بررسی پیکره ای مناسب برای برچسب زنی کلمات در زبان فارسی

هادی امیری*، حسین حجت†، فرهاد ارومچیان‡§

چکیده

یکی از اقدامات اساسی و اولیه در حوزه پردازش زبانهای طبیعی تهیه یک پیکره مناسب می باشد. نویسندگان در این تحقیق سعی در بررسی آماری یک پیکره فارسی به نام پیکره بیجن خان [1] می کنند ، همچنین مجموعه عملیات پیش پردازشی که روی این پیکره جهت کاربردی کردن آن انجام شده است شرح داده می شود. در این تحقیق، به عنوان یکی از کاربردها، از پیکره بیجن خان امکان تهیه یک نرم افزار برچسب زنی کلمات زبان فارسی برررسی شده است. با توجه به تجزیه و تحلیل های آماری ارزشمندی که روی این پیکره انجام شده است و با اصلاحات انجام شده نشان داه می شود که با استفاده از یک روش ابتدایی (تخمین بیشینه احتمال) نیز می توان به دقت قابل قبولی در برچسب زنی خودکار رسید.

قابل ذکر است که پیکره اصلاح شده بیجن خان از گروه تحقیقاتی بردازش زبان های طبیعی در دانشگاه تهران قابل تهیه می باشد.

کلمات کلیدی

پردازش زبانهای طبیعی، برچسبزنی کلمات، پردازش زبان فارسی، تولید پیکره، تخمین بیشینه احتمال

Investigation on a Feasible Corpus for Persian POS Tagging

Hadi Amiri, Hosein Hojjat, Farhad Oroumchian

Abstract

One of the fundamental works in natural language processing is creating a feasible corpus for evaluating effectiveness of different algorithms. In this paper, the authors report creation of test corpus of automatic part of speech tagging purposes based on the Persian tagged corpus of Prof. Bijankhan. This study includes preprocessing , statistical analysis and experiments with simple statistical POS tagging methods done on this corpus. Part of speech tagging experimental results show that even with a simple POS method such as Maximum Likelihood Estimation (MLE), we could reach to an get acceptable accuracy of 93.16 percent.

It should be mentioned that this corpus is available in natural language processing group at university of Tehran.

Keywords

Natural Language Processing, Part of Speech Tagging, Persian Natural Language Processing, Collection Building, Maximum Likelihood Estimation.

* دانشگاه تهران، دانشکده مهندسی برق و کامپیوتر، h.amiri@ece.ut.ac.ir.

† دانشگاه تهران، دانشکده مهندسی برق و کامپیوتر، h.hojjat@ece.ut.ac.ir.

‡ دانشگاه تهران، قطب علمی کنترل و پردازش هوشمند دانشکده برق و کامبیوتر دانشگاه تهران، oroumchian@acm.org.

§ دانشگاه Wollongong دبی، دانشکده فناوری اطلاعات، farhadoroumchian@uowdubai.ac.ae.

۱- مقدمه

یکی از اقدامات اساسی و اولیه در حوزه پردازش زبانهای طبیعی تهیه یک پیکره مناسب می باشد. مجموعه پیکره های تگ زده شده متعددی برای زبان های مختلف تا کنون به وجود آمده است که از مهمترین آنها می توان به پیکره Penn Treebank [7] اشاره کرد.

