برچسب‌گذاری ادات سخن فارسی با استفاده از مدل شبکه‌های فازی

محمد بادپما، فاطمه خورعلی

چکیده

برچسب‌گذاری ادات سخن یکی از مسائل مطرح در حوزه پردازش زبان‌های طبیعی است. هدف در این مطالعه تعیین نقش واژگان در جمله است. برحسب این برچسب‌گذاری ویژگی‌های دستوری و نحوی از واژگان تعریف می‌شود. در این مقاله یک روش بیشتر برآورده توسط ادات سخن فارسی پیشنهاد شده است. در این روش محدودیت‌های روش‌های آمیزه‌ای از استفاده از مجموع کلیدی که توسط واژگان در جمله کرده است باعث شده که تعداد کمی داده آموزشی، مدل فازی پارامترهای قابل اطمینان تری را تخمین می‌زنند. در این روش ابتدا اندازه‌سازی به‌عنوان یک راه حل استفاده می‌شود و سپس فاصله جایگاههای همبستگی و تغییرات همبستگی را با تابع تابعی ثابت نمی‌پذیرد. زده و سپس شکل شبکه فازی تشکیل شده و درجه هر پال در این شبکه با استفاده از یک شبکه عصبی و تابع عضوی مشخص می‌شود. در نهایت بعد از اینکه جمله شبکه فازی برای یک جمله ساخته شده از الگوریتم ویرچر برای تعیین محتوای تنیسی در این شبکه استفاده شده است، نتایج آزمایش روی یکی از کاراکتر آین روش چرخه کرده و نشان می‌دهد که روش پیشنهادی در شرایطی که داده‌های آموزشی کیکی در اختیار باشند، از روی‌های مشابه، مثل مدل‌های مارکوف، مدل‌های بی‌پرندگانی داده می‌گیرد.

واژگان کلیدی: پردازش زبان‌های طبیعی، برچسب‌گذاری ادات سخن، فازی، شبکه عصبی.

Part Of Speech Tagging of Persian Language using Fuzzy Network Model

Mohammad Badpeima, Fatemeh Hourali, Mayam Hourali

1,3 Electrical and Computer Engineering Department, MUT, Tehran, Iran
2 Electrical and Computer Engineering Department, Esfara An University Of Technology, North Khorasan, Iran

Abstract

Part of speech tagging (POS tagging) is an ongoing research in natural language processing (NLP) applications. The process of classifying words into their parts of speech and labeling them accordingly is known as part-of-speech tagging, POS-tagging, or simply tagging. Parts of speech are also known as word classes or lexical categories. The purpose of POS tagging is determining the grammatical category of the words in a sentence. Grammatical and syntactical features of words are determined based on these tags.

The function of existing tagging methods depends on the corpus. As if the educational and test data are extracted from a corpus, the methods are well-functioning, or if the number of educational data is low, especially in probabilistic methods, the accuracy level also decreases. The words used in sentences are often vague. For example, the word ‘Mahrani’ can be a noun or an adjective. Existing ambiguity can be eliminated by using neighbor words and an appropriate tagging method.

Methods in this domain are divided into several categories such as: based on memory [2], rule based methods [5], statistical [6], and neural network [7]. The precision of more of these methods is an average of

*Corresponding author

الگوریتم شبکه عصبی

1397 شماره 4 پیام 38
95% [1]. In the paper [13], using the TnT probabilistic tagging and smoothing and variations on the estimation of the three-words likelihood function, a tagging model has been created that has reached 96.7% in total on the Penn Treebank and NEGRA entities. [14] Using the representation of the dependency network and extensive use of lexical features, such as the conditional continuity of the sequence of words, as well as the effective use of the foreground in the linear models of linear logarithms and fine-grained modeling of the unknown words, on the Penn Treebank WSJ model, 97.24% accuracy is achieved.

The first work in Farsi that has used the word neighborhoods and the similarity distribution between them. The accuracy of the system is 57.5%. In [19], a Persian open source tagger called HunPoS was proposed. This tag uses the same TnT method based on the Hidden Markov model and a triple sequence of words, and 96.9% has reached on the “Bi Jen Khan” corpus.

