Journal of Biomedical Informatics 58 (2015) S60–S66

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Hidden Markov model using Dirichlet process for de-identification

Tao Chen *, Richard M. Cullen, Marshall Godwin

Primary Healthcare Research Unit, Memorial University of Newfoundland, Canada

ARTICLE INFO

Article history:
Received 16 February 2015
Revised 2 September 2015
Accepted 3 September 2015
Available online 25 September 2015

Keywords:
De-identification
Natural language processing
Hidden Markov model
Dirichlet process
Variational method

ABSTRACT

For the 2014 i2b2/UTHealth de-identification challenge, we introduced a new non-parametric Bayesian hidden Markov model using a Dirichlet process (HMM-DP). The model intends to reduce task-specific feature engineering and to generalize well to new data. In the challenge we developed a variational method to learn the model and an efficient approximation algorithm for prediction. To accommodate out-of-vocabulary words, we designed a number of feature functions to model such words. The results show the model is capable of understanding local context cues to make correct predictions without manual feature engineering and performs as accurately as state-of-the-art conditional random field models in a number of categories. To incorporate long-range and cross-document context cues, we developed a skip-chain conditional random field model to align the results produced by HMM-DP, which further improved the performance.

1. Introduction

De-identification allows record-level data collected for healthcare purposes to be available to researchers for secondary analysis while preserving the privacy of individual patients. Where it exists, privacy legislation usually deems de-identification as mandatory for the release of medical data. These retrospective data are attractive to researchers because they require no participant recruitment and provide a large participant pool compared to smaller sample sizes usually associated with prospective datasets. Because manual processing cannot meet the increasing demand for administrative data, automatic algorithms are receiving much attention. Many studies have adopted statistical natural language processing (NLP) methods and have achieved good results. In particular, conditional random field (CRF) models have demonstrated impressive performance in many tests [1,2]. However, CRF usually requires significant effort on feature engineering. The quality of designed features has a great impact on performance but features designed based on training data may not necessarily apply well to new data. Furthermore, a large number of features and parameters introduced through feature engineering increase model complexity which may also prevent the model from generalizing well to new data.

Hidden Markov models (HMM) are simple generative models that have proven effective in many NLP tasks such as Part-of-Speech (POS) tagging and Named Entity Recognition (NER) [3]. These usually do not require much feature engineering. However, their strong independence assumption limits their performance. Recent studies have shown that the use of latent variables can relax the independence assumption, capture underlying semantic information, and provide meaningful features for NLP tasks [4]. In this challenge, we developed a standard HMM into a non-parametric Bayesian model with latent variables, named as HMM-DP, which requires minimum feature engineering for out-of-vocabulary words. In the model, latent variables categorize words into refined categories, which makes the model more expressive and enables the model to capture the variations in the data. Instead of using a pre-fixed number of latent variables, we assume there can be an infinite number of latent variables and let the data determine the optimal number by application of a Dirichlet process prior. The experiment on the 2014 i2b2/UTHealth data demonstrates that the model is effective in understanding local context cues and can be a close competitor to the state-of-the-art CRF models.

Though HMM-DP works well with local context cue modeling, a close examination of the data reveals that long range and cross document context cue modeling are also helpful in improving performance. To take advantage of this observation, we develop a skip-chain CRF on the data produced by HMM-DP. Results of testing show that the system performance, especially recall, can be improved by combining the two models into a pipeline.

* Corresponding author.
E-mail addresses: tao.chen@med.mun.ca (T. Chen), richard.cullen@med.mun.ca (R.M. Cullen), godwinm@mun.ca (M. Godwin).

http://dx.doi.org/10.1016/j.jbi.2015.09.004
1532-0464/

2. Background

De-identification can be modeled as a sequence tagging problem where each word in a document is assigned a tag of 'identifier' or 'non-identifier'. The identifier tag can be further categorized as, for example, NAME, PROFESSION, and LOCATION, as in this challenge. Well studied sequence tagging problems include POS tagging and NER, and in these studies HMM and CRF both have been widely used. However, HMM usually does not offer the same performance as CRF because of the independence assumption. CRF is more flexible because it models the joint probability of tags and words.

There are many proposed improvements over HMM in NLP. One improvement is developed upon a common phenomenon in natural language whereby the words prior to and following a word have significant influence on the meaning of that word. Authors of a previous study introduced bi-direction emission, i.e., the emission probability depends on not only the current and previous tags but also the following tags [5]. In another study, they demonstrated an improvement through relaxing the independence assumption by introducing latent variables for each tag and using latent variables to capture the dependence of the context [6]. These authors have shown that bi-gram HMM with latent variables has been able to outperform tri-gram HMM with bi-directional emission in a POS tagging task. Our model follows this stream of study and employs latent variables to relax the independence assumption and to capture context cues.