پیکره ای که در این تحقیق از آن استفاده شده است بخشی از پیکره برچسب زده شده پروفسور بیجن خان است[1] که در آزمایشگاه زبان شناسی دانشگاه تهران نگهداری می شود. این پیکره از برخی اخبار روزنامه ها و متون معمولی جمع آوری شده است . یکی از ویژگی های بارز این پیکره آن است که هر سند در این مجموعه دارای یک عنوان می باشد. به عنوان مثال اسناد تحت عناوین "سیاسی"، "فرهنگی"، "اقتصادی" و... دسته بندی شده اند. قابل ذکر است که در این پیکره ۴۳۰۰ عنوان مختلف وجود دارد . این دسته بندی بزرگ نشان دهنده کیفیت بالای این پیکره می باشد. در عملیات برچسب زنی از عناوین متون صرف نظر شده است زیرا که هدف بدست آوردن یک نرم افزار برچسب زننده خودکار بود. این پیکره با مجموعه غنی از برچسب ها، شامل ۵۵۰ برچسب مختلف، برچسب زنی شده است . این مجموعه برچسب برای برچسب زدن دقیق و جزئی کلمات به کار گرفنه می شود اما از آن جایی که در برچسب زنی خودکار هدف مشخص کردن کلمات از نظر نوع کلی آنها می باشد و وارد جزئیات نمی شود [7] معمولا از این تعداد برچسب استفاده نمی شود. از طرف دیگر از آنجایی برچسب زنی خودکار بر اساس یادگیری ماشینی می باشد، در نظر گرفتن مجموعه بزرگی از برچسب ها روش های یادگیری ماشینی برای برچسب زنی را در مرحله یادگیری با مشکل مواجه می کند و عملا آن را غیر ممکن می کند . بنابراین در اولین مرحله هدف کاهش تعداد برچسبها می باشد، در بخش دوم شیوه کاهش مجموعه برچسب را تشریح می کنیم.

۲- انتخاب برچسب ها

همانطور که ذکر شد مجموعه بزرگ برچسب ها یادگیری را غیر ممکن می سازد. اکثر ابزارهایی که برای برچسب زنی کلمات به کار گرفته می شود به یک مجموعه برچسب کوچکتر از مجموعه برچسب موجود نیاز دارند. از طرفی روشهایی مانند شبکه های عصبی مصنوعی ،درخت های تصمیم و بسیاری روشهای احتمالاتی دیگر، با این مجموعه بزرگ از برچسب ها نمی توانند فرایند یادگیری را به خوبی انجام دهند و خروجی تولید شده قابل قبول نمی باشد. جهت کاهش اندازه مجموعه برچسب برخی تحلیل آماری روی پیکره انجام شد و طی مراحلی که در زیر تشریح می شود این مجموعه کاهش یافت.

در پیکره بیجن خان برچسب ها به خوبی بازنمایی شده اند به این معنی که هر برچسب در این مجموعه از یک ساختار سلسله مراتبی پیروی می کند، بخشهایی از نام برچسب که در ابتدای نام آن قرار دارند بیان کننده توصیف کلی تری از آن برچسب می باشد و بخشهایی که در انتهای نام برچسب قرار دارند بیان کننده توصیف جزیی تر در مورد آن برچسب هستنند. به عنوان مثال برچسب "N-PL-LOC" (نام برچسب ها به همراه توضیح آنها در جدول شماره ۶ در بخش ضمیمه آورده شده است) دارای سه سطح در ساختار سلسله مراتبی می باشد،سطح اول،"N"، مشخص کننده اسم می باشد، سطح دوم ،"PL"، مشخص کننده نوع جمع می باشد و سطح سوم،"LOC"، مشخص کننده این است که برچسب در مورد مکان است . به عنوان مثالی دیگر "N-PL-DAY" تشریح کننده اسمی است که جمع است و یک تاریخ را توصیف می کند.

عملیات کاهش بر اساس مراحل زیر انجام شد::

