TR－SLT－0070

Unsupervised Chinese Word Segmentation

Rui Yang, Hirofumi Yamamoto

２００４年3月３１日

概要

This report presents my work in ATR during the last one and a half months. We tried to use an unsupervised method base on EM algorithm to get a segmentation result with low total entropy. We haven't got the completed result of the experiment till now, but from the trend line, we can say it is effective. But still we have much work to do to reduce the high complexity.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２ TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

Index :

Abstract

This report presents my work in ATR for one and a half months. We tried to use an unsupervised method base on EM algorithm to get a segmentation result with low total entropy. We haven't got the completed result of the experiment till now, but from the trend line, we can say it is effective. But still we have much work to do to reduce the high complexity.

Introduction

Chinese segmentation is the first step for Chinese Processing. Different from those Indo-European languages, a given Chinese text is a string of characters without any word boundaries between words, such as white space.

Our *purpose* is to get a segmented word string with low total entropy.

We chose an unsupervised method and used an EM algorithm to dynamically change the value until convergence.

We can't get a completed result of the experiment, but from the trend line, we can say it is effective, though we have quite a little work to do.

Background

Chinese segmentation is the first step for Chinese Processing. Different from those Indo-European languages, a given Chinese text is a string of characters without any word boundaries between words, such as white space.

The Chinese segmentation methods can be divided into three kinds:

1. rule-based
2. statistic-based
3. mixture of both above

In Machine translation, we have the equation of

$$P(T|S) = P(S|T)P(T)$$

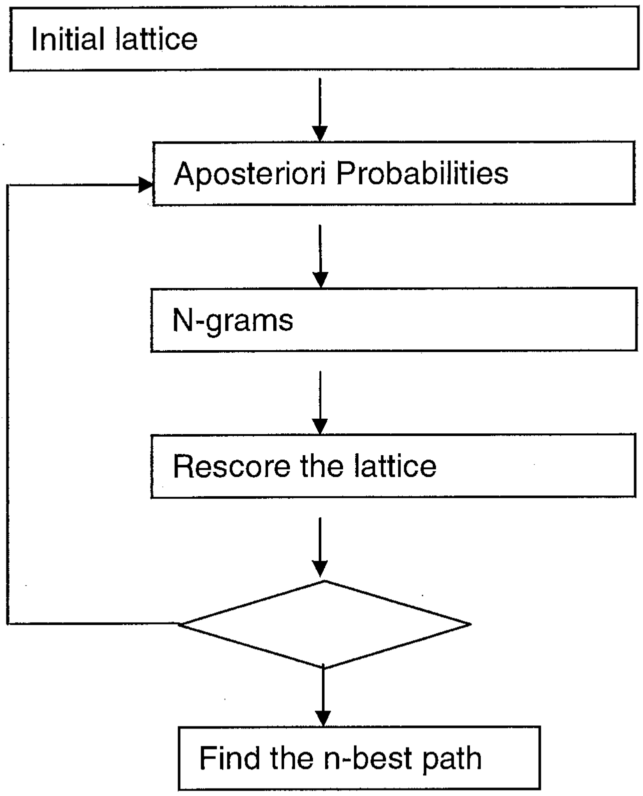
S: source language

T: target language

- Our *purpose* is to reduce the total entropy of P(T)

If we simply consider the main purpose, unsupervised method should give low total entropy. We can used an EM algorithm to dynamically change the value until convergence.