In this paper a statistical based method is proposed for Persian POS tagging. The limitations of statistical methods are reduced by introducing a fuzzy network model, such that the model is able to estimate more reliable parameters with a small set of training data. In this method, normalization is done as a preprocessing step and then the frequency of each word is estimated as a fuzzy function with respect to the corresponding tag. Then the fuzzy network model is formed and the weight of each edge is determined by means of a neural network and a membership function. Eventually, after the construction of a fuzzy network model for a sentence, the Viterbi algorithm as a subset of Hidden Markov Model (HMM) algorithms is used to specify the most probable path in the network.

The goal of this paper is to solve a challenge of probabilistic methods when the data is low and estimation made by these models is mistaken.

The results of testing this method on “Bi Jen Khan” corpus verified that the proposed method has better performance than similar methods, like hidden Markov model, when fewer training examples are available. In this experiment, several times the data is divided into two groups of training and test with different sizes ascending. On the other hand, in the initial experiments, we reduced the train data size and, in subsequent experiments, increased its size and compared with the HMM algorithm.

As shown in figure 4, the train and test set are directly related to each other, as the error rate decreases with increasing the training set and vice versa. In tests, three criteria involving precision, recall and F1 have been used. In Table 4, the implementation of HMM models and a fuzzy network is compared with each other and the results are shown.

Keywords: Natural language processing, Part of speech (POS) tagging, Persian language, Fuzzy, Neural network.

1- مقدمه
برچسب‌زنی اجزای سخن یکی از مهم‌ترین مسائل مطرح در پردازش زبان‌های طبیعی است. عملکرد روش‌های برچسب‌زنی موجب وابستگی زیادی به یکیکی مورد استفاده دارد. به‌طوری که اگر داده‌های آموزشی و آموزن از یک یکه اخراج شده باشد، روش‌ها عملکرد خوبی دارند یا اگر تعداد داده آموزشی کم باشد به‌خاطر اغراق الگوی احتمالی میزان دقت نیز کاهش می‌یابد [1]. برچسب‌زنی و از اگر کار دشواری است، به‌دلیل اینکه اگر داده‌های آموزن از یک یکه دیگری انتخاب شده باشد، عملکرد مانند به‌طور معمول متخلف نخواهد بود. به‌دلیل اینکه اگر داده‌های آموزن از یک یکه دیگری انتخاب شده باشد، عملکرد مانند به‌طور معمول متخلف نخواهد بود. به‌دلیل اینکه اگر داده‌های آموزن از یک یکه دیگری انتخاب شده باشد، عملکرد مانند به‌طور معمول متخلف نخواهد بود. به‌دلیل اینکه اگر داده‌های آموزن از یک یکه دیگری انتخاب شده باشد، عملکرد مانند به‌طور معمول متخلف نخواهد بود. به‌دلیل اینکه اگر داده‌های آموزن از یک یکه دیگری انتخاب شده باشد، عملکرد مانند به‌طور معمول متخلف نخواهد بود. به‌دلیل اینکه اگر داده‌های آموزن از یک یکه دیگری انتخاب شده باشد، عملکرد مانند به‌طور معمول متخلف نخواهد بود. به‌دلیل اینکه اگر داده‌های آموزن از یک یکه دیگری انتخاب شده باشد، عملکرد مانند به‌طور معمول متخلف نخواهد بود. به‌دلیل اینکه اگر داده‌های آمو...
این صورت که تمام پسوندهای جمع کننده به خود وارد چسبیده می‌شوند، در اساس تمام برجسته‌هایی که از ایجاد یک رابط به سیستم با استفاده از این کلاس اضافه و شبکه عصبی وزنه‌های یک گراف به کاربرد می‌رود. در نهایت، این گراف هر یک به مدل شبکه فازی معروف است به الگوریتم ویتریو [4] داده می‌شود سپس این الگوریتم محتمل‌ترین مسیر را انتخاب می‌کند.

درج کردن نتایج این الگوریتم ویتریو به یک صفحه نمایش‌دهنده و نتایج آموزشی با استفاده از الگوریتم ویتریو واقع می‌شود.

برخی از این الگوریتم‌های مختلف می‌توانند به راحتی این محصولات را تولید کنند.

این روش روى مجموعه‌داده بی‌خان [3] مورد آزمایش قرار گرفته و نتایج نشاندهنده این است که وقتی داده‌های آماری کمی در اختیار داشته باشیم، این روش بی‌هر از مدل مخرب می‌تواند بهترین روش بهترین و ارزشمندی است.