- در مرحله اول آن دسته از برچسب هایی را که در ساختار سلسله مراتبی دارای سه یا بیشتر سطح هستنند به برچسب هایی با دو سطح کاهش داده می شوند. بنابراین هر دو برچسب فوق به برچسبی به نام "N-PL" تبدیل می شوند. برچسب جدید مشخص کننده اسم جمع می باشد بدون آنکه چیزی در مورد مکان ویا تاریخ بیان کند. بعد از این مرحله تعداد برچسب های موجود به ۸۱ عدد کاهش یافت.
- در بین برچسب های باقی مانده برخی برچسب های عددی وجود داشت که پس از مراجعه به پیکره متوجه شدیم که این برچسب ها صحیح نیستند و به علت اشتباهاتی که در فرآیند بر چسب زنی به وقوع پیوسته است به وجود آمده اند. برای جلوگیری از کاهش صحت بر چسب زن ها، نام این برچسب ها به برچسب "DEFAULT" تغییر داده شد. بنابراین تعداد برچسب ها به ۷۲ عدد کاهش یافت.
- در مرحله سوم ، برخی از برچسب های دو سطحی غیر ضروری به یک سطح کاهش یافتند. این بر چسب های غیر ضروری جزء آن دسته از برچسب ها هستند که به ندرت در پیکره به کار برده شده اند. پس از این مرحله تعداد برچسب ها به ۴۲ عدد کاهش یافت.
- در این مرحله عملیات کاهش مربوط به دو برچسبی بود که به ندرت در پیکره به کار برده شده بودند (کمتر از ۲ بار) یعنی دو برچسب "N" و "V_SNFL". با توجه به رابطه معنایی که بین این دو برچسب و سایر برچسب ها وجود داشت، این برچسب ها اصلاح شدند. برچسب "N" به اسم مفرد نزدیک است و ما آنرا با برچسب "N_SING" جایگزین کردیم، از طرف دیگر معنی برچسب "V_SNFL" به هیچکدام از برچسب های پیکره نزدیک نیست و ما به آسانی آنرا از مجموعه برچسب ها حذف کردیم. پس از این مرحله تعداد برچسب ها به ۴۰ عدد رسید.

۳- تحلیل پیکره

در اکثر روشهای برچسب زنی کلمات، پیکره به دو دسته داده آموزشی و داده آزمایشی تقسیم می شود، از داده های آموزشی جهت یادگیری استفاده می شود مثل تنظیم پارامترهای برچسب زن، و از داده های آزمایشی جهت ارزیابی برچسب زن استفاده می شود.

در این بخش برخی تحلیل های آماری سودمند ارائه می شود که روی دو مجموعه آموزشی و آزمایشی پیکره انجام شده اند. جدول شماره ۱ نام برخی برچسب ها را همراه با تعداد دفعات تکرار و درصد فراوانی آنها در دو مجموعه آموزشی و آزمایشی و به عبارت دیگر در کل پیکره نمایش می دهد (لسیت کامل توزیع برچسب ها در جدول شماره ۵ در بخش ضمیمه آورده شده است). بررسی این جدول نشان می دهد که برچسب "N_SING" بیشترین تعداد دفعات تکرار را در هر دو مجموعه داراست، این برچسب در مجموعه داده های آموزشی ۸۲۶۵۷۱ بار و در مجموعه داده های آزمایشی ۱۴۰۹۷۵ بار تکرار شده است. از طرف دیگر برچسب "NN" کمترین تعداد دفعات تکرار را در هر دو مجموعه اراست (۲ بار در مجموعه داده های آموزشی و ۰ بار در مجموعه داده های آزمایشی).

جدول(۱): توزیع برخی برچسب ها

نام برچسب	فرآوانی در مجموعه آموزشی	درصد فرآوانی در مجموعه آموزشی	فرآوانی در مجموعه آزمایشی	درصد فرآوانی در مجموعه آزمایشی
MS	8	0.000	253	0.065
NN	2	0.000	0	0.000
OHH	15	0.001	5	0.001
ADJ	21	0.001	1	0.000
NP	42	0.002	10	0.003
DEFAULT	48	0.002	32	0.008
CON	177769	8.056	32523	8.311
ADJ_SIM	192171	8.709	38980	9.961
DELM	217533	9.858	39062	9.982
P	270894	12.276	48964	12.513
N_SING	826571	37.459	140975	36.026

نکته قابل توجه دیگر در این جدول آن است که تنها برچسب های "ADJ_SIM", "CON", "DELM", "N_PL", "N_SING", "P" درصد فراوانی بیشتر از ۶٪ در دو مجموعه داده های آموزشی و آزمایشی دارند و بقیه برچسب ها درصدی کمتر از این مقدار دارند. از طرفی با توجه به اینکه داده های آزمایشی به طورکاملا تصادفی انتخاب شده اند و با توجه به جدول شماره ۱ درصد فرآوانی برچسب ها در دو مجموعه آموزشی و آزمایشی تقریبا مشابه می باشد می توان نتیجه گرفت که توزیع این برچسب ها تقریب خوبی از توزیع برچسب ها در زبان فارسی می باشد. در زیر برخی اطلاعات مفید دیگر از پیکره بیجن خان نشان داده می شود.

