

برچسب‌گذاری ادات سخن زبان فارسی با استفاده از مدل شبکه فازی

محمد بادپیما^۱، فاطمه حورعلی^۲ و مریم حورعلی^{۳*}

^۱گروه آموزشی برق و کامپیوتر، دانشگاه مالک‌اشتر، تهران، ایران

^۲گروه آموزشی برق و کامپیوتر، مجتمع آموزش عالی اسفراین، خراسان شمالی، ایران

چکیده

برچسب‌گذاری ادات سخن یکی از مسائل مطرح در حوزه پردازش زبان‌های طبیعی است. هدف در این مسئله تعیین نقش واژگان در جمله است. برحسب این برچسب‌گذاری ویژگی‌های دستوری و نحوی واژگان نیز مشخص می‌شود. در این مقاله یک روش مبتنی بر آماری برای ادات سخن فارسی پیشنهاد شده است. در این روش محدودیت‌های روش‌های آماری با استفاده از معرفی یک مدل شبکه فازی کاهش پیدا کرده است؛ به طوری که در صورت وجود تعداد کمی داده آموزشی، مدل فازی پارامترهای قابل اطمینان‌تری را تخمین می‌زند. در این روش ابتدا هنجارسازی به عنوان پیش‌پردازش صورت گرفته و سپس فراوانی هر واژه با توجه به برچسب مربوطه به صورت یک تابع فازی تخمین زده و سپس مدل شبکه فازی تشکیل شده و درجه هر یال در این شبکه با استفاده از یک شبکه عصبی و تابع عضویت مشخص می‌شود. در نهایت بعد از این که مدل شبکه فازی برای یک جمله ساخته شد، از الگوریتم ویتربی برای تعیین محتمل‌ترین مسیر در این شبکه استفاده شده است. نتایج آزمایش روی پیکره بی‌جن‌خان کارایی این روش را تأیید کرده و نشان می‌دهد که روش پیشنهادی در شرایطی که داده‌های آموزشی کم‌تری در اختیار باشد، از روش‌های مشابه، مثل مدل مخفی مارکوف عملکرد بهتری دارد.

واژگان کلیدی: پردازش زبان‌های طبیعی، برچسب‌زنی اجزای سخن، زبان فارسی، فازی، شبکه عصبی.

Part Of Speech Tagging of Persian Language using Fuzzy Network Model

Mohammad Badpeima¹, Fatemeh Hourali² & Mayam Hourali^{3*}

^{1,3}Electrical and Computer Engineering Department, MUT, Tehran, Iran

²Electrical and Computer Engineering Department, Esfarayen 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 'Mahrami' 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

* نویسنده عهده‌دار مکاتبات • تاریخ ارسال مقاله: ۱۳۹۵/۹/۳۱ • تاریخ آخرین بازنگری: ۱۳۹۷/۷/۱۴ • تاریخ پذیرش: ۱۳۹۷/۱۰/۱۹ • Corresponding author

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 and 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.

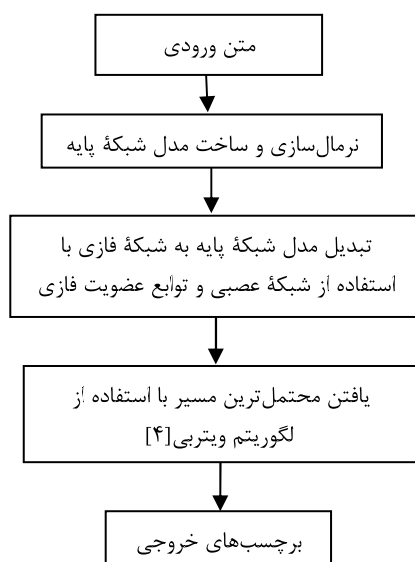
Keywords: Natural language processing, Part of speech (POS) tagging, Persian language, Fuzzy, Neural network.

«محرمی» می‌تواند اسم یا صفت باشد. ابهام موجود را می‌توان با استفاده از واژگان همسایه و یک روش برچسب‌زنی مناسب برطرف کرد. روش‌های موجود در این حوزه به چند دسته مبتنی بر حافظه [2]، مبتنی بر قانون [5]، آماری [6] و شبکه عصبی [7] تقسیم می‌شوند. دقت بیشتر این روش‌ها به‌صورت میانگین ۹۵ درصد است [1]. در مقاله [13] با استفاده از برچسب‌زنی مدل احتمالاتی TnT و هموارسازی و تغییراتی روی تخمین تابع احتمال سه‌واژه‌ای‌ها مدل برچسب‌زنی ایجاد شده که روی پیکره Penn Treebank و NEGRA در مجموع به دقت ۹۶/۷ درصد دست یافته است. در [14] با استفاده از بازنمایی شبکه وابستگی و استفاده گسترده از ویژگی‌های واژگانی همانند پیوستگی شرطی دنباله‌ای از واژگان و همچنین استفاده مؤثر از مقدم در مدل‌های شرطی لگاریتم خطی و مدل‌سازی ریزدانه از ویژگی‌های واژگان ناشناخته، به دقت ۹۷/۲۴ درصد روی پیکره Penn Treebank WSJ رسیده است. در [15] با استفاده از ماشین بردار پشتیبان مدل مؤثر و کارآمدی پیشنهاد شده که روی پیکره Wall Street Journal به دقت

