Error Mining for Parsing Results and Unknown Word Identification in Chinese Word Segmentation

An Extended Research Proposal for MSc HLST Dissertation

A. Student

January 13, 2010

Abstract

Error mining was originally proposed for finding forms that cause incomplete parses of sentences. I generalized it as a technique that identifies patterns representing the differences between two sub-corpora and attempted to apply the method to solve other language processing problems that share similar structures with the parsing one. To build a platform for error mining, I will focus on extending the iterative method of De Kok et al (2009), though maximum entropy based feature extraction methods will also be considered. I will conduct error mining on real corpora for both analyzing parsing results and identifying unknown words in Chinese word segmentation. Finally I will evaluate the performance of the miner and report the results.

1 Description

1.1 Introduction

The concept of error mining first appeared in grammar engineering filed. Though wide-coverage grammars can cover a large number of grammatical and lexical phenomena, they often fail to reach same high accuracy for domain-specific texts as for general domain texts, because of missing lexicon entries, fixed expressions, and grammatical constructs. One type of parsing errors is that the parser can not find an analysis that account for the full sentence. It indicates that the grammar or lexicon is incomplete. As it is a tedious task to find incomplete descriptions for wide coverage grammar by hand, error mining was proposed as a technique to automatically identify problematic grammar or lexicon.

The problem structure of error mining for parsing results is as following: After a corpus is parsed, it is split up into two sub-corpora, which are parsable and unparsable sentences, respectively. If some patterns that only exist in the unparsable sub-corpus can represent the difference between two sub-corpora, these patterns are likely to be the cause of the parsing errors. While error mining was originally designed for identifying these patterns, we generalized it as a technique that finds patterns in one sub-corpus that can evidently influence the binary classification of sentences from the original corpus. Therefore, we can apply error mining to other language processing tasks sharing similar problem structure with the parsing one.

One such task we will focus on is unknown word identification in Chinese word segmentation (Peng et al 2004). Unlike English and other western languages, the basic elements of a Chinese sentence are Chinese characters. Words are made of one or more characters. The problem is that there are only consecutive characters in sentences with no spaces indicating word boundaries. Word segmentation is therefore a key precursor for Chinese language processing. While current systems can well resolve segmentation ambiguities, they often fail to give the correct segmentation
when unknown words, which are words not in the lexicon, occur in the sentences. We can view the segmentation results of a Chinese corpus in the same way as parsing results. Two sub-corpora are now sentences with incorrect segmentation and sentences with correct segmentation. The error mining now is to discover patterns in the sub-corpus of sentences with incorrect segmentation that can represent the difference of two sub-corpora. We expect unknown words to be found in the patterns. And then we can further filter them out using certain criteria.

This project will develop a platform for error mining. The platform will mostly rely on extending the iterative method of De Kok et al. (2009), though other approaches, such as maximum entropy algorithms, will also be considered. Then I will adopt this platform to find causes of parsing failure and to identify unknown words in Chinese word segmentation. I will evaluate the performance of the proposed methods by conducting experiments with large corpora.

1.2 Background

1.2.1 Previous work in error mining

Two major error mining techniques were proposed by Van Noord (2004) and Sagot and de la Clergerie (2006), respectively. Both methods follow the same basic principle: first, a large (unannotated) corpus is parsed. After parsing, the sentences can be split up into a parsable sub-corpora and unparsable sub-corpora. Words or n-grams that occur in the unparsable sub-corpora, but that do not occur in the parsable sub-corpora have a high suspicion of being the cause of the parsing error.

Van Noord (2004) defines an n-gram’s suspicion of being the cause of a parsing error as a ratio:

\[
S(w) = \frac{C(w_i...w_j | \text{error})}{C(w_i...w_j)}
\]

where \(C(w_i...w_j)\) is the total number of occurrences of n-gram \(w_i...w_j\) in sentences of both sub-corpora, and \(C(w_i...w_j | \text{error})\) is the number of occurrences of the n-gram in the unparsable sub-corpora. The longer n-gram \(w_k...w_i...w_j...w_k\) is only considered if its suspicion is higher than each of its substrings. Of course the basis case for n-gram is a word. While this method works well with n-grams that only occur in the unparsable sub-corporus, it also gives n-grams that just accidentally appear in unparsable sub-corpora a high suspicion, which is not desired.

The error mining method proposed by Sagot and de la Clergerie (2006) solves the above problem. It does so by taking the following into account:

- A form also occurs in parsable sentences is less likely to be the cause of an error.
- The suspicion rates of forms in the same sentence are interdependent and their sum is constant.

