IDMC2008
11-12 NOV
Amir Kabir University

بررسی نقش برچسب‌زنی کلمات در نمایه‌سازی و دقت بازیابی برای اسناد و پرس‌وجوهای فارسی

امیر حسین جدیدی نژاد[۱]، فریبرز محمودی[۲]

چکیده

بررسی تاثیر برچسب‌زنی کلمات و تعیین ادات سخن بر کارایی و دقت بازیابی و همچنین حجم نمایه سیستم‌های بازیابی اطلاعات یکی از مباحث داغ در زمینه پردازش زبان‌های طبیعی می‌باشد. تحقیقات گوناگونی تاکنون در زبان انگلیسی جهت بررسی نقشِ ادات سخن و اهمیت آن در حجمِ نمایه و دقتِ بازیابی صورت گرفته است. در این نوشتار برآنیم تا با برچسب‌زنی خودکار اسنادِ پیکره همشهری، بعنوان بزرگترین پیکره استاندارد فارسی، نقش هر یک از ادات سخن را در حجمِ نمایه و همچنین دقتِ بازیابی بررسی نماییم. برای این منظور ابتدا پیکره همشهری برچسب‌گذاری شده و سپس از پیکره برچسب‌گذاری شده جهت تعیین نقش هر برچسب در بازیابی اسناد و پرس‌وجوهای فارسی استفاده شده است. نتیجه این تحقیق، زمینه‌ساز بسیاری از پژوهش‌ها در حوزه‌ی بازیابی اطلاعات فارسی با رویکرد زبان‌شناسی خواهد بود.

کلمات کلیدی

پردازش زبان طبیعی، بازیابی اسناد فارسی، نمایه‌سازی، برچسب‌گذاری خودکار اسناد، برچسب زنی کلمات.

Evaluating Part-of-speech tags in indexing and precision for Persian text retrieval

Amir Hossein Jadidinejad, Hadi Amiri

ABSTRACT

One of the scientific works in the field of information retrieval is studying the effect of term's Part-of-speech (POS) tags on the retrieval precision. In this area of research, a large number of investigations have been done on English texts and it would be very valuable to study the effect of tagging on other languages. In this paper, using two biggest and standard Persian test collections, Bijankhan and Hamshahri, we want to evaluate the affect of the POS tagging on the index size and retrieval precision of Persian text. Output of this paper is initiative for future investigation in automatic Persian text retrieval from linguistic point of view.

KEYWORDS

Natural Language Processing, Persian Text Retrieval, Indexing, Part-of-speech

[۱] دانشجوی کارشناسی ارشد، گرایش نرم‌افزار، دانشگاه آزاد اسلامی قزوین. amir@jadidi.info

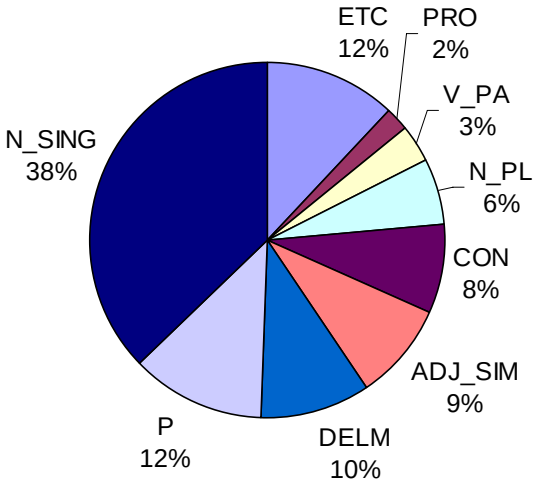
[۲] عضو هیات علمی دانشگاه آزاد اسلامی قزوین. mahmoudi@itrc.ac.ir

۱. مقدمه

برچسب‌زنی کلمات فرآیندی است که طی آن هر کلمه در متن با نحو موردنظر آن نشانه‌گذاری می‌شود. دانستن نحو هر کلمه می‌تواند کمک زیادی به حذف ابهام در تلفظ و معنای لغات و درک صحیح متن نماید. لذا کاربرد بسیار مهمی در بازیابی اطلاعات و پردازش زبان طبیعی ایفا می‌کند. در این نوشتار به مطالعه‌ی تاثیر برچسب‌های نحوی مختلف در زبان فارسی پرداخته و برای این منظور ابتدا پیکره‌ی استاندارد همشهری را با استفاده از مدل‌های آماری و ابزارهای مرتبط [۳]، [۷] برچسب‌گذاری نموده و سپس به بررسی تاثیر برچسب‌های مختلف روی دقت بازیابی بر مبنای استانداردهای TREC [۴] پرداخته‌ایم. البته بایستی توجه نمود که تاثیر برچسب‌های مختلف بسیار وابسته به نوع پرس‌وجو می‌باشد. به عبارت دیگر، بسته به پرس‌وجوی انتخابیِ کاربر و نوع نیاز اطلاعاتیِ او، ممکن است برچسب‌هایی مورد اهمیت واقع شوند که در پرس‌وجوی دیگر، اهمیت چندانی نداشته باشند.