۱- مقدمه

برچسب‌زنی اجزای سخن یکی از مهم‌ترین مسائل مطرح در پردازش زبان‌های طبیعی است. عملکرد روش‌های برچسب‌زنی موجود وابستگی زیادی به پیکره مورد استفاده دارد. به‌طوری‌که اگر داده‌های آموزشی و آزمون از یک پیکره استخراج شده باشند، روش‌ها عملکرد خوبی دارند یا اگر تعداد داده آموزشی کم باشد به‌خصوص در روش‌های احتمالی میزان دقت نیز کاهش می‌یابد [1]. برچسب‌زنی واژگان کار دشواری است؛ به‌دلیل این‌که اگر داده‌های آزمون از یک پیکره دیگری انتخاب شده باشند، عملکرد سامانه به‌طور معمول مطلوب نخواهد بود. نحوه برخورد با واژگانی که در داده‌های آموزشی وجود نداشتند و برای نخستین‌بار دیده می‌شوند نیز یکی دیگر از چالش‌های این حوزه است. برچسب‌زنی واژگان در حجم نمایه‌سازی و بازبایی اطلاعات نقش مهمی دارد. علاوه‌براین برچسب‌گذاری یک ابزار تحلیل نحوی بوده و همچنین به‌عنوان یک روش مناسب برای ابهام‌زدایی به‌شمار می‌آید [1]. واژگانی که در جملات استفاده می‌شوند، اغلب مبهم هستند. به‌عنوان مثال واژه

این صورت که تمام پسوندهای جمع کننده به خود واژه چسبیده می شوند؛ سپس بر اساس تمام برچسب هایی که واژگان می توانند داشته باشند، یک گراف پایه ساخته و سپس با استفاده از توابع عضویت فازی و شبکه عصبی وزن هایی برای هر یک از گره ها و یال ها اختصاص داده می شود. در نهایت این گراف وزن دار که به مدل شبکه فازی معروف است به الگوریتم ویتربی [4] داده می شود سپس این الگوریتم محتمل ترین مسیر را انتخاب می کند.



(شکل-۱): فرایند روش پیشنهادی
(figure-1): Process of proposed method

این روش روی مجموعه داده بی جن خان [3] مورد آزمایش قرار گرفته و نتایج نشان دهنده این است که وقتی داده های آموزشی کمی در اختیار داشته باشیم، این روش بهتر از مدل مخفی مارکوف مرتبه نخست عمل می کند. ساختار مقاله به صورت زیر است:

در ادامه مقاله و در بخش دوم شبکه فازی تعریف می شود. در بخش سوم نحوه ساخت شبکه فازی برای برچسب زنی ادات سخن فارسی بررسی شده است. در بخش چهارم نتایج آورده شده و بخش پنجم نیز شامل نتیجه گیری است.

۲- شبکه فازی

شبکه های فازی در میدان های مختلف فناوری و صنعت کاربرد دارد که از جمله آن ها می توان به پیش بینی [8]، سامانه های پشتیبانی تصمیم [9] و پیدا کردن بیشینه جریان [10] اشاره کرد. تعریف شبکه فازی به شدت به حوزه استفاده

۹۷/۱۶ درصد دست یافته است. در [16] از ویژگی ساخت وازی واژگان ناشناخته استفاده شده و دقت برچسب زنی این واژگان را به هشتاد درصد بهبود داده است در ضمن آزمایش های روی پیکره Chinese Treebank صورت گرفته است. در مقاله [17] با استفاده از trigram ساده برای تخمین احتمالات استفاده شده و یک برچسب زن کد باز با نام HunPos ارائه شده که دقت آن روی پیکره Penn Treebank در بهترین حالت ۹۶/۵۸ درصد است. [18] نخستین کار در زبان فارسی است که از همسایگی واژگان و توزیع شباهت آن ها استفاده کرده و دقت این سامانه ۵۷,۵ درصد است. در [19] یک برچسب زن فارسی کد باز به نام HunPoS ارائه شده و از همان روش TnT بر مبنای مدل مخفی مارکوف و دنباله سه تایی از واژگان استفاده کرده و به دقت ۹۶/۹ درصد روی پیکره بی جن خان رسیده است.

