Chronological Ordering Based on Context Overlap Detection

Mohamed H. Haggag, Department of Computer Science, Faculty of Computers & Information, Helwan University, Cairo, Egypt

Bassma M. Othman, Department of Computer Science, Faculty of Computers & Information, Helwan University, Cairo, Egypt

ABSTRACT

Context processing plays an important role in different Natural Language Processing applications. Sentence ordering is one of critical tasks in text generation. Following the same order of sentences in the row sources of text is not necessarily to be applied for the resulted text. Accordingly, a need for chronological sentence ordering is of high importance in this regard. Some researches followed linguistic syntactic analysis and others used statistical approaches. This paper proposes a new model for sentence ordering based on semantic analysis. Word level semantics forms a seed to sentence level semantic relations. The model introduces a clustering technique based on sentences senses relatedness. Following to this, sentences are chronologically ordered through two main steps; overlap detection and chronological cause-effect rules. Overlap detection drills down into each cluster to step through its sentences in chronological sequence. Cause-effect rules forms the linguistic knowledge controlling sentences relations. Evaluation of the proposed algorithm showed the capability of the proposed model to process size free texts, non-domain specific and open to extend the cause-effect rules for specific ordering needs.

Keywords: Clustering, Natural Language Processing, Overlap Detection, Relatedness Matrix, Sentence Ordering, Word Sense

1. INTRODUCTION

Text Summarization (TS) is considered to be one of the most widely used sub-fields of the Natural Language Processing (NLP) field. A summary is defined as “a text that is produced from one or more texts, that convey important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that” (Dipanjan & Martins, 2007). The researchers of text summarization always focus on the ways of extracting the most suitable sentences to be included in the summary. But they give a little focus on sentence ordering which is rarely considered. In single document summarization, the summary is created from only one document, so the order of sentences would be arranged according to the order of sentences in the original document. But in multi-document summarization, which is
an automatic procedure aimed at extraction of information from multiple texts written about the same topic. As sentences are extracted from different documents, so a strategy to arrange the order of sentences so they would form a readable and coherent summary needs to be found. Multi-document summarization poses a number of new challenges over single document summarization. Researchers have already investigated issues such as identifying repetitions or contradictions across input documents and determining which information is salient enough to include in the summary (Lloret, Ferrández, Muñoz, & Palomar, 2008). One issue that has received little attention is how to organize the selected information so that the output summary is coherent. Once all the relevant pieces of information have been selected across the input documents, the summarizer has to decide in which order to present them so that the whole text makes sense. In single document summarization, one possible ordering of the extracted information is provided by the input document itself. However, Dipanjan and Martins (2007) observed that, in single document summaries written by professional summarizers, extracted sentences do not retain their precedence orders in the summary. Moreover, in the case of multiple input documents, this does not provide a useful solution: information may be drawn from different documents and therefore, no one document can provide an ordering. Furthermore, the order between two pieces of information can change significantly from one document to another.

In this approach, constraints on ordering in the context of multi-document summarization are investigated. Believing that relation between sentences is the most important factor in arranging them in a coherent and readable summary, and believing that sentences must be clustered in relative groups. First step is to construct a sentence relatedness matrix that scores a factor indicates the relation between sentences. Then a simple clustering algorithm is performed to arrange sentences according to their score of relevance stored in the matrix to form an initial order. After the most related sentences are grouped into clusters, overlap detection between sentences is the next step to optimize the ordering mechanism. Finally sentences are ordered chronologically based on the cause-effect relation between sentences according to a set of defined chronological rules.

2. RELATED WORKS

The goal is to determine a most probable replacement of sentences or, in other words, reconstruct speak structure of sentences gathered from multiple sources. When a human is asked to make an arrangement of sentences, he or she may perform this task without difficulty just as writing out thoughts in a text. However, they considered what accomplishes this task since computers are unaware of order of things by nature. There are many searches in ordering sentences techniques. Tiedan Zhu et al. (2012) presented an improved approach to sentence ordering in multi document summarization. They rely on using sentences logical closeness by defining the notation $a \rightarrow b$ to represent that $a$ and $b$ are adjacent. The ‘adjacent’ here means: $a$ and $b$ are coherent enough to be connected together in a document (summary); second, $a$ precedes $b$. Then they defined the sentence-chain which is a chain of adjacent sentences, i.e. $A = (a_1 \rightarrow a_2 \rightarrow \ldots \rightarrow a_{n-1} \rightarrow a_n)$ is a sentence-chain with the length of $n$, where $a_i$ is a sentence. Then $A \rightarrow B = (a_1 \rightarrow a_2 \rightarrow \ldots \rightarrow a_{n-1} \rightarrow a_n \rightarrow b_1 \rightarrow b_2 \rightarrow \ldots \rightarrow b_{m-1} \rightarrow b_m)$. The result of two sentence-chains connected with an arrow is still a sentence-chain.

After defining the logical closeness the measured it according to position in original input: assume there are two documents, Da and Db. They extracted sentence $a$ from document Da and sentence $b$ from document Db to constitute the summary. In document Da, sentence $a_1$ is the one just in front of $a$, and sentence $a_2$ is the one just behind $a$. In document Db, sentence $b_1$ is the one just in front of $b$, and sentence $b_2$ is the one just behind $b$. So in Document Da, $a_1 \rightarrow a_2$. In document Db, $b_1 \rightarrow b_2$.

Sentence $a_1$ and $a_2$ are much coherent with sentence $a$, since they are adjacent in Document
Related Content

Interactive IR in OPAC Environments
Iris Xie (2008). *Interactive Information Retrieval in Digital Environments* (pp. 29-52).
[www.igi-global.com/chapter/interactive-opac-environments/24523?camid=4v1a](www.igi-global.com/chapter/interactive-opac-environments/24523?camid=4v1a)

Improved Parameterless K-Means: Auto-Generation Centroids and Distance Data Point Clusters
[www.igi-global.com/chapter/improved-parameterless-means/75906?camid=4v1a](www.igi-global.com/chapter/improved-parameterless-means/75906?camid=4v1a)

Networking E-Learning Hosts Using Mobile Agents
[www.igi-global.com/chapter/networking-learning-hosts-using-mobile/24169?camid=4v1a](www.igi-global.com/chapter/networking-learning-hosts-using-mobile/24169?camid=4v1a)
Personalization Approaches for Ranking: A Review and Research Experiments
www.igi-global.com/article/personalization-approaches-for-ranking/165376?camid=4v1a