برچسب‌زن اجزای کلام بهبودیافته با دانش زبانی

علی فارسی نژاد، بهرام وزیر نژاد
آزمایشگاه زبان‌شناسی رایانشی ، دانشگاه صنعتی شریف farsinejad@mehr.sharif.edu
آزمایشگاه زبان‌شناسی رایانشی ، دانشگاه صنعتی شریف bahram@sharif.ir

چکیده: برچسب‌زنی اجزای کلام(POS tagging)، یا تشخیص مقوله صرفی یک کلمه، یک گام مهم در پردازش زبان طبیعی است. در این مقاله از دانش زبانی و صرفی برای بهبود عملکرد برچسب‌زن‌های آماری استفاده شده است. بدین منظور، ابتدا چند روش برچسب‌زنی آماری بر روی متن فارسی آزمایش شده‌اند سپس عملکرد آنها با دانش زبانشناسی گنجانده در یک برچسب‌زن عبارت منظم بهبود یافته است.

کلمات کلیدی: پردازش زبان طبیعی، برچسب‌زنی اجزای کلام، عبارت‌های منظم

.

۱. مقدمه

برچسب‌زنی اجزای کلام[۱]، اختصاص مناسب‌ترین مقوله دستوری به هر کلمه در یک متن است. این مسئله از یک جست و جوی ساده در فهرستی که هر کلمه با برچسبش مشخص شده فراتر می‌رود، چون بسیاری از واژه‌های هر متن بیش از یک برچسب دستوری دارند. به عنوان مثال، کلمه "تُند" به زبان فارسی ممکن است صفت یا قید باشد. رفع این ابهامات دستوری، چالش اصلی سیستم‌ها و الگوریتم‌های برچسب‌زن اجزای کلام است. یک سیستم کامپیوتری برچسب‌زن اجزای کلام باید بتواند جمله‌ای مانند "من غذای تند خوردم" را به عنوان ورودی بگیرد و در خروجی جمله برچسب‌خورده یعنی "من/ضمیر غذای/اسم تند/صفت خوردم/فعل" را تولید کند. شکل ۱ بخشی از این اجزای کلام را نشان می‌دهد.

ADV	Adverb	قید
AJ	Adjective	صفت
CL	Classifier	شاخص
CONJ	Conjunction	حرف ربط
DET	Determiner	حرف تعریف
INT	Interjection	حرف صوت
N	Noun	اسم
NUM	Number	عدد
P	Preposition	حرف اضافه
POSTP	Postposition (را)	حرف اضافه پسین
PRO	Pronoun	ضمیر
PUNC	Punctuation	جدا کننده
RES	Residual:Arabic and Latin words, etc	متفرقه
V	Verb	فعل

شکل ۱ فهرست نمونه اجزای کلام

برچسب‌زن‌های اجزای کلام از پیمانه‌های ضروری هر سیستم پردازش زبان طبیعی[۲]، مانند تشخیص گفتار، تبدیل متن به گفتار، ابهام‌زدایی از معنی کلمه، بازیابی اطلاعات، پردازش معنایی، استخراج اطلاعات، ترجمه ماشینی و غیره هستند. به عنوان مثال، با دانستن مقوله صرفی کلمه در یک سیستم تولید ماشینی گفتار می‌توان نوای آهنگ یا تکیه مناسب را تنظیم کرد.

معمولا الگوریتم‌های برچسب‌زنی را به دو دسته روش‌های آماری و قاعده‌بنیاد تقسیم‌بندی می‌کنند. در رویکردهای آماری [۱،۲] ابتدا از روی یک پیکره برچسب خورده با

مقولات دستوری، الگوریتم مورد نظر را آموزش می‌دهیم بدین معنی که احتمال وقوع کلمات را با برچسب‌های مختلف به دست می‌آوریم. در مرحله برچسب‌زنی، بر اساس این احتمالات محاسبه شده، دنباله‌ای از برچسب‌ها انتخاب می‌شود که حاصلضرب معادله (۱)

$$P(WORD \mid TAG) \times P(TAG \mid previous\ tags). \quad (۱)$$

بیشینه کند.

