ساخت پیکره های نشانه گذاری شده با رویکرد وب بعنوان پیکره

مهدی مرادی
کارشناسی ارشد
دانشگاه صنعتی شریف
mehdi_moradi@mehr.sharif.ir

پروانه خسروی زاده
استادیار
دانشگاه صنعتی شریف
khosravizadeh@sharif.ir

بهرام وزیرنژاد
استادیار
دانشگاه صنعتی شریف
bahram@sharif.ir

چکیده : در این تحقیق، اولین پیکره فارسی نظرات برچسب خورده از نظر قطبیت[1] (مثبت و منفی بودن) برای کاربردهای تحلیل نظر کاربر و نیز پیکره برچسب خورده از لحاظ جنسیت برای تشخیص جنسیت نوسنده متن، بصورت خودکار از اسناد html بدست آمده از صفحات اینترنت ساخته شد. ایده بکاربرده شده در این تحقیق، استفاده از تگ های اچ‌تی‌ام‌ال دارای میزان رضایت و نام نویسنده در صفحه نظرات بوده است. کاربرد این دو پیکره جمع آوری شده در این تحقیق استفاده در طبقه بندی و تحلیل متون است .

کلید واژه: پیکره های زبانی، تحلیل نظر[2]، تشخیص جنسیت نویسنده[3]

[1] Polarity
[2] Sentiment analysis
[3] Gender classification

۱. مقدمه :

پیکره های الکترونیکی زبانی بعنوان حجمی گسترده و ساخت یافته از متون یک زبان، از مهم ترین منابع برای انجام مطالعات زبانی هستند که در چند دهه اخیر رشد چشمگیری داشته اند. آندرو کهوهه[۴](۲۰۰۳) و آنتونیت رنوف[۵](۲۰۰۳) می گویند که در واقع بدون این پیکره ها انجام بسیاری از پژوهش های زبان شناختی و تصمیم گیری ها حوزه مهندسی زبان امکان پذیر نمی باشد.از آنجا که ساخت و حاشیه نویسی دستی پیکره ها پروسه ای طولانی مدت و هزینه بر است محققان در صدد یافتن راهی برای کاهش زمان و هزینه ساخت این منابع زیرساختی زبان هستند .عصر ارتباطات و گسترش استفاده از نامه‌های الکترونیکی، پیام‌های فوری (IM)، اسناد، وبلاگ‌ها، مقالات خبری، صفحات خانگی،تالارهای گفتگو و اسناد چاپی، متن را به اصلی‌ترین وسیله‌ی ارائه و انتقال اطلاعت تبدیل کرده است.روزانه میلیون ها کاربر حجم گسترده از این متون را تولید و بروزرسانی می کنند بدون آنکه توجهی به کاربرد زبانشاختی این متون داشته باشند.گیلس ماووریس ده اسکرویور(۲۰۰۲)می گویدمتون قابل خوانش توسط ماشین ،وب را تبدیل به اصلی ترین منبع در حوزه مطالعات زبان شناسی پیکره ای ،زبانشناسی رایانشی، بازیابی اطلاعات و متن کاوی کرده است. محققان این حوزه توانسته اند با کمک محتوای وب ،ابزارهای و منابع مختلفی را برای کارهای علمی پژوهشی توسعه دهند.ما نیز در این تحقیق به کمک ابزارهای پردازش زبان طبیعی پیکره از متون برچسب خورده از نظر "احساسی " را از اینترنت گرداوری کرده ایم.

۲. **انگیزش** :

پیشرفت های اخیر در حوزه پردازش زبان طبیعی به میزان چشمگیری مرهون مطالعات پیکره‌بنیاد بوده است. مشاهدات و ارائه فرضیات زبانشاختی زبان شناسان رایانشی به کمک این پیکره انجام گرفته است.برای غنی تر کردن پیکره معمولا طی فرایندی آنها را حاشیه نویسی یا نشانه گذاری می کنند .حاشیه نویسی و افزودن برچسب های موضوعی[۶]،موازی[۷]، اجزای کلام[۸]،معنایی[۹] ،جنسیتی[۱۰] ، تحصیلاتی[۱۱] ،احساسی[۱۲] لما[۱۳]،گفتمان[۱۴]،

---

[4] Andrew Kehoe

[5] Antoinette Renouf

[6] Genre

وابستگی[۱۵] پیکره ها را غنی تر و کاربردی تر میکند. پیکره های گردآوری شده برای زبان فارسی(پیکره بی جن خان) بسیار محدود و انگشت شمار هستند.