۲. کارهای مرتبط

تاثیر برچسب‌های نحوی مختلف پیشتر در زبان انگلیسی مورد بررسی قرار گرفته است [۵]، [۶]. در این نوشتار سعی داریم ضمن برچسب‌گذاری پیکره استاندارد همشهری، تاثیر برچسب‌های مختلف را در بازیابی اطلاعات فارسی بررسی نماییم. پیکره بی‌جن‌خان [۱] یک پیکره برچسب‌گذاری شده و غنی بوده که شامل برخی از اخبار روزنامه‌ها و متون معمولیِ جمع‌آوری شده می‌باشد. پیکره اولیه حاوی ۵۵۰ برچسب مختلف بوده و از آنجاییکه در کاربردهای بازیابی اطلاعات و برچسب‌زنی خودکار متون بسیاری از برچسب‌های جزیی منجر به مشکلات متعدد در یادگیری ماشین و نویز زیاد در خروجی می‌گردد، لذا در [۲] ضمن بررسی آماری خصوصیات مختلف پیکره بی‌جن‌خان، راهکاری جهت حذف برچسب‌های غیرضروری و کاربردی نمودن آن جهت استفاده در مقاصد بازیابی اطلاعات و پردازش زبان طبیعی صورت پذیرفته است و درنهایت ۴۰ برچسب بعنوان برچسب‌های مهم شناسایی شده است. شکل ۱ توزیع برچسب‌ها در پیکره بی‌جن‌خان را پس از حذف بسیاری از برچسب‌های غیرضروری نشان می‌دهد.



شکل ۱: توزیع برچسب‌ها در پیکره بی‌جن‌خان، پس از حذف برچسب‌های غیرضروری [۲]

۳. نتایج آزمایشات

در این بخش روال آزمایشات انجام شده جهت بررسیِ نقش هر برچسب در دقت بازیابیِ اسناد و پرس‌وجوهای فارسی شرح داده شده است. برای این منظور دو مرحله آزمایش صورت گرفته است. از آنجاییکه قصد داریم دقت هر برچسب را روی پیکره‌ی استاندارد و بزرگ همشهری [۱۰] بررسی نماییم لذا در مرحله اول، کل اسناد موجود در پیکره همشهری مطابق با [۳] برچسب‌گذاری شده‌اند. و سپس در مرحله دومِ آزمایش، به

بررسی نقش هر برچسب در دقت بازیابیِ اسناد و پرس‌وجوهای فارسی پرداخته‌ایم. آزمایشات صورت گرفته حاکی از آن است که برچسب‌های اسم، صفت، فعل و قید بامعناترین برچسب‌های نحوی در زبان فارسی بوده که با استفاده از آن‌ها می‌توان ضمن کاهش حجم نمایه، دقت بازیابی را حفظ نمود. در ادامه این موارد شرح داده می‌شوند.

۱.۳. برچسب‌گذاری پیکره همشهری

[۷] یکی از کاراترین ابزارهای برچسب‌گذاری می‌باشد که با دقت بسیار بالایی قادر به برچسب‌گذاری اسناد بسیاری از زبان‌های طبیعی می‌باشد. با توجه به اینکه این ابزار رویکردی آماری در برچسب‌گذاری خودکار متون دارد لذا یکی از ویژگی‌های منحصربفرد آن، این است که برای یک زبان مشخص طراحی نشده است و در عوض می‌تواند روی هر مجموعه آموزشی دلخواه و به همراه هر مجموعه از برچسب‌ها، آموزش دیده و پس از آن در منبع هدف با دقت بالایی برچسب‌گذاری نماید. پیشتر این ابزار جهت برچسب‌گذاری اسناد زبان‌های انگلیسی [۷، ۹]، فارسی [۳]، آلمانی [۷] و ... مورد استفاده قرار گرفته است. نتایج گزارش شده در [۳] حاکی از آن است که این ابزار توانایی برچسب‌زنی متون فارسی با دقت ۹۶.۶۴٪ را داراست. در این نوشتار، ابتدا بر پایه‌ی تحقیقات صورت گرفته در [۳] تمامی اسناد پیکره همشهری [۱۰] با استفاده از پیکره بی‌جن‌خان [۱] بعنوان پیکره آموزشی، برچسب‌گذاری گردیده است و سپس از پیکره‌ی برچسب‌گذاری شده جهت بررسی نقش هر برچسب در بازیابی استفاده نموده‌ایم. از آنجاییکه پیکره آموزشی مورد استفاده در این بخش کاملا منطبق با پیکره هدف می‌باشد لذا نتایج بدست آمده دور از انتظار نیست. جهت توضیحات بیشتر در این زمینه، به [۳] مراجعه فرمایید.