رویکردهای مبتنی بر قواعد، با استفاده از قوانین و واژگان به رفع ابهام دستوری می‌پردازند. این قوانین می‌توانند یا دست‌نوشته باشند یا یاد گرفته شوند[۳]. روش‌های یادگیری ماشینی از جمله حداکثر آنتروپی، درخت تصمیم، یادگیری مبتنی بر حافظه، نیز در برچسب‌زنی اجزای کلام کاربرد دارند. ترکیبی از این رویکردها نیز می‌تواند استفاده شود. [۴]

اولین کاری که در زمینه برچسب‌زنی دستوری متون فارسی انجام شده، عاصی [۵] در سال ۲۰۰۰ است که با مجموعه برچسبی شامل ۴۵ برچسب، دقت ٪ ۵۷.۵ داشت. این سیستم قادر به ابهام‌زدایی از برچسب کلمات نیست.

بهترین نتایج توسط خانم رجا و همکارانشان [۶] در سال ۲۰۰۷ به دست آمده است. در این کار نتایج چند برچسب‌زن بر روی متون زبان فارسی بررسی شد. نتایج ارائه شده دقت ۹۴ تا ۹۷ درصد را برای کارایی سیستم نشان داد که نشان از کارایی برچسب‌زن‌های آماری برای زبان فارسی دارد. بهترین درصد برای برچسب‌زنهای مدل مخفی مارکوف ٪۹۵ درصد [۷]و برچسب‌زن‌های MLE دقت ٪۹۶ را گزارش کرده‌اند.

۲. پیکره

پیکره متنی[۳] خام مجموعه‌ای از متون یک زبان است. اما پیکره زمانی ارزش دارد که برچسب خورده باشد و اطلاعات و دانش زبانی به آن افزوده شده باشد. در پیکره‌های برچسب‌خورده، کلمات و گروهها با برچسب‌های مورد نیاز، حاشیه‌نویسی یا برچسب‌دهی شده‌اند.

از جمله این برچسب‌ها، برچسب‌های مقوله نحوی[۴] (به کلمات)، برچسب مرز عبارت، برچسب درخت تجزیه (به جملات) و یا برچسب معنایی (به کلمات یا عبارات) است.

کاربرد اصلی پیکره‌های برچسب خورده در آموزش روش‌های یادگیری ماشینی بانظارت است. پیکره متنی که در این آزمایش مورد استفاده قرار گرفته است، پیکره کوچک بیژن خان[۸] است .این پیکره از متون روزنامه‌ها جمع‌آوری شهد است و بیش از دو میلیون کلمه برچسب خورده با مقوله دستوری دارد.

مجموعه برچسب[۵] این پیکره ۴۰ برچسب دارد. قبل از برچسب‌زنی، یک سری پیش پردازش روی پیکره لازم بود، مانند: پیوستن کلمات دارای بیش از یک واحد[۶]، مانند " می کنند، کتاب ها" به هم، تبدیل نویسه‌های قدیمی فارسی پیکره به نویسه‌های استاندارد، تبدیل پیکره از یک کلمه در خط به یک جمله در خط.

ابتدا درجه ابهام کلمات در پیکره بررسی شد. به عنوان مثال، کلمه "بالا" در این پیکره، برچسب‌های متعددی مانند ADJ_SIM، ADV، ADV_NI، N_SING، P، و PRO دارد. اکثریت کلمات موجود در پیکره تنها یک برچسب (۹۱ ٪) و تنها درصد کمی از کلمات بیش از ۳ برچسب مختلف دارند. بررسی پیکره نشان می‌دهد که برچسب N_SING (اسم مفرد) پربسامدترین برچسب این پیکره است.

۳. برچسب‌زن عبارت‌های منظم

کار اصلی ما در این تحقیق، پیاده‌سازی یک برچسب‌زن عبارت منظم (regular expression) برای برچسب‌زنی کلمات ناشناخته (خارج از واژگان)[۷] در زبان فارسی است. فارسی زبانی با تصریف بالاست. در فارسی هم از وندهای تصریفی و هم از وندهای اشتقاقی برای تصریف کلمه استفاده می‌شود. از این وندها می‌توان در مورد رده دستوری یک کلمه قضاوت کرد. به عنوان مثال پسوند تصریفی "ترین" به صفت متصل می‌شود و صفت عالی می‌سازد. یا هر جا در پایان کلمه‌ای پسوند تصریفی "ها" دیدیم می‌توانیم بگوییم که با یک اسم رو به رو هستیم.