۳-۱- تعداد برچسب های متفاوت برای هر کلمه

یکی دیگر از اطلاعات مفید در مورد پیکره بیجن خان تعداد برچسب موجود برای هر کلمه می باشد، به عنوان مثال کلمه "آسمان" در کل پیکره همواره برچسب "N_SING" را گرفته است در حالی که کلمه "بالا" برچسب های متفاوتی را در شرایط متفاوت گرفته است، برچسب هایی نظیر ..ADJ. جدول شماره ۲ نشان دهنده تعداد کلماتی در پیکره است که به ترتیب دارای ۱۰،...،۳،۲،۱ برچسب می باشند.

دانستن این اطلاعات دید خوبی نسبت به نقش کلمات در زبان فارسی به ما می دهد.

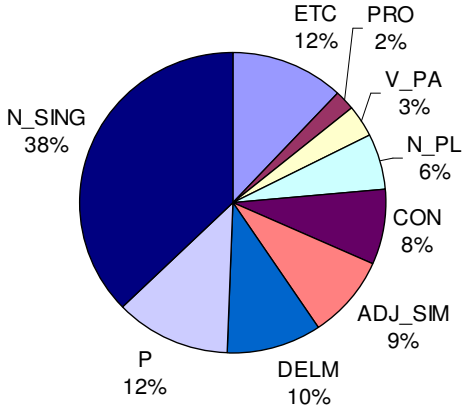
جدول(۲) : تعداد بر چسب های متفاوت

تعداد کلمات	تعداد برچسب های متفاوت
62568	1
4838	2
493	3
118	4
49	5
14	6
7	7
2	8
2	9
1	10

۳-۲- دسته بندی برچسب های کم تکرار در یک گروه

مطالعه جدول شماره یک نشان می دهد که تعداد زیادی از برچسب ها تعداد دفعات کمی در پیکره به کار برده شده اند لذا قرار دادن آنها در یک گروه جهت داشتن دید بهتری نسبت به توزیع کلمات در پیکره می تواند مفید باشد. ما این برچسب ها را در یک گروه به نام ETC قرار دادیم و برچسب هایی که برای این گروه انتخاب کردیم آن دسته از برچسب ها هستند که تعداد دفعات تکرار آنها در پیکره کمتر از ۵۰۰۰ دفعه می باشد. با این گروه جدید توزیع برچسب ها در پیکره به صورت شکل ۱ می باشد.

شکل (۱): توزیع بر چسب ها



۴- روش احتمال بیشینه

تا کنون روش های مختلفی جهت تگ زنی کلمات ارائه شده است که از مهمترین آنها می توان از روشهای تگ زنی مبتنی بر حافظه[1] [4] مانند روش MBT [5] و Trigrams'n'Tags [6] نام برد. یکی از روشهایی که برای تگ زنی به کار گرفته می شود روش تخمین احتمال بیشینه[2] است. در این بخش با استفاده از روش تخمین احتمال بیشینه به برچسب زنی کلمات می پردازیم. در این روش برای هر کلمه در مجموعه داده های آموزشی، برچسبی انتخاب می شود که بیشتر از سایر برچسب ها به آن کلمه در کل مجموعه نسبت داده شده باشد. برای این منظور احتمالات بیشینه را برای هر کلمه و برچسب های آن محاسبه می شود و سپس برچسبی را که بیشترین احتمال را دارد به عنوان برچسب انتخابی برای آن کلمه انتخاب می شود [2,3]. جهت ارزیابی این روش داده های آزمایشی را در نظر می گیریم. به ازای هر کلمه در این مجموعه ما برچسب انتخابی را به آن نسبت می دهیم و سپس روش را مورد ارزیابی قرار می دهیم. جدول شماره ۳ این ارزیابی را نشان می دهد، در این جدول برای هر کلمه دیده نشده (کلمه ای که قبلا در بین داده های آموزشی دیده نشده است) برچسب "DEFAULT" در نظر گرفته می شود.

در جدول شماره ۳ درصد صحت این روش برای کلمات دیده شده و دیده نشده در مجموعه داده های آزمایشی نشان داده شده است. برای محاسبه این درصد باید تعداد دفعاتی را که برچسب انتخابی به درستی به کلمات نسبت داده شده است، محاسبه شود، پس از آن درصد جهت براساس فرمول ۱ به دست می آید.

درصد صحت= تعداد کلمات / ۱۰۰*(تعداد برچسب های انتخابی که صحیح نسبت داده شده اند) (۱)

جدول شماره ۳ شرایطی را نشان می دهد که ما به کلمات دیده نشده برچسب "DEFUALT" را نسبت داده ایم که تنها در ۷ مورد برای این کلمات درست عمل کرده ایم، سایر کلمات دیده نشده برچسبی غیر از برچسب "DEFUALT" را داشته اند. با این روش صحت کلی به 90.43% می رسد.

جدول (۳): تخمین احتمال بیشینه با برچسب "DEFAULT" برای کلمات دیده نشده

کلمات	تعداد	تعداد برچسب های صحیح زده شده	درصد صحت
دیده شده	3700716	353862	95.45
دیده نشده	20594	7	0
مجموع	391310	353862	90.43

یکی از راه حل های ساده برای به دست آوردن درصد صحت بهتر این است که به کلمات دیده نشده به جای برچسب "DEFUALT" برچسبی که بیشترین تعداد دفعات تکرار را در مجموعه داشته است، یعنی برچسب "N_SING" را نسبت می دهیم. نتایج این ایده در جدول ۴ نشان داده شده است. صحت کلی به 93.16% افزایش پیدا می کند.

جدول (۴): تخمین احتمال بیشینه با برچسب "N_SING" برای کلمات دیده نشده

کلمات	تعداد	تعداد برچسب های صحیح زده شده	درصد صحت
دیده شده	370716	353862	95.45
دیده نشده	20594	10713	52.02
مجموع	391310	364575	93.16

همانطور که در این جدول نشان داده شده است صحت کلی به 93.16% افزایش پیدا می کند.

۵- نتیجه گیری

ما در این تحقیق عملیات لازم جهت کاهش مجموعه برچسب ها و تولید یک مجموعه برچسب مناسب جهت برچسب زنی کلمات در زبان فارسی ارائه کردیم. همچنین برخی تجزیه و تحلیل های آماری مفید روی پیکره بیجن خان انجام شد و یک روش ساده اما پایه ای جهت برچسب زنی کلمات مورد ارزیابی قرار گرفت. گروه تحقیقاتی پردازش زبان های طبیعی در دانشگاه تهران در حال کار روی روش های متفاوت برچسب زنی کلمات با استفاده از پیکره بیجن خان می باشند.

سپاسگزاری

با تشکر فراوان از دکتر فیلی که ما را در جهت تهیه پیکره راهنمایی کردند و همچنین با تشکر فراوان از دکتر بی جن خان که با فعالیت ارزنده خود چنین پیکره با ارزشی را برای زبان فارسی به وجود آورده اند.

ضمایم

جدول(۵): توزیع برچسب ها

نام برچسب	فرآوانی در مجموعه آموزشی	درصد فرآوانی در مجموعه آموزشی	فرآوانی در مجموعه آزمایشی	درصد فرآوانی در مجموعه آزمایشی
ADJ	21	0.001	1	0.000
ADJ_CMPR	5968	0.270	1475	0.377

1.199	4693	1.020	22503	ADJ_INO
0.217	849	0.260	5743	ADJ_ORD
9.961	38980	8.709	192171	ADJ_SIM
0.256	1001	0.287	6342	ADJ_SUP
0.057	224	0.059	1291	ADV
0.203	793	0.109	2398	ADV_EXM
0.045	177	0.087	1917	ADV_I
0.044	173	0.068	1495	ADV_NEGG
0.834	3265	0.845	18635	ADV_NI
0.221	863	0.343	7564	ADV_TIME
0.300	1175	0.105	2318	AR
8.311	32523	8.056	177769	CON
0.008	32	0.002	48	DEFAULT
9.982	39062	9.858	217533	DELM
1.563	6115	1.803	39783	DET
0.140	547	0.117	2575	IF
0.001	2	0.005	111	INT
0.052	204	0.128	2823	MORP
0.028	111	0.011	250	MQUA
0.065	253	0.000	8	MS
6.375	24945	6.139	135474	N_PL
36.026	140975	37.459	826571	N_SING
0.000	0	0.000	2	NN
0.003	10	0.002	42	NP
0.003	12	0.012	271	OH
0.001	5	0.001	15	OHH
12.513	48964	12.276	270894	P
0.032	125	0.034	755	PP
2.062	8067	2.438	53792	PRO
0.009	37	0.013	296	PS
0.683	2673	0.578	12745	QUA
0.135	528	0.149	3281	SPEC
0.610	2386	0.611	13484	V_AUX
0.029	113	0.047	1044	V_IMP
1.790	7003	3.335	73591	V_PA
1.842	7209	1.599	35286	V_PRE
2.686	10512	1.868	41226	V_PRS
1.336	5228	1.296	28592	V_SUB
	140975		826571	بیشترین
	0		2	کمترین
	391310		2206627	مجموع

جدول(۶): تعریف برچسب ها

نام برچسب	توزیع
ADJ	صفت
ADJ_CMPR	صفت تفضیلی
ADJ_INO	صفت مفعولی
ADJ_ORD	صفت ترتیبی
ADJ_SIM	صفت ساده
ADJ_SUP	صفت عالی
ADV	قید
ADV_EXM	قید مثال
ADV_I	قید پرسشی
ADV_NEGG	قید نفی
ADV_NI	قید غیر پرسشی
ADV_TIME	قید زمان
AR	عربی
CON	حرف ربط
DEFAULT	بر چسب پیش فرض
DELM	تمام جدا کننده
DET	حرف تعریف
IF	ادات شرط
INT	حرف صوت
MORP	تکواژ
MQUA	سور گستر
MS	علامت ریاضی
N_PL	اسم جمع
N_SING	اسم مفرد
NN	گستره عدد
NP	گروه اسمی
OH	حرف ندا
OHH	منادی
P	حرف اضافه
PP	گروه حرف اضافه ای
PRO	ضمیر
PS	جمله واره
QUA	سور
SPEC	کیفیت نما
V_AUX	فعل کمکی
V_IMP	فعل امری
V_PA	فعل ماضی
V_PRE	فعل اسنادی
V_PRS	فعل حال
V_SUB	فعل التزامی

مراجع

[۱] بی جن خان،**"نقش پیکره های زبانی در نوشتن دستور زبان: معرفی یک نرم افزار رایانه ای"**، مجله زبانشناسی، سال ۱۹، شماره ۲، پاییز و زمستان ۱۳۸۳

[2] Allen, J., *Natural Language Understanding*, Second Edition. Benjain/Cummings Publishing Company, 1995.

[3] Manning, C. D., Schutze, H., Foundations of Statistical Natural Language Processing, MIT Press, 1999.

[4] Daelemans, W., J.Zavrel, *Recent Advances in Memory-Based Part of Speech Tagging*, *VI Simposio Internacionale de Comunicacion,* 1999.

[5] Daelemans, W., J.Zavrel, P.Berck, S.Gillis, *MBT: A Memory-Based Part of speech Tagger Generator*, In Proc. of the Fourth Workshop on Very Large Corpora, Copenhagen: ACL SIGDAT, 14-27, 1996.

[6] Brants, *T.: TnT a Statistical Part of Speech Tagger*, In Proc. of the sixth conference on applied natural language processing (ANLP-2000), 2000.
[7] Marcus M., Santorini B., Ann M., *Building a large annotated corpus of English: The Penn Treebank*, Computational Linguistics, Vol.19, No. 2, 1993.

زیرنویس‌ها

[1] Memory Based Part of Speech Tagging
[2] Maximum Likelihood Estimation