۲.۳. بررسی نقش هر برچسب در دقت بازیابی

جهت مطالعه‌ی تاثیر برچسب‌های نحوی مختلف، از مجموعه‌ای شامل برچسب‌های نحوی مطالعه شده در [۲] و [۸] استفاده نموده‌ایم. در هربار اجرای آزمایش، کلمات متناظر با برخی از برچسب‌ها را در متن هر سند نگه داشته و بقیه را از متن اسناد حذف نموده‌ایم و درنهایت بر مبنای اسناد تغییر داده شده نمایه‌سازی و بازیابی نموده‌ایم. در نهایت نتایج آزمایشات مختلف با یکدیگر مقایسه شده‌اند. جدول ۱ نتایج دقت و فراخوانی را به همراه برخی از معیارهای مطرح در بازیابی اطلاعات نشان می‌دهد.

تحقیقات پیشین [۵] حاکی از اهمیت اسامی در بازیابی اطلاعات دارند لذا اسامی را بعنوان عنصر پایه‌ای در آزمایشات مدنظر قرار داده‌ایم. همانطور که در جدول ۱ نشان داده شده است، صفات اهمیت فوق‌العاده‌ای در بازیابی اطلاعات فارسی داشته و منجر به رشد چشمگیری در دقت بازیابی در تمامی معیارهای مطرح گردیده است. این مهم در مقایسه با نتایج گزارش شده در زبان انگلیسی [۶] قابل تامل است. از طرفی با توجه به اینکه صفات تنها بخش نسبتا اندکی از مستندات فارسی را شامل می‌شوند (حدود ۹٪ در پیکره بی‌جن‌خان، مطابق شکل ۱)، لذا نگهداری آنها سربار کمی را به نمایه تحمیل کرده و در مقابل تاثیر بسیار زیادی در دقت بازیابی دارد. بنابراین صفات نیز مانند اسامی یکی از مهمترین برچسب‌های زبان فارسی بوده و ترجیح می‌دهیم تا در نمایه حضور داشته باشد.

اگرچه ضمایر بهبود اندکی را در بازیابی شامل می‌شوند اما ترجیح می‌دهیم آنها را در نمایه نگهداری کنیم چراکه حجم بسیار اندکی از متون فارسی را ضمایر تشکیل می‌دهند (حدود ۲٪ در پیکره بی‌جن‌خان، مطابق شکل ۱). بنابراین نگهداری ضمایر نیز می‌تواند مفید واقع شود.

جدول ۱: نتایج دقت/فراخوانی روی پیکره برچسب‌گذاری شده همشهری

فراخوانی	دقت			
	اسم	اسم/صفت	اسم/صفت/فعل	اسم/صفت/فعل/قید
۰.۰	۰.۶۹۱۵	۰.۷۶۸۷	۰.۷۷۴۵	***۰.۷۹۰۵***
۰.۱	۰.۵۶۶۶	۰.۶۴۷۸	۰.۶۵۴۷	***۰.۶۵۷۳***
۰.۲	۰.۵۰۰۶	۰.۵۷۸۹	۰.۵۹۰۵	***۰.۵۹۰۸***
۰.۳	۰.۴۲۳۴	۰.۵۱۵۳	۰.۵۲۰۷	***۰.۵۲۳۰***
۰.۴	۰.۳۶۶۹	۰.۴۵۷۲	۰.۴۷۱۸	***۰.۴۷۱۶***
۰.۵	۰.۳۲۶۹	۰.۴۰۹۳	۰.۴۲۷۲	***۰.۴۲۷۶***
۰.۶	۰.۲۶۹۱	۰.۳۳۶۵	۰.۳۴۴۷	***۰.۳۴۵۲***
۰.۷	۰.۱۸۶۱	۰.۲۴۵۸	۰.۲۶۲۳	***۰.۲۶۲۹***
۰.۸	۰.۱۴۴۵	۰.۱۹۲۸	۰.۱۹۶۰	***۰.۲۰۴۴***
۰.۹	۰.۰۶۲۵	۰.۰۹۰۰	۰.۰۹۱۸	***۰.۰۹۱۷***
۱.۰	۰.۰۳۸۱	۰.۰۴۹۲	۰.۰۴۹۸	***۰.۰۴۹۲***