Framework of the Method



Initial lattice

Aposteriori Probabilities

N-grams

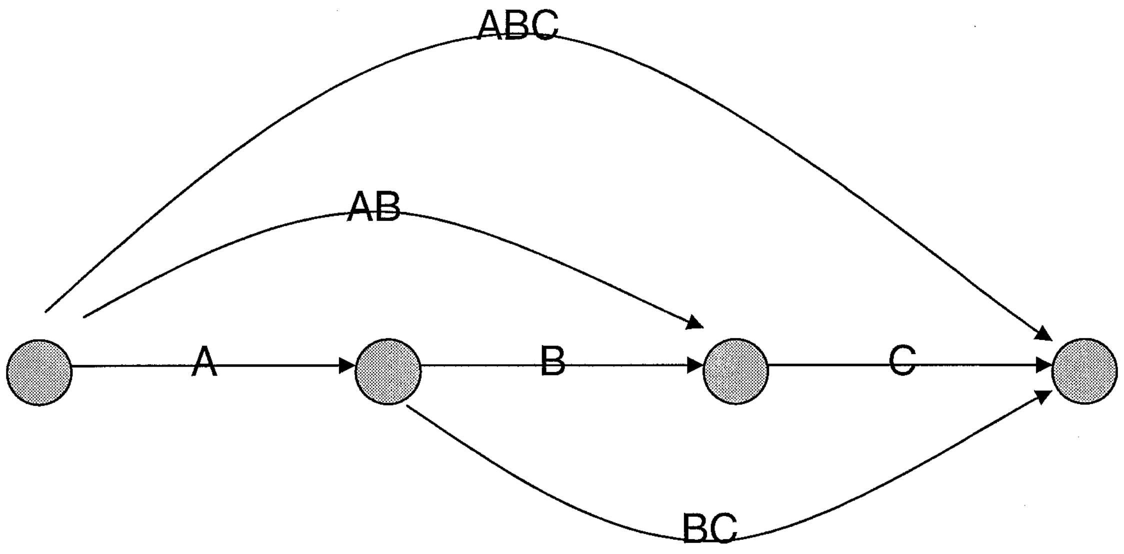
Rescore the lattice

Find the n-best path

Word Lattice

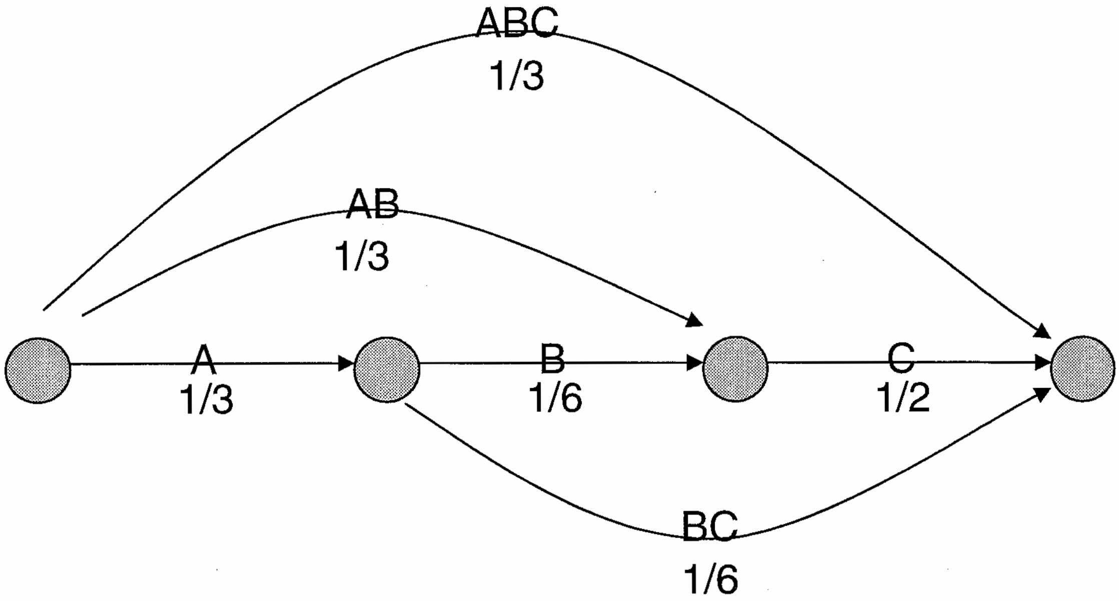
Word lattice is a word graph defined by lexicon. It gives a searching space to find a suitable path. When there is no lexicon, word lattice will be the whole set of probable path, which can be character, word, or phrase.

A simple example of word lattice is as following,

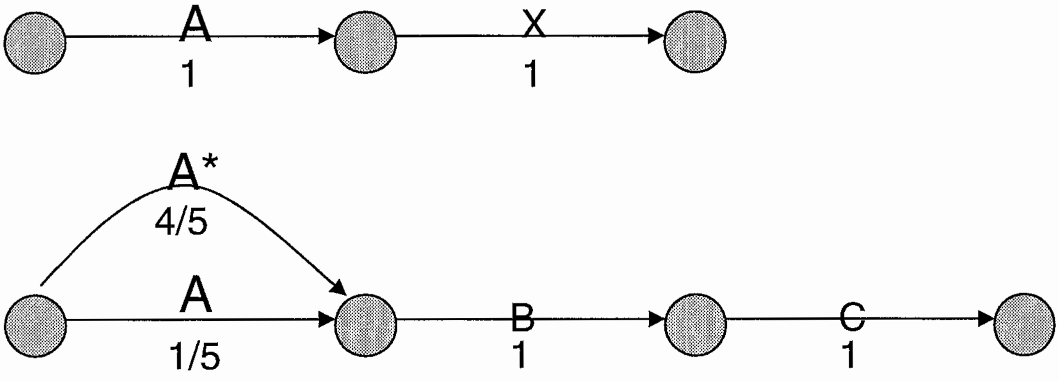


Aposteriori Probabilities

We conclude the probabilities from the initial lattice as *Aposteriori Probabilities.*



But the initial values sometimes are not reasonable; in this case, we build iteration procedure to change them dynamically. The new value comes from the old probability inside lattice.



$$P_{A\,new} = (0.2+1) / 5 = 0.24$$

$$P_{A^{*}new} = 0.8 / 5 = 0.16$$

In this example, the arc A occurs once in the first utterance, but only has a probability of 1/5 in the second utterance. The arc A then should be re-scored use the value 0.24.

Environment Settings

- Corpus:

  Part of BTEC, with 30,000 sentences

- Lexicon:

  90,000 words.

Future Work

We haven't got the completed result of the experiment, but from the trend line, we can say it is effective, although we have much work to do.

1. BTEC corpus has a high consistency. It's hard to distinguish the open and close test, if the test set also comes from the corpus, so the experiment should be moved to use another corpus.

2. Since we use an EM algorithm for the iteration, it is possible to drop into local maxima.

3. Do the same procedure without a lexicon is another idea.

4. To decrease the complexity of the experiment, we should,

handle the size and quality of the lexicon;
handle the length of utterances

Acknowledgments

I would like to thank *Yamamoto Hirofumi* and *Li Weishan* for all the directions they have given to me.

And also, I would like to thank *Frank Soong* and *Zhang Ruiqiang* for their important help.