a metric would not be informative.

Table 5.7 shows the results of new word generation using training text that has been tagged using various methods. The method trained on gold-standard tagged text clearly performs best, with the lowest total number of generated words and the largest proportion correct. This is despite the fact that this method discovers a significantly higher number of morphological strategies than the others, indicating that the strategies it has found are restricted appropriately by the inclusion of part of speech information. Overall, this method also generates the
Training text | No. strategies | No. new words generated | No. correct new words | Precision (%)
---|---|---|---|---
No tags | 135 | 136,441 | 38,227 | 28.05%
Gold-standard tags | 205 | 37,870 | 22,572 | 59.60%
Baseline POS cluster tags | 135 | 136,305 | 38,228 | 28.05%
Induced POS cluster tags | 122 | 42,746 | 19,119 | 44.73%
Induced POS cluster tags with morphology | 125 | 38,933 | 18,312 | 47.03%

Table 5.8: Accuracy of word generation according to spellchecker

largest number of correct new words. It is noteworthy that the methods trained on part of speech cluster tagged text (both with and without the incorporation of morphological information) also perform significantly better than the methods trained on untagged and baseline part of speech cluster tagged text. Clearly even the incorporation of noisy part of speech information improves the accuracy of the morphological strategies. From a comparison of the performance of the methods trained on part of speech cluster tagged text with and without morphological information, it is clear that the morphological information has restricted the strategies discovered by WWM, resulting in a higher proportion of correct generated words, but also decreasing the total number of correct words.

As discussed above, the similarity in the taggings produced by the methods trained on untagged text and on baseline cluster tagged text leads to almost identical performance of the two methods. Clearly both of these methods lead to over-generation of new words, producing more than three times the number of new words than the other methods. Examples of nonsense words produced by applying inappropriate morphological strategies include somewhat understandable but invalid forms such as: entailness, enterprisation, chestly, nightized, 29thness, enthusiasticity, chemistryman, blueable and many more obscure examples such as everythingic, rulingst, sidingrs, hurdleern, workshopt.

To improve on the limited coverage of the CELEX database, newly generated words were also evaluated using an automated spellchecker. This was done using the spellchecker ‘spell’ on the Sun OS5.9 operating system. This spellchecker checks words against a spelling list and checks whether they are derivable from a word in the list by applying certain inflections, prefixes or suffixes. The spelling list is based on various sources and has better coverage of proper names and technical terms than a dictionary. The results when comparing generated words against this spellchecker are presented in Table 5.8.

Evaluating against the spellchecker, the accuracy figures for generated new words are higher than those resulting from comparison with CELEX, providing a better indication of the system’s overall accuracy in word generation. The various methods are ranked the same with respect to one another when comparing against the spellchecker and CELEX. However, the methods trained on untagged text and baseline part of speech cluster tagged text perform slightly better in relation to the other methods. Neuvel and Fulop report 70 - 80% accuracy in new word generation by WWM trained on an English lexicon of approximately 3,000 words including detailed part of speech information. However it is not clear how they make validity judgements for new words. Even trained on gold-standard text, accuracy figures for WWM in this thesis are not this high. Nonetheless, regardless of the resource against which new word accuracy is assessed, the ranking of the different methods are unchanged.
Chapter 6

Conclusions and Future Work

6.1 Conclusion

Induction methods in natural language processing avoid any requirement for annotated training data. As a result their application extends to languages and linguistic structures for which such data is not available. However, unannotated text provides such learning methods with limited information from which to learn.

This thesis investigated the effect of bootstrapping part of speech induction and morphological induction methods in order to provide both methods with additional information. This was done through the incorporation of explicit annotations from one learner into the training text provided to the other.

The results of bootstrapping the two learners using English training data are encouraging. The approach taken to incorporating morphological annotations into a part of speech induction system used in this thesis was aimed at improving the clustering of words that occurred infrequently in the training text. The resulting improvement in the accuracy of the part of speech ‘tags’ assigned to infrequent words and indeed over the whole lexicon confirmed that these annotations provide useful information for this task. Unfortunately the linguistically informed morphological annotations were outperformed by a naive approach to morphological annotation, when incorporated into the part of speech induction component. However, this may be explained by the relative lack of complexity of the morphological system of English, which may weaken the potential power of linguistically accurate morphological annotations.

The effect of incorporating part of speech annotation into morphological induction is a little hard to assess due to the lack of widely applicable evaluation metrics for morphology induction methods and the many aspects of the task of morphological learning. However, in the two areas evaluated for this study - the identification of morphologically related words and word generation - there were some indications of performance improvements as a result of the incorporation of part of speech information. This information lead to a decrease in the coverage and number of morphological relationships that were discovered, but an increase the accuracy of those relationships that were discovered. From comparison with the performance of the morphological induction system on gold-standard part of speech tagged text, which improved both coverage and accuracy, it is clear that accurate part of speech information can lead to an overall improvement in this aspect of performance of a morphology induction system.

The incorporation of induced part of speech tags into the morphological induction system resulted in a clear improvement in the accuracy of word generation. The results of this evaluation also indicated that the incorporation of morphological information into a part of speech induction method, in turn improved its performance when incorporated back into the morphological induction system.

In general, the results of this study tend to indicate that the part of speech induction system is affected by the level of coverage of the morphological annotations and that better coverage can lead to performance improvements on a knowledge-free part of speech induction system.
In contrast, the morphological induction system seems to benefit from highly accurate part of speech information and improved tag accuracy leads to better coverage when discovering morphologically related words.

This thesis has shown that combining part of speech induction and morphological induction by incorporating annotations produced by one induction method into the other can benefit both induction methods. The correlation between linguistic information that is conveyed via syntax and morphology indicates potential advantages arising from interactions between the two induction components. Whilst the exact extent of the performance improvements depends on the induction task under examination, overall combining the two induction methods can and does lead to improved performance.

6.2 Future Work

A number of avenues for exploration in future work have arisen out of this study. The part of speech induction algorithm used in this study only incorporates morphological information after the final iteration of the clustering process. A more successful strategy might be to move infrequent words according to their morphology after each iteration. By alternately clustering frequent and infrequent words, this strategy has the potential to iteratively improve the both clustering of infrequent words and also the clustering of other words.

Further morphological information in the form of a word’s stem might also provide evidence for the assignment of a word to a particular word class. For example if walked has been assigned to a certain cluster, then this may be evidence for the assignment of walking to the same cluster. Clearly in a language such as English which contains a large amount of derivational morphology this evidence may not be very strong. However in languages with richer inflectional morphology, such evidence may be of great benefit.

There are a number of obvious extensions to the morphological induction system: in particular, modifying the string comparison procedure to enable the discovery of different kinds morphology. A first step might include other kinds of affixes (e.g. prefixes and circumfixes), but the incorporation of the discovery of non-concatenative morphology would be a more significant step. It would also be interesting to perform the same experiments using a phonetically transcribed training text to gauge the effects of orthographic normalisation.

Additionally, a comparison of the performance of both induction components using training text from a variety of languages would enable an evaluation of the effect of language typology in these methods and provide insights into their inherent strengths, weaknesses and limitations. It is quite possible that the linguistic structure of English is less suitable for combining these two methods than that of some other languages. The availability of morphologically annotated corpora in a wider range of languages, would also enable more accurate assessment of performance improvements that can be achieved through the incorporation of morphological information into other learners.

A final important extension to the work reported in this thesis would involve running the same experiments using a significantly larger amount of training data. The expectation would be that the performance of both induction components would improve as a result. In turn these performance improvements would indicate that benefits may be derived from the combination of the two components in a framework in which they are more heavily dependent upon on another. For example, a co-training framework, in which the components iteratively improve the quality of the annotations provided to one another. This iterative process could be expected to improve the performance of both components significantly.
Bibliography


Appendix A

Calculation of Perplexity in a Class-Based Bigram Language Model

The derivation steps leading to the simplified formula for the perplexity of a class-based bigram language model as discussed in section 4.1.1 are shown below.

Perplexity, $B$, is given by the reciprocal of the geometric average of the per word probabilities in the text. So, where the text is $k$ words long:

$$ B = \text{Pr}(w_1^k)^{-\frac{1}{k}} $$  \hspace{1cm} (A.1)

taking the logarithm of both sides of the equation gives:

$$ \log B = \log \text{Pr}(w_1^k)^{-\frac{1}{k}} $$  \hspace{1cm} (A.2)

$$ = -\frac{1}{k} \log \text{Pr}(w_1^k) $$  \hspace{1cm} (A.3)

substituting bigram probability estimates for $\text{Pr}(w_1^k)$ gives:

$$ = -\frac{1}{k-1} \prod_{i=2}^{k} \log \text{Pr}(w_i|w_{i-1}) $$  \hspace{1cm} (A.4)

$$ = -\frac{1}{k-1} \sum_{w_{i-1},w_i} C(w_{i-1}, w_i) \log \text{Pr}(w_i|w_{i-1}) $$  \hspace{1cm} (A.5)

which, in a class-based model is given by:

$$ = -\frac{1}{k-1} \sum_{w_{i-1},w_i} C(w_{i-1}, w_i) \log Pr(c_i|c_{i-1}) \cdot Pr(w_i|c_i) $$  \hspace{1cm} (A.6)

$$ \approx - \left[ \sum_{w_{i-1},w_i} \frac{C(w_{i-1}, w_i)}{(k-1)} \left[ \log \text{Pr}(w_i|c_i) + \log \text{Pr}(c_i) \right] ight] $$  \hspace{1cm} (A.7)

$$ + \sum_{w_{i-1},w_i} \frac{C(w_{i-1}, w_i)}{(k-1)} \left[ \log \text{Pr}(c_i|c_{i-1}) - \log \text{Pr}(c_i) \right] $$

$$ \approx - \left[ \sum_{w_i} \sum_{c_{i-1}} \frac{C(w_{i-1}, w_i)}{(k-1)} \log \text{Pr}(w_i|c_i) \frac{Pr(c_i)}{Pr(c_i)} \right] $$  \hspace{1cm} (A.8)
\begin{align*}
&\approx - \left[ \sum_{w_i} \Pr(w_i) \log \Pr(w_i) + \sum_{c_{i-1}, c_i} \Pr(c_{i-1}, c_i) \log \frac{\Pr(c_{i-1}, c_i)}{\Pr(c_{i-1}) \Pr(c_i)} \right] \quad (A.9)
\end{align*}

removing the factor $-\frac{1}{k}$ and substituting the maximum likelihood estimates gives:

\begin{align*}
&\approx \sum_{w_i} C(w_i) \cdot \log C(w_i) + \sum_{c_{i-1}, c_i} C(c_{i-1}, c_i) \cdot \log \frac{C(c_{i-1}, c_i)}{C(c_{i-1}) \cdot C(c_i)} \quad (A.10)
\end{align*}

Equation (A.9) relies on the approximations $\sum_{w_{i-1}} \frac{C(w_{i-1}, w_i)}{k} \approx \Pr(w_i)$ and $\frac{C(c_{i-1}, c_i)}{k-1} \approx \Pr(c_{i-1}, c_i)$ which hold for large $k$. Additionally $\Pr(w_i | c_i) = \Pr(w_i, c_i) = \Pr(w_i)$ holds as $w_i$ is assigned to class $c_i$. For further discussion see Manning and Schutze (1999) and Martin et al. (1998).