در رابطه با این داده مدل شبکه فازی تعريف می‌شود. در یک بخش سوم، نحوه ساخت شبکه فازی برای بررسی‌های ایجاد مخرب می‌تواند به شرح زیر باشد:

- ۲- شبکه فازی

شبکه‌های فازی در مدل‌های مختلف فناوری و صنعت کاربرد دارد که از جمله آنها می‌توان به پیش‌بینی [8] و سامانه‌های استراتژیک اشاره کرد. تعريف شبکه فازی به‌طور خودکار، به دو کلیه این شکل گذاشته می‌شود.
جدول 1: تعداد برچسب‌ها متغییر

<table>
<thead>
<tr>
<th>تعداد برچسب‌ها متغییر</th>
<th>تعداد وازگان</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92858</td>
</tr>
<tr>
<td>2</td>
<td>3828</td>
</tr>
<tr>
<td>3</td>
<td>293</td>
</tr>
<tr>
<td>4</td>
<td>118</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

به عنوان مثال اگر مجموعه $\mathcal{V}$ به صورت زیر باشد:

$$\mathcal{V} = \left\{ \left( \text{دلما}, 0.1 \right), \left( \text{دبی}, 1.0 \right), \left( \text{سپاه}, 1.0 \right), \left( \text{اشیاء}, 0.0 \right), \left( \text{سیاه}, 0.0 \right), \left( \text{بحث}, 1.0 \right), \left( \text{بحث}, 0.0 \right), \left( \text{بحث}, 0.0 \right), \left( \text{بحث}, 0.0 \right) \right\}$$

و اگر مجموعه بالهای فازی $\mathcal{E}$ به صورت زیر باشد:

$$\mathcal{E} = \left\{ \left( \text{دلما}, 0.1 \right), \left( \text{دبی}, 1.0 \right), \left( \text{سپاه}, 1.0 \right), \left( \text{اشیاء}, 0.0 \right), \left( \text{سیاه}, 0.0 \right), \left( \text{بحث}, 1.0 \right), \left( \text{بحث}, 0.0 \right), \left( \text{بحث}, 0.0 \right), \left( \text{بحث}, 0.0 \right) \right\}$$

شکل 2: مثالی از شکل‌های فازی برای عبارات موردنظر

(Figure-2): An example of a fuzzy network for a given phrase
در این صورت این مجموعه توالی‌هایی به بردار عددهای تبدیل می‌کنم که نتیجه‌های این تبدیل یک بردار ورودی به صورت زیر خواهد بود. در این بردار به نحوه تبدیل برچسب‌های استفاده‌شده در مجموعه‌دمای جهان، به‌صورت مجموعه‌ای جهان است.

\[ X = \{0, 0.1, 0.2, \ldots, 0.9\} \]

می‌توانیم ساختار شبکه فازی برای تبدیل هر یک از مجموعه‌های مورد نظر استفاده‌شده در داده‌باین‌های مورد نظر استفاده می‌شود.

\[ W = \begin{bmatrix} w_1 & w_2 & \cdots & w_N \end{bmatrix} \]

\[ T = \begin{bmatrix} t_1 & t_2 & \cdots & t_M \end{bmatrix} \]

در اینجا به‌عنوان نشان دهنده تعداد وزن‌های و \( M \) و \( N \) به چهار مورد استفاده راهنما به هر گره استفاده می‌شود. برای تخمین این تابع عضویت از احتمال رخدادن وزن \( W \) به چهار برچسب \( T \) استفاده می‌شود. در اینجا به‌عنوان نشان دهنده استفاده می‌شود. صورت کلی این احتمال در رابطه 5 استفاده می‌شود.

\[ \mu_k(w, t) = \begin{cases} \Pr(t|w) & \text{if } \Pr(t|w) > 0, \\ 0 & \text{otherwise} \end{cases} \]

\[ \Pr(t|w) = \frac{c(w, t)}{\sum_i c(w, t_i)} \]

در رابطه 5 هنگامی که یک وزن در یک اجزای آموزشی وجود داشته باشد، مقدار آن متقابلی رابطه محاسبه می‌شود. در غیراین صورت زمانی که از این پیکردهای آموزشی نباید، برای آن مقدار یکی از این مقدارها \( k \) عددی باشد. از یک‌باره‌ای که برای کنترل گرایش استفاده می‌شود و در آموزش‌ها \( k \) مقدار ثابت 9 را دارد. 