برچسب زنی ادات سخن زبان فارسی چالش های خاص خود را دارد که از جمله آن ها می توان به مواردی که در ادامه آمده اشاره کرد. شناسایی واژه های ناشناخته در متن و مشکلات خط فارسی، وجود "وند" های بسیار که به طور معمول به واژه می چسبند و باعث به وجود آمدن مشکلاتی برای واژگان با ریشه یکسان می شوند، مشکلاتی که شکل یکسان برخی تکواژها ایجاد می کنند، برخی از تکواژها همراه با واژگان یا افعال می توانند دست کم به سه شکل مختلف ظاهر شوند که خود این مسئله عامل تأثیرگذار در برچسب گذاری است و همچنین موارد دیگری که در این جا به صورت گزینشی به برخی از آن ها اشاره شد. مسئله برچسب گذاری ادات سخن نه تنها در شناسایی نقش واژگان مبهم تأثیر دارد، بلکه در کاربردهایی مانند سنتز و تشخیص گفتار، ترجمه ماشینی، بازیابی اطلاعات و غیره نقش مهمی را ایفا می کند. از یک دیدگاهی دیگر متدهای موجود برای حل این مسئله به دو دسته برچسب زنی درشت دانه و ریزدانه تقسیم می شوند که در دسته ریزدانه نقش واژگان به صورت دقیق مشخص می شود. به عنوان مثال اگر واژه ای صفت تشخیص داده شود، نوع صفت در ریزدانه مشخص می شود. روش پیشنهادی در این مقاله از دسته ریزدانه بوده و از ترکیب روش های آماری، شبکه عصبی و شبکه فازی تشکیل شده است. این مقاله دو نقطه قوت دارد. یکی شناسایی واژگان ناشناخته و دیگری رسیدن به نتایج مطلوب با توجه به داده آموزشی پائین می باشد.

شکل (۱) فرایند روش پیشنهادی را نشان می دهد. در روش پیشنهادی ابتدا هنجار سازی ساده صورت می گیرد، به

(جدول-۱): تعداد برجسب‌های متفاوت

(Table-1): Number of different labels

تعداد واژگان	تعداد برجسب‌های متفاوت
۶۲۵۶۸	۱
۴۸۳۸	۲
۴۹۳	۳
۱۱۸	۴
۴۹	۵
۱۴	۶
۷	۷
۲	۸
۲	۹
۱	۱۰

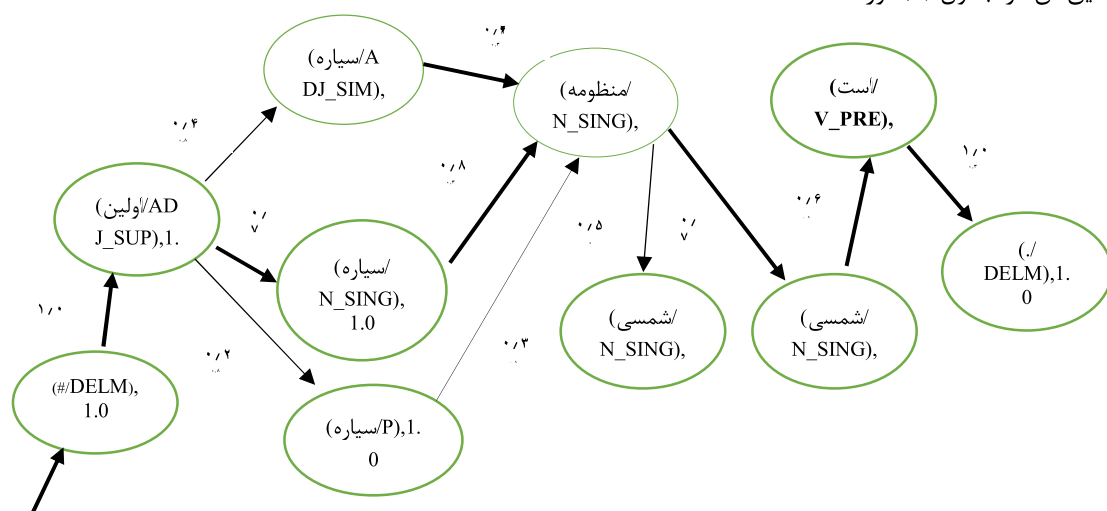
به‌عنوان مثال اگر مجموعه \tilde{V} به‌صورت زیر باشد.