۳. پیکره قطبیت

یکی از پیکره های پرکاربرد،پیکره های نشانه گذاری شده از نظر احساسات (ذهنی ونه عینی) نویسنده است .کاربرد این نوع پیکره ها بیشتر بعنوان داده آموزشی برای برنامه های طبقه بندی کننده متون به احساسات منفی و مثبت است.این برنامه های مبتنی بر رویکردهای یادگیری ماشینی[۱۶] پس از آموزش دیدن از روی این متون ،قادر خواهند بود متون جدید بدون برچسب و نشانه را در یکی از کلاس های تعریف شده طبقه بندی کنند .این پیکره ها می توانند کابردهای مختلف تجارتی(تحلیل نظر کامنت های مصرف کنندگان محصول یا خدمات برای مثال گوشی موبایل یا هتل وغیره)،سیاستی(تحلیل نظر افراد جامعه از طریق بررسی کامنت هایی که در یک سایت خبری برای مطلبی سیاسی گذاته اند) ویا اجتماعی (یافتن افراد یا جامعه های هم فکر و نظر در شبکه های اجتماعی).

۴. پیکره تفکیک شده براساس جنسیت :

این نوع پیکره ها بر اساس جنسیت نویسنده (مرد یا زن) آن برچسب گذاری شده اند.از این نوع پیکره ها می توان برای دسته بندی خودکار متون از نظر جنسیتی در کاربردهای تجاری (بررسی میزان محبوبیت یک کالا در بین زنان و مردان) و استفاده کرد .

۵. مواد و روش ها :

رویکردهای پیکره ای به وب به دسته الف)وب بعنوان پیکره[۱۷] و ب)وب برای پیکره[۱۸]تقسیم می شوند که در رویکرد اول توسط ابزارهایی (مانند google API) بصورت مستقیم طی جستجو بصورت برخط[۱۹] از خود

---

[7] parallel
[8] POS tags
[9] Sense
[10] gender
[11] education class
[12] sentiment
[13] lemma
[14] Discourse
[15] dependency
[16] Machine learning
[17] Web as corpus
[18] Web for corpus

بعنوان پیکره ای پویا[20] استفاده می شود و در رویکرد دوم بخشی از محتوای وی توسط ابزارهای خیزشگر[21]مورد خیزش قرار گرفته و پس ازذخیره سازی، محتوی بصورت برون- خط[22] بعنوان پیکره ای ایستا[23] مورد استفاده قرار می گیرد.روش اتخاذ شده دراین مطالعه رویکرد دوم بوده است.

۵,۱ .ساخت پیکره :

پیکره ها و ابزارهای مرتبط با آنها که تاکنون برای زبان انگلیسی ساخته شده اند قابل مقایسه با زبان های دیگر نیست .ذکر تمام این پیکره ها از مجال این مقاله خارج است ولی دیوید لی[24] لیستی طویل از پیکره های انگلیسی و غیر انگلیسی را در سایتش تهیه کرده است .معدود پیکره های موجود برای زبان فارسی عمری کمتر از یک دهه دارند .از پیشگامان گردآوری پیکره برای زبان فارسی بی جن خان (۱۳۸۶) و عاصی (۱۳۸۳) می باشند.جدیدترین پیکره فارسی dotIR است که بر اساس یک خیزش بزرگ ۸,۵ میلیون صفحه ای از دامنه ir. بدست آمده است.

۵,۲. دامنه :

برای گردآوری پیکره از صفحات وب ابتدا بایستی دامنه(سایت) یا دامنه های مورد نظر برای انجام خیزش مشخص گردند.سپس به کمک خیزشگرهای خودکار(مانند Heritrix)یا زبان های برنامه نویسی (مانند python) صفحات مرتبط با آن دامنه جهت انجام پردازش های بعدی بارگذاری می شود.