In the de-identification task, a context cue plays an important role because training data usually do not contain all identifier words and some words can be identifiers or not depending on the context. In CRF modeling, context cues are captured by examining a word window around the potential identifier word and modeling a joint probability among words and tags within the window. In HMM, context cues are modeled through transition probability between tags, and as a result the meaningfulness and detailedness of tags determine how well HMM can understand context cues. Ideal training data would not only provide tags for identifiers but also label cue words, for example, "Dr." for the name identifier and "live" for the location identifier, but such data is unlikely to be available. Manually deriving all cue words would not be practical. A common automatic solution is to assign new tags to words that are specially positioned around identifiers, for example, assigning a non-identifier word with a special tag if it is immediately prior to or following an identifier. The effectiveness of this approach has been demonstrated previously [7] but it does not perform well if context cues are more than one position away or contain more than one word. Our model utilizes latent variables to classify words into more detailed sub-tags and the classification process considers surrounding words and sub-tags to allow it to capture more complicated context cues.

HMM with latent variables constructs a mixture model and choosing the right number of latent variables is a difficult task. The likelihood of a general mixture model increases with the number of latent variables. However, a model with a large number of latent variables may appear to fit well with training data, although it is likely to be over-fitting and perform poorly with new data. An engineering approach is to use validation data to determine the optimum number of latent variables: test the model on validation data with different numbers of latent variables and pick the number with the best result. This problem can also be addressed by adding regularities, for example, adding a Bayesian prior to latent variables, like LDA [4]. The likelihood of such a Bayesian model is no longer a monotonic function of the number of latent variables. In practice, the optimal number is determined by running the estimation multiple times with different numbers of latent variables and choosing the number producing the maximum likelihood [8]. Non-parametric modeling introduces a different approach to the problem and we use a Dirichlet process as a prior for the number of latent variables. In this non-parametric model, we assume there can be an infinite number of sub-tags in the data and let the characteristics of the data determine the best number.

3. Model

In this section, we review the Dirichlet process, introduce the HMM-DP model, and then discuss how to learn the model and apply the model for prediction.

3.1. Dirichlet process and HMM-DP model

In the model, a Dirichlet process (DP) is used as a Bayesian prior for latent variables represented with a stick-breaking process. The stick-breaking representation of DP contains two components. The first component is the stick-breaking process, which is achieved by first generating a countable infinite collection of stick-breaking portions $v_1, v_2, \ldots$, where each $v_i \sim Beta(1, \alpha)$ and then letting

$$\pi_i = v_i \prod_{j<i} (1 - v_j)$$

It can easily show $\sum \pi_i = 1$ and the stick-breaking process is denoted as $GEM(\alpha)$. The second component is a countable infinite collection of atoms, $\eta_1, \eta_2, \ldots$, matching with the stick-breaking portions, where each η_i is generated from a probability distribution β. Given these two components, DP defines a distribution:

$$P(B) = \sum \pi_i \delta_{\eta_i}$$

where δ_{η_i} equals 1 if $\eta_i \in B$, else 0. DP is denoted as $DP(\alpha, \beta)$. Notice a sample π from stick-breaking process $GEM(\alpha)$ can define a parameter of an infinite multinomial distribution since it sums to 1. In the model a tag has an infinite number of sub-tags that form an infinite multinomial distribution. We use the stick-breaking process as a prior for the distribution of the sub-tags. A sample η from $DP(\alpha, \beta)$ also defines a parameter of an infinite multinomial distribution given β is a stick-breaking process. The transition probability from sub-tags to sub-tags forms an infinite multinomial distribution and the transition probability should be related to the distribution of the destination sub-tag, that is, it is more likely to transit to a sub-tag that is important in the tag. Here we use DP as a prior for the transition probability and β is the distribution of destination sub-tag. This construction allows the transition probability to consider the relative importance of the sub-tags.

The generative process of HMM-DP is defined as such: each tag contains an infinite number of sub-tags; the sub-tag of a tag determines the next tag; the sub-tag of a tag and the next tag determine the sub-tag of the next tag; the sub-tag of a tag determines the word at the position. The complete distribution is as follows,

$$\begin{aligned} p(w, s, \pi, \phi^E, \phi^T, \phi^t, z) = & \prod_s p(\pi_s|\alpha) \times \prod_z p(\phi^E_{sz}|\alpha^E) p(\phi^T_{sz}|\alpha^T) \\ & \times \prod_{s,z,s'} p(\phi^t_{s,z,s'}|\alpha^t, \pi_s) \times \prod_d p(s_0) p(w_0) p(z_0) \\ & \times \prod_n p(s_n|\phi^T_{s_{n-1}, z_{n-1}}) \times p(z_n|\phi^t_{s_{n-1}, z_{n-1}, s_n}) \\ & \times p(w_n|\phi^E_{s_n, z_n}) \end{aligned}$$

where s denotes tag, z denotes sub-tag, d denotes document, w denotes word, π is the weight for sub-tags, ϕ^E is the emission probability of a sub-tag, ϕ^T is the transition probability from sub-tag to tag, and ϕ^t is the transition probability from sub-tag to sub-tag. The