$$\tilde{V} = \left\{ \begin{aligned} &((\#/\text{DELM}), 0.1), ((\text{اولین}/\text{ADJ_SIM}), 1.0), ((\text{سیاره}/\text{ADJ_SIM}), 1.0), \\ &((\text{سیاره}/\text{N_SIGN}), 1.0), ((\text{سیاره}/\text{P}), 1.0), ((\text{منظومه}/\text{N_SIGN}), 1.0), \\ &((\text{شمسی}/\text{N_SIGN}), 1.0), ((\text{شمسی}/\text{ADJ_SIM}), 1.0), ((\text{است}/\text{V_PRE}), 1.0), \\ &((\text{است}/\text{DELM}), 1.0) \end{aligned} \right\}$$

و اگر مجموعه یال‌های فازی \tilde{E} به‌صورت زیر باشد.

$$\tilde{E} = \left\{ \begin{aligned} &((\#/\text{DELM}), (\text{اولین}/\text{ADJ_SUP}), 1.0), ((\text{اولین}/\text{ADJ_SUP}), (\text{سیاره}/\text{ADJ_SIM}), 0.4), \\ &((\text{سیاره}/\text{P}), (\text{سیاره}/\text{ADJ_SUP}), 0.2), ((\text{اولین}/\text{ADJ_SUP}), (\text{سیاره}/\text{N_SIGN}), 0.7), \\ &((\text{سیاره}/\text{N_SIGN}), (\text{سیاره}/\text{ADJ_SIM}), 0.26), ((\text{سیاره}/\text{N_SIGN}), (\text{سیاره}/\text{P}), 0.8), \\ &((\text{سیاره}/\text{P}), (\text{سیاره}/\text{N_SIGN}), 0.3), ((\text{منظومه}/\text{N_SIGN}), (\text{شمسی}/\text{N_SIGN}), 0.5), \\ &((\text{منظومه}/\text{N_SIGN}), (\text{شمسی}/\text{ADJ_SIM}), 0.7), ((\text{است}/\text{V_PRE}), (\text{است}/\text{V_PRE}), 0.6), \\ &((\text{است}/\text{V_PRE}), (\#/\text{DELM}), 1.0), ((\text{سیاره}/\text{N_SIGN}), 1.0) \end{aligned} \right\}$$

شبکه فازی حاصل از دو مجموعه در شکل (۲) نشان داده شده که محتمل‌ترین مسیر با خطوط پررنگ مشخص شده است.



(شکل-۲): مثالی از شبکه فازی برای عبارت موردنظر

(Figure-2): An example of a fuzzy network for a given phrase

آن بستگی دارد. در اینجا یک تعریف جدیدی از شبکه فازی با توجه به زمینه کاری موردنظر، ارائه شده است. یک شبکه، گرافی جهت‌دار با وزن‌هایی روی لبه‌های آن است [11]. برای بازی‌سازی این شبکه می‌توان از همان روش بازی‌سازی روابط گراف‌ها در [12] استفاده کرد که به شبکه حاصل از این بازی‌سازی، شبکه فازی گفته می‌شود.

فرض کنید $G=(V,E)$ یک گراف جهت‌دار است که در آن V و E به‌ترتیب نشان‌دهنده گره‌ها و یال‌ها باشند؛ در این صورت برای تعریف شبکه فازی می‌توان فرض کرد که V یک زیرمجموعه‌ای از ضرب $W*T$ است که در آن T و W به‌ترتیب نشان‌دهنده واژگان و برجسب‌ها هستند. در این صورت گره‌های فازی می‌تواند به‌صورت رابطه (۱) تعریف شود که در آن $\mu_{\tilde{V}}(w,t)$ نشان‌دهنده تابع عضویت است.

$$\tilde{V} = \{((w,t), \mu_{\tilde{V}}(w,t)) | (w,t) \in W \times T\} \quad (1)$$

هم‌چنین می‌توان فرض کرد که E یک زیرمجموعه‌ای از ضرب $V \times V$ است. در این صورت یال‌های فازی می‌تواند توسط رابطه (۲) تعریف شود. $\mu_{\tilde{V}}(v_i, v_j)$ یک تابع عضویت از (v_i, v_j) است.

$$\tilde{E} = \{((v_i, v_j), \mu_{\tilde{V}}(v_i, v_j)) | (v_i, v_j) \in V \times V\} \quad (2)$$

بدین ترتیب شبکه فازی \tilde{G} ساخته می‌شود.

مجموعه داده بی‌جن‌خان دارای چهل برجسب بوده و حدود ۲/۶ میلیون واژه برجسب خورده است. بسیاری از این واژگان در این مجموعه داده بیش از یک برجسب دارند که آمار دقیق آن در جدول (۱) آورده شده است.