MAP	۰.۳۰۱۱	۰.۳۶۸۵	۰.۳۷۷۰	***۰.۳۷۹۱***
GMAP	۰.۱۸۲۹	۰.۲۸۹۲	۰.۳۰۲۸	***۰.۳۰۵۳***
R-PREC	۰.۳۵۷۹	۰.۴۲۰۵	۰.۴۲۰۲	***۰.۴۲۱۲***

۴. نتیجه‌گیری و کارهای آینده

در این نوشتار به بررسی نقش و جایگاه هر یک از برچسب‌های نحویِ فارسی پرداخته و تاثیر آنها را در دقت بازیابی بررسی نمودیم. برای این منظور ابتدا کلِ پیکره همشهری را توسط یک متد آماری و با استفاده از پیکره بی‌جن‌خان بعنوان پیکره آموزشی، برچسب‌گذاری نموده و سپس نقش هر یک از برچسب‌های نحوی را در دقت بازیابی بررسی نمودیم. در نهایت می‌توان نتیجه گرفت که مجموعه‌ی برچسب‌های اسم، صفت، فعل و قید مجموعه‌ای کارا در بازیابی اطلاعات فارسی بوده که ضمن پایین آوردن حجم نمایه و عدم ذخیره‌سازی کلمات کم‌کاربرد، دقت بازیابی را تا حد زیادی حفظ می‌کند لذا می‌تواند بعنوان یک مجموعه مناسب جهت نمایه‌سازی اسناد فارسی استفاده گردد.

یکی از کاربردهای مهمی که در این نوشتار به آن پرداخته نشد، اعمال روشی مشابه برای پرس‌وجوهای فارسی می‌باشد. مسلما برچسب‌گذاری پرس‌وجوها و بررسی نقش هر برچسب در بازیابی کار ارزشمندی است که می‌تواند یافته‌های این نوشتار را کامل کند. علاوه بر آن بررسی و پیاده-سازی سایر روش‌های برچسب‌گذاری خودکار متون و استفاده از روش‌هایی جهت بهبود دقت نیز می‌تواند به کیفیت بازیابی کمک کند.

۵. مراجع

[۱] بی‌جن‌خان، نقش پیکره‌های زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای، مجله زبانشناسی، سال ۱۹، شماره ۲، پاییز و زمستان ۱۳۸۳.

[۲] هادی امیری، حسین حجت، فرهاد ارومچیان، بررسی پیکره‌ای مناسب برای برچسب‌زنی کلمات در زبان فارسی، دوازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران، ۱۳۸۶.

[3] 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.

[4] National Institution of Standards and Technology: http://trec.nist.gov/.

[5] Chirag Shah and Pushpak Bhattacharyya. A Study for Evaluating the Importance of Various Parts of Speech (POS) for Information Retrieval (IR). International Conference on Universal Knowledge and Language (ICUKL) 2002.

[6] Klavans, J. and Kan, M. Role of verbs in document analysis. In Proceedings of the 17th international Conference on Computational Linguistics - Volume 1 (Montreal, Quebec, Canada, August 10 - 14, 1998). International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 680-686. 1998.

[7] T. Brants,TnT a Statistical Part of Speech Tagger, In Proc. of the sixth conference on applied natural language processing (ANLP-2000), 2000.

[8] Farhad Oroumchian, Samira Tasharofi, Hadi Amiri, Hossein Hojjat, Fahime Raja. Creating a Feasible Corpus for Persian POS Tagging. Technical Report, no. TR3/06, University of Wollongong in Dubai, 2006.

[9] R. Mihalcea, "Performance Analysis of a Part of Speech Tagging Task," in Proc. Computational Linguistics and Intelligent Text Processing, Gelbukh A. Editor, Centro de Investigacin´en Computacin´IPN, México, 2003.

[10] A. Aleahmad, H. Amiri, F. Oroumchian, and M. Rahgozar. "Hamshahri: A standard Persian text collection". White Paper, Database research Group, University of Tehran, 2008.