صفحات فارسی که کاربران نظراتشات را در مورد محصولی اظهار کرده باشند معدود است. در بیشتر موارد نظرات فاقد تگ یا نشانه ای(برای مثال شکلک) دال بر مثبت یا منفی بودن نظر کاربر هستند در این مواقع از افرادی خواسته می شود تا بعد از خواندن نظر با برچسب منفی یا مثبت آنرا نشانه گذاری کنند .ایده استفاده شده در این تحقیق استفاده از نظراتی بوده است که خود کاربران بعد از نوشتن نظر ،میزان رضایت از محصول مورد نظر را نیز به درصد وارد کرده اند .سایت هلوکیش (hellokish) به بازدیدکنندگان اطلاعاتی درمورد هتل ها،رستوران ها مراکز فروش جزیره کیش ارائه می کند،بخشی از سایت به *نظرات درباره اقامت و هتلها* اختصاص

---

[19] On-line
[20] Dynamic
[21] Crawler
[22] Off-line
[23] Static
[24] David lee

داده شده است .کاربران در این بخش با وارد کردن نام، محل زندگی و متن نظر، میزان درصد رضایتشان را نیز بصورت مضربی از ده وارد می کنند.

۵,۳. خیزش :

برای خیزش در این سایت و استخراج تمامی صفحات حاوی نظرات، برنامه ای تحت زبان برنامه نویسی پایتون نوشته شد.این برنامه با وصل شدن به این سایت،صفحات تعریف شده در محدوده مورد نظر را بارگذاری و در یک فایل متنی ۱۰ مگابایتی ذخیره کرد .

۵,۴. استخراج نظرات از صفحات :

صفحات ذخیره شده در مرحله قبل علاوه بر متن نظرات حاوی بخش های زاید نیز هستند. فرمت این صفحات html است یعنی قسمت های مختلف (نام هتل، متن نظر، نام نویسنده، میزان رضایت و ...) هرکدام با تگ خاصی مشخص شده است. شکل یک بخشی از یک صفحه حاوی نظر در این سایت را نشان میدهد.برای پارس(دسترسی به قسمت های مختلف متن) کردن صفحات html از برنامه BeautifulSoup استفاده شد.این برنامه پارسری است که قابلیت های بالایی در پارس صفحات HTML/XML دارد و بصورت کتابخانه ای در پایتون وارد[۲۵] می شود.

[25] **import**

```
<div id="comment">
<div class="titr">مهسا</div>
<div id="PublishComm" dir="rtl">من 17 تا 20 تیر هتل آریان بودم. اتاق ها و سرویس بهداشتي خیلي تمیز و خوبي داشت.
صبحانه معمولي بود. تنوع زیادي نداشت و هر روز تکراري بود. یه روز از رستوران هتل ناهار بشقاب دریایي سفارش دادم
که شامل میگو و ماهي شیر به صورت کبابي و سوخاري شده میشد که میشه گفت خوشمزه بود. قیمتش هم 15 هزار تومان بود.
در کل میشه گفت هتل خوبي بود. برخورد کارکنان و مسوول گردشگري محترمانه و خوب بود. من تاکید زیادي روي تمیزي اتاق ها
مخصوصا حمام داشتم که هتل آریان این ویژگي ها رو خیلي خوب داشت. به پاساژها هم نزدیک بود. تا پاساژ مروارید 10 دقیقه
پیاده راه بود. تا زیتون هم یه کم پیاده روي داشت. در کل من از هتل راضي بودم.         <br> <br />
<span class="month">امتیاز دهي:<br/></span>
اتاقها: 10 از 10<br />
خدمات: 10 از 10<br />
رستوران: 8 از 10<br />
<br />
تاریخ سفر:

شهر محل سکونت: تهران
<td width="49" class="tdata-lable">80%</td>
<td width="69" class="tdata-lable" dir="rtl">رضایتمندي:</td>
</tr>
</table>
<table width="304" border="0" align="center" cellpadding="0" cellspacing="0">

<td width="77" height="30" class="tdata-lable">
 بلي</td>        <img src="../images/yes.png" alt="yes" width="20" height="20" /> 
<td width="227" class="tdata-lable">آیا این هتل را براي اقامت توصیه مینمایید؟</td>
```

(شکل ۱)

بخشی از یک صفحه حاوی نظر در سایت هلوکیش.تگ های مشخص کننده متن نظر،نام نظردهنده و میزان رضایت بترتیب با رنگهای سبز،صورتی و بنفش مشخص شده اند

.

