Information Sciences 181 (2011) 1538–1551

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A trigram hidden Markov model for metadata extraction from heterogeneous references

Bolanle Ojokoh [a,b], Ming Zhang [a,*], Jian Tang [a]

[a] School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, PR China
[b] Department of Computer Science, Federal University of Technology, P.M.B. 704 Akure, Nigeria

ARTICLE INFO

Article history:
Received 24 March 2010
Received in revised form 25 December 2010
Accepted 2 January 2011
Available online 11 January 2011

Keywords:
Metadata extraction
Hidden Markov models
Bibliography
Second order
Shrinkage

ABSTRACT

Our objective was to explore an efficient and accurate extraction of metadata such as author, title and institution from heterogeneous references, using hidden Markov models (HMMs). The major contributions of the research were the (i) development of a trigram, full second order hidden Markov model with more priority to words emitted in transitions to the same state, with a corresponding new Viterbi algorithm (ii) introduction of a new smoothing technique for transition probabilities and (iii) proposal of a modification of back-off shrinkage technique for emission probabilities. The effect of the size of data set on the training procedure was also measured. Comparisons were made with other related works and the model was evaluated with three different data sets. The results showed overall accuracy, precision, recall and F1 measure of over 95% suggesting that the method outperforms other related methods in the task of metadata extraction from references.

1. Introduction

The dramatic growth of digital libraries in recent years has not only simplified access to existing information sources, but has also made the task of finding, extracting and aggregating relevant information difficult. In the bibliographic research community, several researches are being conducted on citation analysis, grouping and social networks creation for subsequent mining. A prerequisite to such tasks is accurate reference metadata extraction process.

References are most commonly found in the late section of an article; this section is often labeled "References", "Bibliography" or "List of References", and information that is normally contained in this section includes the author names, title, journal, volume, number (issue), year, and page information. These have constituted an important kind of metadata valuable for literature search, analysis, and evaluation [9].

Automatic reference extraction is particularly difficult because of the problems of inconsistent formatting, semantically overloaded punctuations and field separators, and existence of many dramatically different reference styles.

Inspired by the work of Yin et al. [30], where the inner emission probability was computed according to the bigram sequence relationship of words within the same field, we describe a method that utilizes trigram HMMs with more priority to the words emitted in transitions to the same state for the task of metadata extraction from references. We propose a three dimensional transition matrix in which the probability of transitioning to a new state depends not only on the current state according to the traditional HMM but also on the previous state. Our method improves on those adopted by previous researches, by recommending a new approach for smoothing transition probabilities, a modified shrinkage technique for smoothing emission probabilities and optimization of the emission vocabulary.

* Corresponding author. Tel.: +86 13601036730; fax: +86 10 62765822.
E-mail addresses: bolanleojokoh@yahoo.com (B. Ojokoh), mzhang@net.pku.edu.cn (M. Zhang), tangjian_0@126.com (J. Tang).

0020-0255/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2011.01.014

The rest of this paper is organized as follows: Section 2 discusses related work. In Section 3, an overview of the trigram hidden Markov model is presented. Section 4 describes the approach adopted here in solving the problem of metadata extraction and the system architecture while emphasizing the contributions of this paper. In Section 5, the experimental results are presented. Finally, Section 6 concludes the paper.

2. Related work

Several researches have been conducted on extraction of metadata from references. The methods adopted generally fall under two broad categories: rule and knowledge-based methods [6,10,18] and machine learning methods [1,2,4,7,8,28,29]. Ojokoh [18] used rule-based approach for the task from a number of reference styles. Day et al. [6] utilized a knowledge based approach for solving the problem. Recently, Gupta et al. [10] used a combination of regular expression based heuristics and knowledge based systems to extract metadata from the references of some biological science papers. Cortez et al. [4] and Cortez and da Silva [5] employed unsupervised machine learning approach for metadata extraction with relatively good results. Some related studies involving scholarly digital libraries have been conducted by many researchers. Automated reference extraction systems will be relevant for these studies. For instance, Shi et al. [26] worked on anchor text extraction; Papavlasopoulos [20] worked on evaluating the scientific impact of journal; and Kerne et al. [13] developed applications that could be useful to represent the output of automated metadata extraction applied to specific context like in our research. Several researchers have applied the hidden Markov model for different information extraction tasks with good performance when applied to the tasks in both structured, semi structured and free text [17]. For instance, HMMs have been used for extraction of gene names and locations from scientific abstracts [14]; named-entity recognition [1]; and extraction of addresses [2,28]. Seymore et al. [24] explored the use of HMM for learning model structure from data, but centered their research on the task of extracting information from the headers of computer science research papers. The reference metadata extraction problem has been solved using HMMs by some researchers. Connan and Omlin [3] developed a variety of models for different reference styles while Yin et al. [30] explored a bigram HMM which used word sequential relation and position information in text fields in addition to only word frequency used by traditional HMMs. Geng and Yang [8] proposed the use of special states for delimiters and HTML tags while extracting metadata from Web references. More recently, Hetzner [12] used HMMs for metadata extraction from the same data set used in this work. He assigned two states to each label and one state to each delimiter implementing the traditional HMM. He further recommended a better method for model selection; mapping optimization and the employment of a second-order HMM considered in our work.

Our research was therefore partly motivated by the need to solve the problems proposed in [12], as it creates a more efficient HMM for metadata extraction. It is the first to use a full second-order HMM with more priority to the words emitted in transitions to the same state for the task of metadata extraction from references. Peng and McCallum [21] applied HMM to metadata extraction with only second-order transitions. The inclusion of word sequential relation and position information in text fields has been adopted with success [1,30]. While those works focused on these at the first order level, this work pays more attention to the existence of such within the same state and at a higher order level.

Some researchers outside the metadata extraction domain have proved that higher-order HMMs can achieve higher performance than first-order HMMs. This was shown in the contexts of speech recognition [15] and part-of-speech tagging [29]. More recently, Shahin [25] confirmed that using second-order HMMs enhanced speaker authentication performance compared to first-order models. Generally, higher-order HMMs have been proved to incorporate more information into the training procedures of practical problems solved by them. Embedding more information into the learning process, especially if there are many training data, leads to accuracy of estimations.

3. The trigram hidden markov model

A hidden Markov model (HMM) is a finite state automaton comprising of stochastic state transitions and symbol emissions. The automaton models a probabilistic generative process whereby a sequence of symbols is produced by starting at a designated start state, transitioning to a new state, emitting one of a set of output symbols selected by that state, transitioning again, emitting another symbol, and so on, until a designated final state is reached. Associated with each of the set of states, $S = \{s_1,\ldots,s_n\}$, are a probability distribution over the symbols in the emission vocabulary $V = \{w_1,\ldots,w_m\}$, and a probability distribution over its set of outgoing transitions.

When using HMM to perform metadata extraction, the goal is to determine the most likely sequence of states that generates a given sequence of symbols. The Viterbi Algorithm is a common method for calculating this [23]. It would solve this problem in just $O(TN^2)$ time, where T is the length of the sequence and N is the number of states [7]. After obtaining the results, each word is then put in its corresponding field according to the state sequence.

The hidden Markov model proposed in this work is different from the traditional HMM which utilizes the common Viterbi algorithm [23]. It is a new model referred to as trigram. It is a full second-order HMM, but with the placement of more priority to the words emitted in transitions to the same state. The new transition matrix is three dimensional, stating that the probability of transitioning to a new state depends not only on the current state according to the traditional HMM but also on the previous state. This allows more information to be incorporated into the model.