Comparative Analysis of Term Distributions in a Sentence and in a Document for Sentence Retrieval

Kyoung-Soo Han and Hae-Chang Rim

Dept. of Computer Science and Engineering, Korea University
1, 5-ga, Anam-dong, Seoul 136-701, Korea
{kshan, rim}@nlp.korea.ac.kr
WWW home page: http://nlp.korea.ac.kr/~kshan/publications/cicling2004/

Abstract. Most of previous works of finding relevant sentences applied document retrieval models to sentence retrieval. However, the performance was very poor. This paper analyzes the reason of this poor performance by comparing term statistics in a document with those in a sentence. The analysis shows that the distribution of within-document and within-sentence term frequency is not similar, and the distribution of document frequency is similar to that of sentence frequency. Considering the discrepancy between the term statistics, it is not appropriate that document retrieval models, as they stand, are applied to sentence retrieval.

1 Introduction

It is necessary to find relevant information for many text processing systems including information retrieval, text summarization, and question answering. Recently, TREC 2002 novelty track defined the related task consisting of two phases[1]. The first phase was to first filter out all non-relevant information, defined in the track to be sentences, from a ranked list of documents. In the second phase, a system threw away any redundant information, defined also to be sentences, from the list of relevant sentences. The relevant information not containing any redundancy was called novel information. In this paper, we focus on the first phase, relevant sentence retrieval.

Most of systems participating in the TREC novelty track applied traditional document retrieval models to relevant sentence retrieval. However, their performance of finding relevant sentences was very poor, which were negative effect on the second phase.

To improve the sentence retrieval performance, Allan tried various techniques with traditional document retrieval models such as vector space model and language model[2]. While the pseudo-relevance feedback was useful to improve performance, almost all his trial was unsuccessful.

In this paper, we analyze the characteristics of sentence retrieval and examine the reason why the traditional document retrieval models suffered from the poor performance in sentence retrieval.