Available online at www.sciencedirect.com

ScienceDirect

IERI Procedia 10 (2014) 190 – 195

Procedia IERI

www.elsevier.com/locate/procedia

2014 International Conference on Future Information Engineering

Improve Bayesian Network to Generating Vietnamese Sentence Reduction

Ha Nguyen Thi Thu[a*], Dung Vu Thi Ngoc[b]

[a]*Information Techonology Faculty, Vietnam Electric Power University, Hanoi, Vietnam*
[b]*HaiDuong Center for Continuing Education, HaiDuong, Vietnam*

Abstract

Sentence reduction is one of approaches for text summarization that has been attracted many researchers and scholars of natural language processing field. In this paper, we present a method that generates sentence reduction and applying in Vietnamese text summarization using Bayesian Network model. Bayesian network model is used to find the best likelihood short sentence through compare difference of probability. Experimental results with 980 sentences, show that our method really effectively in generating sentence reduction that understandable, readable and exactly grammar.

Keywords: Sentence reduction, natural language processing, text summarization, Bayesian network, probability;

1. Introduction

Today, most text summarization systems based on extracted sentences to generate a summary and we called extracted approach [9], [12], [13], [14]. With this approach, the weight of sentence is calculated based on some features that we think it is important like: term frequency, sentence position, sentence length... And

* * Corresponding author. Tel.: +84906113373
E-mail address: hantt@epu.edu.vn.

doi:10.1016/j.ieri.2014.09.076

then, sentences will be sorted by its weight and extracted based on the rate (extraction rate). A text summary is including sentences that has maximum weight from the original text. With this approach, text summary will synthesis of the discrete sentences from original text, it can be:

- Text summarization is seamlessly because of sentences are not linked by content in the text, specially, when the extraction rate is smaller, it will be greater discrete.
- Text summarization is confusing sometimes it can be loosen important information in original text by some sentences that have been not extracted.

Therefore, we have chosen the sentence reduction approach for processing in sentence level, remove not important words in the sentence and generate new sentence and creating summary. Target text will overcome some disadvantages that have been analyzed above [5], [6], [7].

This paper present a Vietnamese sentence reduction method based on Bayesian network, each word in original text is considered as a node of the network. Reduced sentence is generated by find a path that is the shortest and greatest weight, we called it: the best likelihood path.

The next structure of this paper: In Section 2, is overview of related work, the methodology of sentence reduction based on Bayesian network method in section 3, Section 4 is the experimental results and finally is conclusion.

2. Related works

Almost related works focus in building lexical rules model or syntax parser tree. First Aho and Ullman using synchronous context free grammars (SCFGs)[21]. Wu in 1997 has proposed a method that included inversion transduction grammar [22] and some other relate research with CFG like head transducers by Alshawi, Bangalore and Douglas in 2000.

Knight and Marcu proposed a noisy channel of sentence compression. They use two components: P(y) is language model and P(x|y) is a channel model. P(x|y) capturing the probability of original sentence x and target compression y. After that, using decoding algorithm searches for maximizes of P(x)P(x|y). This channel model is a SCFG, parallel corpus is used to extracted rules, and weights estimated using maximum likelihood [9].

With Vietnamese sentence reduction approach, most of the methods are applied from English method. However, the performance of this method is not high when applied to Vietnamese language. Because of single syllable language, In Vietnamese, words can not be determined based on space..... So that, they often use extraction approach for building Vietnamese text summarization systems, there are some method that use reduction approach but it is not high effectively.

3. Methodology of sentence reduction based on bayesian network

3.1. Bayesian Network

Bayesian networks is the one of probabilistic graphical models. When represent the uncertain knowledge can used graphical models. In Bayesian model, each node is a random variable, and the edges between the nodes represent prbalilistic dependencies among the corresponding random variables. This probabilistic can be computed from history data [2].

If B is a Bayesian network B, so that B is an annotated acyclic graph and B represents a joint probability distribution over a set of random variables V. This network is defined:

B=<G, Θ>

In which: