A New Approach Using Hidden Markov Model and Bayesian Method for Estimate of Word Types in Text Mining

Adem Doganer, Biostatistics and Medical Informatics, Kahramanmaras Sutcu Imam University, Kahramanmaras, Turkey
Sinan Calik, Fırat University, Elazig, Turkey

ABSTRACT

Determining the structure of words in the text for the operations such as automated information extraction and text summarization of the text is essential. In computers, textual analysis to define the type of the word is considered as a vital advantage. Defining the types of words provides an estimate of the sequence of words in the sentence. In this article, estimating types of Turkish words is provided by developing a Hidden Markov Model and a Bayesian-based new model. In this model, an algorithm is developed which separates the suffixes of the words and grouping the words by counts of characters that suffixes of the words receive. A text composed of 584 Turkish words is used for the testing the dependability of the model. The model has achieved a high success rate in predicting the types of Turkish words.

KEYWORDS
Bayesian Methods, HMM, Modeling, Text Mining, Word Type

INTRODUCTION

Improvements in information technologies alongside the development of Internet technology, the web has led to the formation of great size of information in the stack. The information contained in the web environment is usually present as text. Individually examination of the text in the web environment by humans is not possible in terms of time. Therefore, intensive studies have been conducted on the development of technology that can automatically evaluate these texts.

In order to evaluate the texts in the web environment, there is a requirement for intelligent systems. Understanding the texts of web environment by evaluating is not possible for the computers. Hence in attempt to evaluate the texts by the computers, new model developments that could estimate based upon the words contained in the text is a necessity. One of the important factors that determines the structure and the contents of the text is the words. It provides an opportunity in terms of evaluating structure and lineage in sentence of words in the text. Estimating the types of words contained in the text with a model will ensure an important progress in the analysis of text content.

Different algorithms and models have been used in many studies to improve information extraction from text and text mining methods. He and Ling (2006) have used ontology-based metadata model by implementing the entity-relationship models and the classification of metadata information. Data mining methods are frequently used in the analysis of the text. Alhajj (2003) worked on an algorithm to perform mining on text documents by using the entity-relationship model. Direct access to regarding information from text documents, text filtering, text summary and data extraction can

DOI: 10.4018/IJKSS.2017100102

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
be achieved. In general, these operations expressed as text mining. Text mining methods are taken into consideration by many researchers. Tsai and Chang (2013) have used support vector machine algorithms for the model that takes into account the examples of selections. A large number of models can be administered to text classification. Ghiassi et al. (2012) worked on a dynamic neural network model for automatic text classification. Chen and Chen (2011) have benefited from the similarity measurements and chi-square statistics for text classification. Jiang et al. (2010) have worked on the text classification by using a feature extraction-based graph mining method. Klose et al. (2000) worked on document similarities on the text access subject.

Information extraction is one of the major areas of text mining. Different models have been developed on information extraction. Ko and Seo (2008) have obtained successful results in the operations such as text summarization and sentence extraction by benefitting from statistical techniques. Downey et al. (2010) worked on the uncontrolled data extraction from the web environment with the probabilistic model analysis. Data extraction models can be used on the texts in Turkish Language. Many researchers worked on algorithms and models that can extract information from texts in Turkish by using these methods (Tur et al., 2003; Tatar & Cicekli., 2011; Tatar, 2011; Adalı, 2009).

Stochastic model is one of the most commonly used methods in text mining. Stochastic models are widely used to provide information in various studies (Vlad et al., 2003; Hromic & Atkinson, 2012). Different statistical methods were used to investigate the morphological aspects of the words. Aeljmi et al. (2011) studied morphological aspects of Arabic words by HMM. Papageorgiou (1994) offers different model from the current approaches for the segmentation of words in the Japanese Language by using HMM.

Since Turkish has an agglutinating language structure feature, words can take roots as well as suffixes. Meaning of the Turkish words can be changed by added suffixes. On the other hand, some suffixes do not cause any changes in the meaning of the word.

The study is intended to estimate the count of characters of the Turkish words according to the type of suffixes received by the word. Model was used to 584 word-text. An algorithm that separates words from suffixes and classifies them by counts of characters received is developed to estimate the types of words. Hidden Markov Model, Baum-Welch algorithm and Bayesian methods are used for estimating the structure of words upon the counts of characters received by suffixes of the words.

The word types in a text must be estimated in order to automatically information extraction from a text in Turkish. Turkish has a very complex structure compared to other language groups. While a word belongs to a certain word type in most languages, word types in Turkish may vary depending on their suffixes and the count of characters in these suffixes. In addition, word types may differ when the count of additional characters in a word increases. Because it is quite complex to classify word types in a Turkish text, numerous machine learning and statistical methods fail to automatically information extraction in a Turkish text. Although Artificial Neural Network (ANN) model was applied to the classification of word types in a Turkish text, it did not yield the expected outcome. ANN model aimed to classify word types based on their syntactical order. When applied to a number of examples, however, it tries to overfitting the syntactical order in a sentence. Support Vector Machines (SVM) reaches a high success rate in various classification tasks. Nevertheless, because word types are probabilistically distributed in a Turkish text, SVM model still fails to classify them. Decision Trees (DT) model also succeeds in classification when a text has a relatively low number of text types. However, because a word may belong to different word types in the same Turkish text, it is unlikely to accurately classify word types. Cluster analysis and discriminant analysis yields limited results the number of variables defining word types is low. There are two process sequences in HMM model as observation sequence and state sequence. Observation sequence is used to estimate transition probabilities among states. Thus, HMM model seems suitable for the estimation of word types in a Turkish text. Word types in a Turkish text represent states while the additional count of characters represent observations. It is aimed to estimate word types through the count of additional characters in a word. A word may sometimes have different additional characters in a Turkish text,
Possibility of Fractal Aesthetics
www.igi-global.com/chapter/possibility-fractal-aesthetics/69465?camid=4v1a

The Crisis and the Complexity of Knowing
www.igi-global.com/article/crisis-complexity-knowing/69959?camid=4v1a