**Introduction**

Automatic Morphological Analysis (CMA) is still a much discussed topic in natural language

processing.1 The objective of CMA is to understand the inner mechanism of word form formation in a language. A morphological analyzer can provide valuable information for other computer based linguistic tasks such as lemmatization, syntactic parsing, machine translation, information retrieving, text clustering and many others.

Goldsmith (2001) classifies the work in automatic morphological analysis into four categories.

The classification is done with an emphasis on how the morphological rules are obtained. The first category identifies morpheme boundaries on the basis of the degree of predictability of the n+1st letter given the first n letters. The second employs n-gram grammar and high likelihood to obtain morpheme internal structure. The third category endeavors to discover rules through phonological relationships between pairs of related words. The fourth one seeks for an analysis for a language which is the most concise and consequently boils down to a set of rules for the language.12

Two factors are essential to achieve accurate automatic morphological analysis. One factor is

the construction of a set of morphological rules and the other is the morphological analysis

procedure. The absence or underperformance of either of them impairs the overall ability of the

morphological analyzer. Thus in constructing an algorithm for the task, both factors should be

addressed.5

Machine learning is a promising alternative to obtain morphological rules. This method can

avoid problems such as costly human labor, rule inconsistency and can provide additional statistical information which can be used in morphological analysis procedure. Based on information type employed in the machine learning task, we can usefully obtain two classes: supervised learning and unsupervised learning. The class of supervised learning (Bosch, 1999; Wicentowski, 2004) makes use of lexical database with morphological information. A good example of such lexical database is CELEX2. The class of unsupervised learning (Goldsmith, 2001; Kurino et al, 2004; Creutz, 2005)1 uses only a wordlist, with or without information of frequency along the way.

In regard of morphological analysis procedure, there are two popular techniques. One

technique is stemming, represented by Porter (1980)3. The algorithm given in Porter (1980) consists of two stages: (1) the de-suffixing step which subtracts predefined endings from words, and the recoding step which adds possible ending part to the string obtained in the previous stage. These two stages can be done in sequence or simultaneously. One distinct feature of the algorithm is that it doesn’t use a dictionary, which makes it very sufficient in analysis.

The other technique is morphological parsing, represented by Model of Two-level Morphology

proposed in Koskenniemi (1983). This model considers morphological analysis and morphological transformation in different phonological situations and encodes the correspondence between surface form and lexical form with finite-state transducer. For example:

Lexical form: ***s p e c i f y + s***

Surface form: ***s p e c i f i e s***

The first morphological analysis system that adopts this model is KIMMO (Karttunen, 1983). The system has two parts: rules and lexical information such as morpheme structure and

morphosyntactic constraint.

The morphological analyzer illustrated in this paper falls into the first class of Gold(2001)

classification. The system aims at high accuracy of morphological analysis of English language

with morphological rules obtained through unsupervised machine learning. The analyzer applies

letter transitional probability proposed in Keshava&Pilter(2005) in morphological rule learning and in disambiguation of morphological analysis as well. An initial evaluation of the analyzer shows a promising result with an 88.42% precision, 78.46 recall and 83.14% F-score, which transcends the best results of English language reported in Unsupervised Segmentation of Words into Morphemes – Challenge 2005.

**Analysis Procedure Control**

A wide-covering and correct set of affix rules is prerequisite for accurate morphological analysis.

But it alone does not guarantee a successful analysis. The procedure in which the analysis is done is also crucial.

**3.1 Disambiguation**

Like word segmentation in Chinese, there are ambiguities in morphological analysis. A good

understanding of the types of ambiguities certainly helps to solve the ambiguities. In nature, the

morphological analysis is analogous to Chinese word segmentation. The categorization of

ambiguity in Chinese segmentation may also apply here. Thus we have two types of ambiguity in morphological analysis: intersectional ambiguity and intersectional ambiguity.

**3.1.1 Intersectional Ambiguity**

Intersectional ambiguity stands for the type of strings which have more than one possible

intersectional analysis. For example, given a string ABCD, B, CD, C, and D are all potential

affixes.4 Thus the key of intersectional ambiguity is to decide where the morphological boundary is.

Вступление  
Автоматический Морфологический Анализ (CMA) - все еще очень обсуждаемая тема на естественном языке  
обработка.1 Целью CMA является понимание внутреннего механизма формирования словоформ в языке. Морфологический анализатор может предоставить ценную информацию для других компьютерных лингвистических задач, таких как лемматизация, синтаксический анализ, машинный перевод, поиск информации, кластеризация текста и многие другие.  
  
