Mandarin Question Sentence Detection: A Preliminary Study

Ping-Jer Yeh and Shyan-Ming Yuan

Department of Computer and Information Science
National Chiao Tung University
1001 Ta Hsueh Road, Hsinchu 300, Taiwan
{pjyeh, smyuan}@cis.nctu.edu.tw

Abstract. Detecting Mandarin question sentences is both interesting and difficult. To tackle this new topic, our strategy is first to try to increase recall and then precision. To achieve higher recall, we not only review relevant linguistic literature but also re-examine relevant issues from a new statistical and corpus point of view, and discover more comprehensive and precise question-related words than before. Next we present our statistical approaches and procedure, and discuss our findings. We achieve good recall and modest precision in the preliminary study, and pioneer the computational study of indefinitives.

1 Introduction

This paper presents a new topic in the field of natural language processing (NLP): Mandarin question sentence detection. Its importance is twofold. The first is human-computer dialogue system, and the second is punctuation processing. First, a non-toy human-computer dialogue or question answering system needs to distinguish between background information and foreground queries, in order to interact more like human-to-human conversation. Imagine that you are asking a digital assistant for help:

(1) I have installed and configured Wine, but Wine cannot find MS Windows on my drive. Where did I go wrong?

If the assistant program fails to pinpoint the foreground query and surrounding context, how can it work out a search plan to answer your “where” question?

Second, punctuation has been neglected in the NLP. For example, speech recognition software maps acoustic signals to text, but it seldom places appropriate punctuation marks in the output text. Word processors have built-in or plug-in spelling and grammar checkers, but they seldom try to check punctuation.

1 To our knowledge, no such prior art exists.
2 This phrase is excerpted from The Wine FAQ.
URL: http://www.winehq.com/site/docs/wine-faq/index
The reason why it has been neglected is that, punctuation is such a complex coding device that challenges computers. It is, as defined in *The American Heritage Dictionary* (Pickett et al. 16), “the use of standard marks and signs in writing and printing to separate words into sentences, clauses, and phrases in order to clarify meaning.” Therefore, it involves not only syntactic but also semantic and pragmatic levels of processing. For example,

(2) a. Is this yours? syntactic level
    b. I beg your pardon? semantic level
    c. This is yours? I don’t think so. pragmatic level

Sentence (2a) is obviously a question because of its verb BE-initial syntactic pattern. Sentence (2b), which begins without a verb BE, an auxiliary verb, or a WH word, is regarded as a question only if the meaning of the word “pardon” is taken into account. Furthermore, Sentence (2c) is regarded as a question only if the pragmatic context is taken into account.

It is even more challenging for the Mandarin language because there is no syntactically decisive and reliable marker and word order in Mandarin question sentences (Cheng 2), let along semantic and pragmatic clues.

The goal of this paper is to detect Mandarin question sentences. To put it more concretely, our task, in respect of training and validation, is to label unpunctuated input text with appropriate question marks. This paper is organized as follows. Section 2 reviews linguistic literature on Mandarin question sentences. Section 3 describes our feature-selection stages and findings, and in the meantime re-examines literature from a different angle: statistical point of view. Section 4 presents our training procedures and discusses our findings. Section 5 concludes our main contributions.

2 Linguistic Background

2.1 Use of Mandarin Question Marks

Modern Mandarin punctuation system, inspired by the western culture, was stabilized and formalized in the 20th century (Wu 17). Since then, prescriptive guidelines have been announced by authorities in major Mandarin-speaking regions, including Taiwan and mainland China (MPC 15; GB/T 15834-8). In general, question marks are used at the end of three kinds of question sentences: interrogative, dubitative, and rhetorical questions. For example,

(3) a. 這是什麼？ interrogative
    What is this?
 b. 那麼用功的學生，會考不上學校？ dubitative
    Can such a diligent student fail the school entrance exams?
 c. 難道你還不了解我？ rhetorical
    Don’t you understand me?

However, these vague statements touch only superficial mood issues. For NLP purpose, we need more information on their linguistic structures.