۳- ساخت شبکه فازی برای برچسب زنی ادات سخن

فرض کنید زبان فارسی مجموعه متناهی از واژگان W و برچسبها T به صورت روابط (۳) باشد:

$$\begin{aligned} W &= \{w^1, w^2, w^3, \dots, w^N\} \\ T &= \{t^1, t^2, t^3, \dots, t^M\} \end{aligned} \quad (3)$$

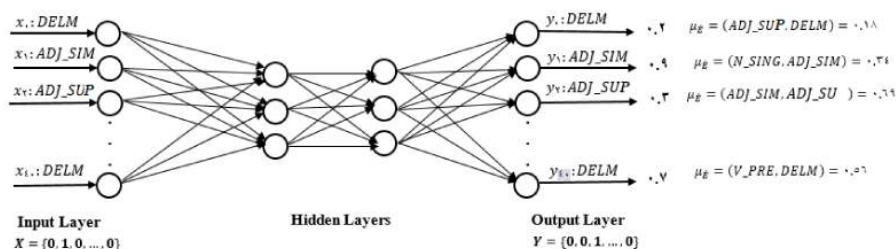
که N و M در اینجا به ترتیب نشان دهنده تعداد واژگان و برچسبها است. رابطه (۴) [20] تابع عضویت مربوط به هر گره است که در آن $w \in W$ و $t \in T$ می باشد. برای تخمین این تابع عضویت از احتمال رخ دادن واژه w با برچسب t استفاده می شود. صورت کلی این احتمال در رابطه (۵) [20] نشان داده شده است.

$$\mu_F(w, t) = \begin{cases} \sqrt[k]{\Pr(t|w)}, & \Pr(t|w) > 0 \\ 1, & 0.W \end{cases}, k \simeq 0.9 \quad (4)$$

$$\Pr(t|w) = \frac{C(w, t)}{\sum_t C(w, t)} \quad (5)$$

در رابطه (۴) هنگامی که یک واژه در پیکره آموزشی وجود داشته باشد، مقدار آن مطابق رابطه $\sqrt[k]{\Pr(t|w)}$ محاسبه می شود، در غیر این صورت زمانی که واژه ای در پیکره آموزشی نباشد، برای آن مقدار یک را برمی گرداند. از پارامتر k هم برای کنترل گرادین استفاده می شود و در آزمایشها مقدار ثابت ۰/۹ را دارد. $C(w, t)$ در رابطه (۵) نیز نشان دهنده تعداد واژه w با هر برچسبی در مجموعه داده آموزشی است. برای تعیین وزن یالها یا همان توالی مجاز برچسبها که به قواعد دستوری زبان مربوط می شود از شبکه عصبی استفاده شده است. به عنوان مثال اگر توالی برچسبها به صورت زیر باشد:

$$X = \{\text{DELM}, \text{ADJ_SIM}, \text{ADJ_SUP}, \text{N_SING}, \dots, \text{DELM}\}$$



(شکل-۳): ساختار شبکه عصبی برای تخمین وزن یالها
(Figure-3): Neural network structure to estimate the weight of the edges

در این صورت این مجموعه توالی را به بردار عددی تبدیل می کنیم که نتیجه حاصل از این تبدیل یک بردار دودویی به صورت زیر خواهد بود. طول این بردار با توجه به تعداد برچسبهای استفاده شده در مجموعه داده بی جن خان، چهل است.

$$X = \{0, 0, 1, 0, \dots, 0\}$$

ساختار شبکه نیز با توجه به مجموعه داده مورد استفاده، ساخته شده است. به همین دلیل ورودی و خروجی این شبکه با توجه به تعداد برچسبهای استفاده شده در داده بی جن خان دارای چهل نرون است؛ در ضمن دو لایه مخفی در هر لایه سه نرون در نظر گرفته شده و از تابع پرسپترون در آزمایشها استفاده شده است. ساختار کلی شبکه در شکل (۳) قابل مشاهده است. در این روش از مرتبه نخست مدل مخفی مارکوف استفاده شده است. در مرتبه یک توالی برچسبها فقط یکی از این نرونها فعال و مابقی غیرفعال هستند. شبکه با توجه به وزنهای آموزش دیده چهل مقدار را در خروجی ظاهر می کند که این مقادیر به دست آمده به عنوان خروجی شبکه عصبی به تابع عضویت (۶) [20] داده می شود. این تابع عضویت یک مقداری بین صفر و یک برمی گرداند. به عنوان مثال این خروجی با توجه به بردار ورودی X با مقدار فعال ADJ_SUB به این معنی است که بعد از این برچسب، کدام برچسب می تواند ظاهر شود. یالهای خروجیهای شبکه هستند. مقدار y_{\max}, y_{\min} کمینه و بیشینه خروجی شبکه و β نیز مقدار ثابت ۰/۹ دارد.