با طبقه‌بندی دانش ما از چنین پسوندهایی، یک برچسب‌زن عبارت منظم بر اساس شکل کلمه با زبان برنامه‌نویسی پایتون[۸] پیاده‌سازی شد.

ابتدا دقت الگوهای شهودی اولیه در پیکره *واژگان زایا* [۹] آزمایش شده است. در این پیکره فهرستی از کلمات فارسی، به همراه وندهای تصریفی و اشتقاقی و جایگاه انها

آمده است. به عنوان مثال، قاعده "اسم + ترین = صفت عالی" در ۹۸٪ موارد درست بوده است. در پیاده‌سازی برچسب‌زن فقط از قواعد با اطمینان بیش از ۹۵٪ استفاده شده است. برخی از این پسوندها در جدول ۱ آمده است.

جدول ۱: : برخی قواعد دستور زبان فارسی در قالب پسوندهای شناخته شده برای بکار گیری در برچسب زن نقش دستوری

تر	ADJ_CMPR	**پسوندهای صفت و قیدساز**
ترین	ADJ_SUP	
پذیر، دار ، ناک	ADJ	
انه، اً	ADV	
ها،های، ایی،هایم،هایت، هایش،هایمان، هایتان، هایشان، ین،ات، اتی، ان، انی	N_PL	**پسوندهای اسم‌ساز**
دان، زار، ستان، چی، چه، بان،اک، ار	N	
ام، ای، یم، ید، اید، اند،بودم،بودی، بود، بودیم، بودید، بودند	V_PA	**پسوندهای فعل‌ساز**
است، ست	V_PRE	
	ADJ_CMPR	

۴.آزمایش

ابتدا سه برچسب‌زن آماری N-gram یعنی unigram, bigram, trigram پیاده‌سازی شد.در برچسب‌زن‌هایNgram هدف یافتن دنباله برچسبی است که عبارت (۲) را بیشینه کند.

$$t_1^n = \text{argmax}\ P(w_1^n \mid t_1^n)\ P(t_1^n) \quad (۲)$$

در برچسب‌زن unigram فقط کلمه و برچسبش را در نظر می‌گیریم. در برچسب‌زن bigram برای یافتن برچسب هر کلمه، کلمه و برچسب قبل، و در برچسب‌زن trigram دو کلمه و برچسب قبل را در نظر می‌گیریم

سپس با روش Back off یا عقب‌گرد، ترکیب‌های مختلفی از این برچسب‌زن‌ها با هم زنجیر شده است . در روش عقب‌گرد، اگر مثلا یک دنباله سه‌تایی از برچسب‌ها پیدا نشد، از دنباله‌های دوتایی در برچسب‌زن بعدی استفاده می‌شود. نحو نامگذاری برچسب‌زن‌ها بدینگونه است که در برچسب‌زن UBT، یک برچسب‌زن ترایگرم به بایگرم و بایگرم به یونی‌گرم عقب‌گرد می‌کند.

برای برچسب‌زنی مدل مخفی مارکوف ، ما از TnT tagger [۱۰] استفاده کردیم که بر اساس مدل مرتبه دوم مارکوف کار می‌کند و می‌تواند برای زبان‌های مختلف و مجموعه برچسب‌های مختلف آموزش ببیند.

در نهایت، دقیق‌ترین این برچسب‌زن‌های آماری را با برچسب‌زن عبارت منظم (regular expression) ترکیب کردیم تا نتایج را بهبود دهیم.

۵. نتایج

هر کدام از برچسب‌زن‌ها روی ۹۰٪ جملات پیکره بیجن‌خان آموزش دیده و روی ۱۰٪ باقیمانده آزمایش شدند.

جدول ۲ نشان‌دهنده دقت هر یک از برچسب‌زنهای Ngram به تنهایی است. برچسب‌زن unigram که فقط به کلمه و برچسبش نگاه می‌کند و کاری به برچسب‌های قبل ندارد، بالاترین دقت را دارد.