Голдсмит (2001) классифицирует работу по автоматическому морфологическому анализу на четыре категории.  
Классификация проводится с акцентом на то, как получаются морфологические правила. Первая категория определяет границы морфемы на основе степени предсказуемости n + 1-й буквы с учетом первых n букв. Второй использует грамматику n-граммы и высокую вероятность получения внутренней структуры морфемы. Третья категория стремится обнаружить правила посредством фонологических отношений между парами связанных слов. Четвертый ищет анализ для языка, который является наиболее кратким и, следовательно, сводится к набору правил для языка.  
  
Два фактора необходимы для достижения точного автоматического морфологического анализа. Одним из факторов является  
построение набора морфологических правил, а другой - морфологический анализ  
процедура. Отсутствие или неэффективность любого из них ухудшает общую способность  
морфологический анализатор. Таким образом, при построении алгоритма для задачи оба фактора должны быть  
addressed.5  
  
Машинное обучение является многообещающей альтернативой для получения морфологических правил. Этот метод может  
избегать таких проблем, как дорогостоящий человеческий труд, несоответствие правил и может предоставить дополнительную статистическую информацию, которая может быть использована в процедуре морфологического анализа. Основываясь на типе информации, используемой в задаче машинного обучения, мы можем с пользой получить два класса: контролируемое обучение и неконтролируемое обучение. Класс контролируемого обучения (Bosch, 1999; Wicentowski, 2004) использует лексическую базу данных с морфологической информацией. Хорошим примером такой лексической базы данных является CELEX2. Класс неконтролируемого обучения (Goldsmith, 2001; Kurino et al, 2004; Creutz, 2005) 1 использует только список слов, с информацией о частоте или без нее по пути.  
  
Что касается процедуры морфологического анализа, существуют две популярные методики. Один  
Техника основывается, представленная Портером (1980) 3. Алгоритм, приведенный в работе Портера (1980), состоит из двух этапов: (1) этап удаления суффиксов, который вычитает предопределенные окончания из слов, и этап перекодирования, который добавляет возможную конечную часть в строку, полученную на предыдущем этапе. Эти два этапа могут быть выполнены последовательно или одновременно. Отличительной особенностью алгоритма является то, что он не использует словарь, что делает его очень достаточным для анализа.  
  
Другой метод - это морфологический анализ, представленный моделью двухуровневой морфологии.  
предложено в Коскенниеми (1983). Эта модель рассматривает морфологический анализ и морфологическое преобразование в различных фонологических ситуациях и кодирует соответствие между формой поверхности и лексической формой с помощью конечного преобразователя. Например:  
Лексическая форма: s p e c i f y + s  
Форма поверхности: вид сверху  
  
Первой системой морфологического анализа, которая принимает эту модель, является KIMMO (Karttunen, 1983). Система состоит из двух частей: правила и лексическая информация, такая как структура морфемы и  
морфосинтаксическое ограничение.  
  
Морфологический анализатор, проиллюстрированный в этой статье, относится к первому классу золота (2001).  
классификация. Система нацелена на высокую точность морфологического анализа английского языка  
с морфологическими правилами, полученными посредством неконтролируемого машинного обучения. Анализатор применяется  
Переходная вероятность буквы предложена в Keshava & Pilter (2005) при изучении морфологических правил, а также при устранении неоднозначности морфологического анализа. Первоначальная оценка анализатора показывает многообещающий результат с точностью 88,42%, 78,46 отзыва и 83,14% F-показателя, что превосходит лучшие результаты по английскому языку, о которых сообщалось в разделе «Неуправляемая сегментация слов в морфемы - задача 2005».  
  
Контроль процедуры анализа  
Обширный и правильный набор правил аффиксов является необходимым условием для точного морфологического анализа.  
Но одно это не гарантирует успешного анализа. Процедура, в которой проводится анализ, также имеет решающее значение.  
  
3.1 Неоднозначность  
Подобно сегментации слов в китайском языке, в морфологическом анализе есть неясности. Хороший  
Понимание типов неоднозначностей, безусловно, помогает решить эти двусмысленности. В природе  
морфологический анализ аналогичен сегментации китайского слова. Категоризация  
Неоднозначность в китайской сегментации также может применяться здесь. Таким образом, мы имеем два типа неоднозначности в морфологическом анализе: неоднозначность пересечения и неоднозначность пересечения.

3.1.1 Межсекционная неоднозначность  
Поперечная неоднозначность обозначает тип строк, которые имеют более одной возможной  
межсекционный анализ. Например, учитывая строку ABCD, B, CD, C и D являются потенциальными  
аффиксы.4 Таким образом, ключ межсекционной неоднозначности состоит в том, чтобы решить, где находится морфологическая граница.