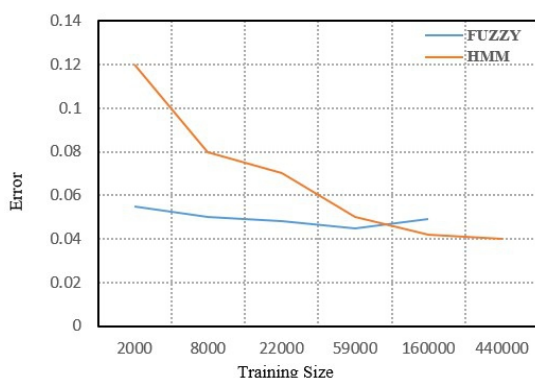
$$\mu_E(w, t) = \frac{y_j - \beta \cdot y_{\min}}{y_{\max} - \beta \cdot y_{\min}} \quad (6)$$

به خود اختصاص دادند. در این آزمایش چندین بار داده‌ها به دو دسته آموزشی و آزمایشی با اندازه‌های مختلف به صورت صعودی تقسیم شده‌اند. به عبارت دیگر در آزمایش‌های اولیه اندازه داده آموزشی را کمتر در نظر گرفتیم و در آزمایش‌های بعدی اندازه آن را افزایش داده و با الگوریتم HMM مقایسه کردیم که نتایج در شکل (۴) نشان داده شده است.

(جدول-۳): توزیع برچسب‌ها

(Table-3): Distribute labels

TAGS	Percentage distribution
N_SING	۳۸٪
P	۱۲٪
DELM	۱۰٪
ADJ_SIM	۹٪
CON	۸٪
N_PL	۶٪
N_PA	۴٪
PRO	۲٪
and etc.	۱۲٪



(شکل-۴): نتایج مقایسه فازی و مدل مخفی مارکوف

برای مرتبه نخست

(Figure-4): Compare Fuzzy Results and Hidden Markov Model for First Order

همان‌طور که در نمودار مشخص است، مجموعه آموزشی و خطا با هم دیگر رابطه مستقیم دارند؛ به طوری که با افزایش مجموعه آموزشی، خطا کاهش و با کاهش آن خطا زیاد می‌شود.

در آزمایش‌ها از سه معیار دقت، فراوانی و F1 استفاده شده و روش محاسبه این سه معیار به صورت زیر است:

$$P = \frac{M}{N} \quad (7)$$

$$R = \frac{M}{L} \quad (8)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (9)$$

شبکه عصبی قواعد دستور زبان مربوطه را یاد گرفته و به صورت مطلوب با ابهامات برخورد می‌کند. تبدیل بردار توالی برچسب‌ها برای یک جمله به بردار دودویی باعث می‌شود که تعداد حالت‌های زیادی از قواعد دستور زبان تولید شود.

در ادامه مثالی از روند کار، ارائه شده است. البته در این مثال برای سادگی و واضح تر شدن روند کار فرض شده که پیکره بی‌جن‌خان دارای شش برچسب و در نتیجه شبکه عصبی هم دارای شش نورون در ورودی و خروجی است.

با این مفروضات جمله "تهران پایتخت ایران است." را در نظر بگیرد و فرض کنید که این جمله را خواهیم با شش برچسب، برچسب‌گذاری کنیم و همچنین فرض کنید که برای واژه "پایتخت" دو برچسب L_5, L_6 مشخص شده و بقیه واژگان نیز یک برچسب داشته باشند. در این صورت مجموعه برچسب‌های واژگان جمله به صورت زیر خواهد بود:

$$\left\{ (L_6), (L_4 \text{ است}), (L_3 \text{ است}), (L_1 \text{ ایران}), (L_5, L_6 \text{ پایتخت}), (L_5 \text{ تهران}) \right\}$$

بنابراین تعداد قواعدی که از جمله مربوطه به عنوان داده آموزشی برای شبکه عصبی می‌توان تولید کرد بر طبق جدول (۲) است.