To be able to handle the suspicion of an n-gram within its context, this method introduces the notion of observation suspicion, which is the suspicion of an n-gram within a given sentence. The suspicion of an n-gram, outside the context of a sentence, is then defined to be the average of all observation suspicions. The observation suspicions themselves are dependent on the n-gram suspicions, making the method an
We ignored many technical details of the iterative process here, but the important information is that the method is based on the recursive dependence between suspicion and observation suspicions and it works with only unigrams and bigrams, since it encounters data sparse problems when trying longer n-grams.

De Kok et al (2009) take sparseness into account, producing n-grams that are as long as necessary to identify problematic patterns, but not longer. It is implemented by combining previous two methods and introducing an expansion function.

We will extend the current methods by re-defining patterns that represent the difference between two sub-corpora, i.e. modeling the language of two sub-corpora. There are various techniques in language modeling, among which the most popular one is n-grams (Rosenfeld 2000), due to its simplicity and effectiveness. We will generalize n-grams to fit the context of individual tasks. For example, for error mining in parsing results, we have already introduced the hybrid n-grams of words and part-of-speech tags to represent more abstract patterns and have achieved preliminary results.

1.2.2 Review of unknown word identification in Chinese word segmentation

In Chinese word segmentation, unknown words cause segmentation errors because out-of-vocabulary words in an input text normally are incorrectly segmented into pieces of single character word or shorter words.

Chen and Bai (1998) reports that according to an examination of a group of testing data which is a part of the Sinica corpus, 4572 occurrences out of 4632 unknowns were incorrectly segmented into sequences of shorter words, and each sequence contained at least one monosyllabic word. That is, 60 of the unknown words were segmented into sequences of multi-syllabic words only. Therefore, occurrences of monosyllabic words (i.e., single character words) in the segmented input text may denote the possible existence of unknown words.

We will use this finding as the starting point for unknown word detection. The linguistic elements and patterns will be defined and a proper pattern expansion strategy will be proposed, based on the study of the types and traits of unknown words. Then the iterative process is used to filter out unknown words from candidates generated by the expansion strategy.

1.3 Aims and objectives

The overall aim is to build a platform for effective error mining, and to apply the proposed techniques to cause of parsing errors analysis and unknown words identification in Chinese word segmentation. The objectives include:

1) Proposing patterns to model the differences of sub-corpora for two target problem domains.
2) Extending pattern expansion methods for two target applications.
3) (Optional) Tailoring maximum entropy based algorithms for error mining.
4) Implementing the error mining platform, realizing the core iterative process,
pattern models, pattern expansion functions, as well as other necessary pre-processing and post-processing functions.

5) Conducting experiments on finding causes of parsing errors, using large corpora in Dutch and other languages if possible; evaluating the performance of the error mining platform.

6) Conducting experiments on unknown word identification in Chinese word segmentation and evaluating the performance.

7) Writing the thesis.

1.4 Method

The main algorithms of the error mining platform will base on iterative methods described in 1.2.1, with further extensions and modifications according to two target tasks. A special tailored maximum entropy method based on public accessible implementation will also be introduced, subject to the time availability. The forms of patterns to represent sub-corpora differences will be designed on an n-gram language model basis, considering information on several levels, such as word surface level and part of speech level. Based on the study of types and characteristics of unknown words in real corpora, a linguistically heuristic strategy will be proposed as the pattern expansion function for the unknown word identification task.

1.5 Deliverables

The deliverables will be:

1) An error mining platform as described in 1.3, developed in C++ using Qt4.5.

2) An HLST thesis reporting all the methods, experiments and evaluations in this project.

2. Work Plan

2.1 Sub tasks

1) To design and implement basic error mining platform for analyzing parsing results.

2) To extending the error mining platform and to conduct iterative error mining on parsing results of large corpora.

3) (Optional) To incorporating maximum entropy based methods into the platform, to conduct maximum entropy based error mining on parsing results with same corpora as in 2) and to compare the results.

4) To further review literatures on the unknown word identification problem, and to design patterns and pattern expansion strategies.

5) To conduct error mining experiments on unknown word identification in Chinese word segmentation, using standard corpora and to evaluate the performance.

6) To write the thesis
2.2 Schedule

- Jan 15th~ Feb 15th (4 weeks): Finishing Subtask 1)
- Feb 16th~ Feb 30th (2 weeks): Finishing Subtask 2)
- March 1st~ March 15th (2 weeks): (Optional) Finishing Subtask 3)
- March 16th~ April 15th (4 weeks): Finishing Subtask 4)
- April 16th~ April 30th (2 weeks): Finishing Subtask 5)
- May 1st~ May30th (4 weeks): Finishing Subtask 6)

References