جدول ۲: صحت برچسب زن های N-gram روی داده آزمون

برچسب‌زن	دقت
unigram	93.74
bigram	39.67
Trigram	20.37

جدول ۳ دقت ترکیب‌های مختلف برچسب‌زن‌های Ngram را نشان می‌دهد. از بین ترکیب‌های مختلف بر خلاف انتظار ما که UBT بیشترین دقت را داشته باشد، برچسب‌زن UTB که یک برچسب‌زن دوتایی به یک برچسب‌زن سه‌تایی و در نهایت یک‌تایی عقب‌گرد می‌کند دقیق‌ترین است.

جدول ۳: صحت برچسب زن های ترکیب N-gram روی داده آزمون

برچسب‌زن	دقت
UBT	93.92
UTB	94.01
BUT	88.06
BTU	87.59
TUB	72.27
TBU	71.90

جدول ۴ اولا نشان می‌دهد که دقت برچسب‌زن HMM از دیگر برچسب‌زن‌های آماری بیشتر بوده است. دوما مشخص می‌کند برچسب‌زن REGEX تا چه اندازه درصد برچسب‌زن‌های دیگر را بهبود داده است.

جدول ۴: : ترکیب برچسب زن UTB با مدل دانش زبانی REGEX

برچسب‌زن	دقت
UTB	94.01
HMM	95.23
+REGEX	96.72

۶. جمع‌بندی

در این مقاله با آموزش یک برچسب‌زن مدل مارکوف مخفی از مرتبه دو، و ترکیب آن با یک برچسب‌زن دانش‌بنیاد به یک برچسب‌زن برای کلمات فارسی رسیدیم.

این مقاله نشان می‌دهد که برای برچسب‌زنی اجزای کلام در زبان فارسی می‌توان با استفاده از روش‌های آماری مانند Ngrams، مدل مخفی مارکوف و ترکیب آنها با دانش زبانی به نتیجه قابل قبولی رسید.

کار بعدی می‌تواند تزریق دانش زبانی بیشتر به بخش برچسب‌زن عبارت‌های منظم و بالابردن دقت برچسب‌زن‌های آماری با تربیت آنها روی پیکره بزرگ بیجن‌خان باشد.

مراجع

[1] Christopher D. Manning, Hinrich Scheutze. Foundations of Statistical Natural Language
Processing. 1st MIT Press; 1999.

[2] Sang-Zoo Lee, Juni ichi Tsujii, Hae-Chang Rim. Lexicalized Hidden Markov Models for Part-of-Speech Tagging. In Proceedings of 18th International Conference on Computational Linguistics, Saarbrucken, Germany, August 2000.

[3] Eric David Brill. A corpus-based approach to language learning. Ph.D. Dissertation, Departmentof Computer and Information Science, University of Pennsylvania, 1993.

[4] Shamsfard, Mehrnoush and Hakimeh Fadaee. 2008. A hybrid morphology-based pos tagger for Persian. In N.Calzolari, Ed., Proceedings of LREC'08, Marrakech, Morocco.

[5] Assi, M. and Haji Abdolhosseini, M. (2000) "Grammatical tagging of a Persian corpus", International Journal of Corpus Linguistics, Vol. 5, No. 1, pp.69-82.

[6] Samira Tasharofi, Fahimeh Raja, Farhad Oroumchian, Masoud Rahgozar. Evaluation of Statistical Part of Speech Tagging of Persian Text. International Symposium on Signal Processing and its Applications, Sharjah, (U.A.E.), 2007.

[7] Morteza Okhovvat and Behrouz Minaei Bedgoli, (2011) A hidden Markov model for Persian part-of-speech tagging". Procedia Computer Science: World Conference on Information Technology 3:977-981

[8] BijanKhan, M. (2004). The role of the corpus in writing a grammar: An introduction to a software. *IranianJournal of Linguistics*.

[9] واژگان زایای زبان فارسی، *زبان فارسی و رایانه*، ج۱، انتشارات سمت، تهران. اسلامی، محرم و همکاران. ۱۳۸۹

[10] Brants, T., "TnT - A Statistical Part-of-Speech Tagger", In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, April 29 – May 3, 2000 Seattle, WA.

[1] Part of speech tagging
[2] Natural language processing
[3] Corpus
[4] Part of speech
[5] tagset

[6] Multi Unit Tokens
[7] Out Of vocabulary words
[8] Python