(جدول-۲): نحوه تولید داده‌های آموزشی برای شبکه عصبی

(Table-2): How to generate educational data for the neural network

Input(Data)		Output(Label)	
$\{L_1, L_2, L_3, L_4, L_5, L_6\}$		$\{L_1, L_2, L_3, L_4, L_5, L_6\}$	
L_5 , تهران	{0,0,0,0,1,0}	L_2 , پایتخت	{0,1,0,0,0,0}
L_2 , پایتخت	{0,1,0,0,0,0}	L_1 , ایران	{1,0,0,0,0,0}
L_1 , ایران	{1,0,0,0,0,0}	L_3 , است	{0,0,1,0,0,0}
L_3 , است	{0,0,1,0,0,0}	L_4 , است	{0,0,0,1,0,0}
L_4 , است	{0,0,0,1,0,0}	L_6 , *	{0,0,0,0,0,1}

بعد از این که وزن گر‌ها و یال‌های شبکه تعیین شد از الگوریتم ویتربی برای تعیین محتمل‌ترین استفاده می‌کنیم.

۴ - نتایج آزمایش‌ها

روش پیشنهادی روی داده‌های بی‌جن‌خان مورد آزمایش قرار گرفت.

در جدول (۳) نیز فراوانی برچسب‌ها بر حسب درصد آورده شده است. در این جدول هشت برچسب به صورت مجزا با درصد بیان شده و برچسب‌های باقی‌مانده همان‌طور که در سطر آخر قابل مشاهده است، حدود دوازده درصد را

6- References

۶- مراجع

[۱] محمدرضا فیضی درخشی، فرهنگ فیروزی، مهدی رحیمی، "مقایسه کارهای انجام‌شده برای برچسب‌گذاری ادات سخن زبان فارسی"، زبان‌شناسی رایانشی، سومین همایش ملی زبان‌شناسی رایانشی، دانشگاه صنعتی شریف، ۱۳۹۳.

[1] M. R. Feizi Derakhshi, F. Firozi, M. Rahimi, "Comparison of Works Performed on the Persian Part-of-Speech Tagging," *Computational Linguistics, 3rd National Conference on Computer Linguistics*, Sharif University of Technology, 2014.

[۲] مهدی حسینی، سیستم برچسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکره متنی زبان فارسی، کارشناسی ارشد، علم و صنعت، تهران، ۱۳۸۷.

[2] M. Hosseini, "Automatic labeling system and automatic disambiguation of the components of the word for the textual form of Persian language," MA, Iran University of Science And Technology, Tehran, 2008.

[3] M. BijanKhan, "The Role of the Corpus in Writing a Grammar: An Introduction to a Software", *Iranian Journal of Linguistics*, 19(2), 2004.

[4] G. D. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, pp. 268-278, 1973.

[5] E. Brill, "A simple rule-based part of speech tagger", In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP-92)*, pp. 153-155, 1992.

[6] K. W. Church, "A stochastic PARTS program and noun phrase parser for unrestricted text", In *Proceedings of Applied Natural Language Processing*, pp. 136-143, 1988.

[7] J. Benello, A. W. Mackie, and J. A. Anderson, "Syntactic category disambiguation with neural networks," *Computer Speech and Language*, vol.3, pp.203-217, 1989.

[8] H. Hidekiyo and Y. Nishkawa, "Fuzzy network technique for technological forecast-ing", *Fuzzy Sets and Systems*, pp. 99-113, 1984.

[9] H. Kawamura, "Fuzzy network for decision support systems", *Fuzzy Sets and Systems*, pp. 59-72, 1993.

[10] S. Chanas and, W. Kolodziejczyk, "Maximum flow in a network with fuzzy arc capacities", *Fuzzy Sets and Systems*, pp. 165-173, 1982.

که در این جا M, N به ترتیب تعداد کل برچسب‌های اختصاص‌داده‌شده و تعداد برچسب‌های صحیح اختصاص داده‌شده توسط مدل است. همچنین L تعداد کل برچسب‌های موجود در داده استاندارد است. در جدول (۴) پیاده‌سازی دو مدل HMM و شبکه فازی باهم مقایسه شده و نتایج نشان داده شده‌اند.

(جدول-۴): ارزیابی مدل مخفی مارکوف و شبکه‌های فازی

(Table-4): Evaluation of Hidden Markov Model and Fuzzy Networks

روش	دقت	فراخوان	F1
HMM	76%	80%	78%
Fuzzy Network	84%	81%	82%

۵- نتیجه‌گیری و کارهای آینده

مسئله برچسب‌گذاری ادات سخن نه تنها در شناسایی نقش واژگان مبهم تأثیر دارد، بلکه در کاربردهایی مانند سنتر و تشخیص گفتار، ترجمه ماشین، بازیابی اطلاعات و غیره نقش مهمی را ایفا می‌کند. درواقع هدف، تعیین نقش واژگان در جمله است. برحسب این برچسب‌گذاری، ویژگی‌های دستوری و نحوی واژگان نیز مشخص می‌شود.