نظرات با توجه به درصد وارد شده به دو دسته تقسیم شدند، نظراتی که درصد رضایت آنها کمتر یا مساوی ۳۰٪ بود و نظراتی که درصد رضایتشان بالاتر یا مساوی ۷۰٪ بود .تنها نظراتی مورد توجه بودند که کاربر بعد از نوشتن متن،میزان رضایت را نیز از منوی کشوی بالای نظر انتخاب کرده بود.

۵,۵. مشخصات داده :

کل نظرات ثبت شده در این سایت در تاریخ ۹۰/۸/۱۲، ۳۳۱۲ نظر بوده است،که از این تعداد ۶۴۲ نظر دارای درصد رضایت بوده اند. ۱۰۲ نظر در دسته منفی و ۴۴۷ نظر در دسته نظرات مثبت قرار گرفتند .هر نظر در یک فایل متنی جداگانه با فرمت utf-8 در فولدر مربوطه اش(نظرات مثبت در فولدر pos و نظرات منفی در فولدر neg) ذخیره شد. میانگین کلمات در هر نظر ۱۰۹ کلمه بود.

۵,۶. استخراج داده های تفکیک شده از لحاظ جنسیتی:

برای استخراج نظرات مرتبط با نظر دهندگان زن و مرد،ابتدا با جستجو در چندین سایت و وبلاگ فارسی، اسامی زن و مرد استخراج شد. بعداز حذف اسامی تکراری ،لیستی از اسامی فارسی زن و مرد تفکیک شده تهیه گشت .تعداد اسامی دختران ۳۲۵۴ نام و تعداد اسامی پسران ۲۶۵۱ نام بود .در صفحات مربوط به نظرات در سایت هلوکیش نام هر نظر دهنده در درون تگ <"div class="titr> (در شکل ۱با خط قرمز مشخص شده است) قرار دارد،اگر نام های موجود در این تگ ها در یکی از لیست های اسامی وجود داشت،نظر مورد نظر استخراج شده و در یک فایل متنی مربوط به آن جنسیت ذخیره شد.

نتیجه گیری :

با انجام این تحقیق مشخص شد که وب قابلیت استفاده برای ساخت پیکره های زبانی را دارد. دو پیکره ساخته شده در این تحقیق کاربردهای مختلفی میتوانند داشته باشند. هرچند به منظور کاربردهای تحلیل نظر و تشخیص نویسنده تبدیل زبان گفتاری به نوشتاری لازم نیست،ولی بعنوان کارهای آینده و برچسب زنی این متون میتوان این متون را به زبان معیار تبدیل کرد .

منابع :

B. Pang, L. Lee, and S. Vaithyanathan, (2002) “Thumbs up? Sentiment classification using machine learning techniques,” in Proc Conf on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86.

Gilles-Maurice De Schryver. (2002) Web for/as Corpus: A Perspective for the African Languages**.** *Nordic Journal of African Studies 11(2)* 266-282.

http://sourceforge.net/projects/archive-crawler
http://www.python.org
http://www.crummy.com/software/BeautifulSoup/
http://www.esm.ir
http://koodakname.persianblog.ir

Kehoe, A. & A. Renouf(2002) WebCorp: Applying the Web to Linguistics and Linguistics to the Web**.** WWW2002 Conference, Honolulu, Hawaii.

Morley B., A. Renouf & A. Kehoe(2003) Linguistic Research with the XML/RDF aware WebCorp Tool WWW2003 Conference, Budapest.

Renouf, A. (2003) ‘WebCorp: providing a renewable data source for corpus linguists’, in S. Granger & S. Petch-Tyson (eds.) Extending the scope of corpus-based research: new applications, new challenges. Amsterdam: Rodopi.

بی‌جن‌خان، محمود (۱۳۸۶)، *پیکره متنی زبان فارسی*، پژوهشکده پردازش هوشمند علایم.

عاصی ، مصطفی (۱۳۸۳)،پایگاه داده های زبان فارسی در اینترنت،پژوهشگاه علوم انسانی و مطالعات فرهنگی

احسان درودی، هما برادران هاشمی، ابوالفضل آل احمد، علی محمد زارع بیدکی، امیرحسین حبیبیان، فرزاد مهدیخانی، آزاده شاکری، مسعود رهگذر، مجموعه محک استاندارد برای تحقیقات بازیابی اطلاعات وب فارسی، ارسال شده برای داوری در مجله مهندسی برق و کامپیوتر ایران، پاییز ۱۳۸۸.