\[ c(w, t) = \begin{bmatrix} c_{1}(w, t) & c_{2}(w, t) \end{bmatrix} \]

مقدار ثابت 9 دارد.

\[ \frac{\mu_k(w, t)}{\mu_{\min}(w, t)} = \frac{\max\{\mu_k(w, t), 0\}}{\max\{\mu_{\min}(w, t), 0\}} \]

مقدار ثابت 9 دارد.

\[ X = \{\text{DELM, ADJ_SIM, ADJ_SUP, N_SING, \ldots, DELM}\} \]

(شکل-3): ساختار شبکه عصبی برای تخمین وزن‌های الأولی

(Figure-3): Neural network structure to estimate the weight of the edges

سال 1397 شماره 4 صفحه 38

38
۴ - نتایج آزمایش‌ها

روش پیشنهادی روي داده‌های پی چن خان مورد آزمایش قرار گرفت.

در جدول (۴) نیز فراخوانی بر حسب مجموعه آموزشی درصد آورده شده است. در این جدول هشت بر حسب مجموعه آموزشی درصد داده‌های پیشنهادی پایین‌تر از مقدار مطرح می‌باشد. برای پاسخ‌های مثبت می‌تواند در سطح آخر قابل مشاهده شود. حاصل دوباره درصد را

\[
P = \frac{M}{N},
\]

\[
R = \frac{M}{L},
\]

\[
F_1 = \frac{2 \times P \times R}{P + R}
\]

بداعتایی که وزن‌گرده و بالایی شکه تعمیم شد از الگوریتم ویترین برای تعیین محتمل نرین استفاده می‌کنیم.
6- مراجع

[1] محمدرضا فیضی درخشنده، فریدن فیروزی، مهدي رحیمی، "مقایسه کارایی انجامشده برای پرچسب‌گذاری ادات سخن زبان فارسی", زبانشناسی رایانه، سومین همایش ملی زبان‌شناسی رایانه، دانشگاه صنعتی شریف، 1362.


قانون حورعلی مدرث کارشناسی خود را در رشته مهندسی برق دانشگاه صنعتی شهید啤酒 در سال 1385 و مدرث کارشناسی ارشد خود را در سال 1388 از دانشگاه صنعتی سهند تبریز دریافت کرده است. این‌شان در حال حاضر عضو هیئت علمی رشته مهندسی برق مجمع آموزش عالی استواران است. زمینه‌های پژوهشی مورد علاقه ای‌شان عبارتند از برداش‌تکنیک و برداش‌تصویر و پیکسل‌ینی‌کاپی‌بکتر. باسلامت اگو و شبکه‌های عصبی.

محمدرضا رایانامه ایشان عبارت است از:

Hourali@esfarayen.ac.ir

مروی حورعلی مدرث کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات گرایش تجارت الکترونیک از دانشگاه علم و صنعت ایران در سال 1385 و مدرث کارشناسی ارشد خود را در گرایش مهندسی فناوری اطلاعات از دانشگاه تربیت مدرس اخذ کرده است. ای‌شان در حال حاضر عضو هیئت علمی رشته هوش مصنوعی دانشگاه صنعتی مالک اشتر تهران است. زمینه‌های پژوهشی مورد علاقه ای‌شان عبارتند از برداش‌تکنیک و شبکه‌های عصبی، تحلیل‌اتلاعات در شبکه‌های اجتماعی و ساختارهای فازی، تحلیل رایانامه‌ای ایشان عبارت است از:

Mhourali@mut.ac.ir


محمد بادیم‌آقا مدرث کارشناسی خود را در رشته مهندسی نرم‌افزار از دانشگاه ژنگان در سال 1392 و مدرث کارشناسی ارشد خود را در سال 1395 از دانشگاه صنعتی مالک اشتر تهران اخذ کرده است. زمینه‌های پژوهشی مورد علاقه ای‌شان عبارتند از برداش‌تصویر و پیداکاری کامپیوتری، باسلامت اگو و شبکه‌های عصبی.

نشانی رایانامه ایشان عبارت است از:

badpeima_mohammad@mut.ac.ir