در این مقاله یک روش مبتنی بر شبکه فازی ارائه شده است. در این روش برخی از چالش‌های روش‌های احتمالی بررسی و سپس از شبکه‌های عصبی و مدل شبکه فازی برای رفع این چالش‌ها استفاده شده است. در این روش محدودیت‌های روش‌های آماری با استفاده از معرفی یک مدل شبکه فازی کاهش پیدا کرده است؛ به‌طوری‌که در صورت وجود تعداد کمی داده آموزشی مدل فازی پارامترهای قابل اطمینان‌تری را تخمین می‌زند؛ درواقع این مقاله دو نقطه قوت دارد. یکی شناسایی واژگان ناشناخته و دیگری رسیدن به نتایج مطلوب با توجه به داده آموزشی پایین است. روش ارائه‌شده می‌تواند در شرایطی که اندازه داده آموزشی کم باشد، کارایی بهتری داشته باشد.

نتایج آزمایش روی پیکره بی‌جن‌خان کارایی این روش را تأیید کرده و نشان می‌دهد که روش پیشنهادی در شرایطی که داده‌های آموزشی کمتری در اختیار باشد، از روش‌های مشابه، مثل مدل مخفی مارکوف، عملکرد بهتری دارد.

در کارهای آینده برای حل این مسئله می‌توان از شبکه‌های عصبی بازگشتی یا ساختارهای یادگیری عمیق استفاده کرد.



فاطمه حورعلی مدرک کارشناسی خود را در رشته مهندسی برق از دانشگاه صنعتی شاهرود در سال ۱۳۸۵ و مدرک کارشناسی ارشد خود را در سال ۱۳۸۸ از دانشگاه صنعتی سهند تبریز اخذ کرده است. ایشان در حال حاضر عضو هیئت علمی رشته مهندسی برق مجتمع آموزش عالی اسفراین است. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: پردازش تصویر و ویدئو، بینایی کامپیوتر، بازشناسی الگو و شبکه‌های عصبی. نشانی رایانامه ایشان عبارت است از:

Hourali@esfarayen.ac.ir



مریم حورعلی مدرک کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات، گرایش تجارت الکترونیک از دانشگاه علم و صنعت ایران در سال ۱۳۸۵ و مدرک دکترای 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: پردازش متن و زبان طبیعی، تحلیل اطلاعات در شبکه‌های اجتماعی و سامانه‌های فازی. نشانی رایانامه ایشان عبارت است از:

Mhourali@mut.ac.ir

- [11] R. Sedgewick, "Algorithms in C," Addison-Wesley Publishing Company, 1990.
- [12] H.-J. Zimmermann, "Fuzzy Set Theory and Its Applications," Kluwer-Nijhoff Publishing, pp. 61-82, 1985.
- [13] T. Brants, "TnT – a statistical part-of-speech tagger," In *Proceedings of the 6th Conference on Applied Natural Language Processing*, 2000, pages 224–231.
- [14] K. Toutanova, D. Klein, Ch. D. Manning and Y. Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network", 2003,
- [15] J. Giménez, and L. Màrquez, "A general pos tagger generator based on support vector machines," In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.
- [16] H. Tseng, D. Jurafsky, and Ch. Manning. "Morphological features help POS tagg-ing of unknown words across language varieties," *Fourth SIGHAN Work-shop on Chinese Language Processing*, 2005, pp. 32-39.
- [17] P. Halacsy, A. Kornai, and C. Oravecz, "HunPos – an open source trigram tagger," In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Posters Prague, Czech Republic, 2007.
- [18] S. Mostafa ASSI and M. Haji Abdolhosseini, "Grammatical Tagging of a Persian Corpus," Institute for Humanities and Cultural Studies, 2000.
- [19] S. Mojgan, "A Statistical Part-of-Speech Tagger for Persian," Department of Linguistics and Philology, NODALIDA 2011, Riga, Latvia, May 11–13, 2011.
- [20] K. Jae-Hoon, and G. Chang Kim, "Fuzzy network model for part-of-speech tagging under small training data," *Natural Language Engineering* 2.02 (1996), pp. 95-110.



محمد بادپیما مدرک کارشناسی خود را در رشته مهندسی نرم‌افزار از دانشگاه زنجان در سال ۱۳۹۲ و مدرک کارشناسی ارشد خود را در سال ۱۳۹۵ از دانشگاه صنعتی مالک اشتر اخذ کرده است.

زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: پردازش تصویر و ویدئو، بینایی کامپیوتر، بازشناسی الگو و شبکه‌های عصبی.

نشانی رایانامه ایشان عبارت است از:

badpeima_mohammad@mut.ac.ir