بسمه تعالی

درخواست ارائه پیشنهاد (RFP) پروژهٔ ایجاد پیکره و ابزار برچسب‌گذار نقش معنایی (SRL) برای زبان فارسی

زیرمجموعه: طرح جویشگر

نسخه: ۱,۰

مهر ماه ۹۵

سطح دسترسی: عمومی

فهرست مطالب

مقدمه

پردازش زبان فار سی نیز در سال‌های اخیر مورد توجه محققان ب سیاری در ایران و سایر نقاط جهان قرار گرفته ا ست. متأ سفانه یکی از گلوگاه‌های پردازش زبان فار سی در د سترس نبودن منابع زبانی کافی و ابزار پیش‌پرداز شی معتبر برای این زبان است.

درک زبان توسط ماشین آن گونه که برای انسان اتفاق می‌افتد هدف اصلی بسیاری از پژوهش‌های پردازش زبان طبیعی ا ست. برای درک معنا لازم ا ست تا تحلیل‌ها در سطوح مختلف زبان انجام گیرد. تا کنون فعالیت‌های خوبی در سطح صرف و نحو انجام گرفته ا ست و برچ سب‌گذاری صرف و نحو تا حدود زیادی د ستر سی به معنا را محقق می‌کنند اما به دلیل تنوعات زبانی و پیچیدگی‌های آن برای تحلیل زبان کافی نیستند. بنابراین برای دستیابی به معنا نیاز است تا فراتر از صرف و نحو رفته تا بتوان تحلیل‌ها را در سطح معنا نیز انجام داد. استخراج نقش‌های معنایی یکی از گام‌های اصلی در بازنمایی معنی متن است. نقش‌های معنایی، ارتباط معنایی بین فعل و آرگومان‌های آن در جمله را مشخص می‌کنند.

به طور کلی روش‌های تعیین نقش‌های معنایی را می‌توان به دو دسته روش‌های مبتنی بر قاعده و روش‌های مبتنی بر یادگیری(آماری) تقسیم‌بندی کرد. در روش های مبتنی بر قواعد، تحلیل های معنایی متن به کمک لغتنامه‌ها، گرامرها و سایر منابع معنایی انجام می‌گیرد. این منابع بیشتر به صورت دستی تهیه می‌شوند.

در همین اواخر به منظور برچسب‌زنی نقش‌های معنایی استفاده از روش‌های یادگیری ماشین مورد توجه قرار گرفته. در این د سته روش‌ها از پیکره‌هایی که جملات آنها به صورت د ستی برچ سب‌گذاری معنایی شده‌اند جهت استخراج قواعد به صورت خودکار استفاده می‌شود. مهمترین پیکره‌هایی که در زبان انگلیسی برای این منظور ایجاد شده‌اند عبارتند از FrameNet و PropBank.

برای زبان فارسی نیز فعالیت‌های متعددی انجام شده است. به عنوان نمونه، موسسه نور پیکره گزاره‌های معنایی زبان فار سی را تهیه کرده ا ست که شامل ۳۰۰۰۰ جمله و حدود ۵۰۰ هزار کلمه ا ست. همچنین فرهنگ ظرفیت معنایی افعال زبان فار سی که تو سط مو سسه نور تهیه شده شامل حدود ۱۴۰۰۰ فعل ساده و مرکب زبان فار سی همراه با ظرفیت معنایی آنها می‌با شد. همچنین در پژوه شگاه ارتباطات و فناوری اطلاعات تحقیقات گ سترده‌ای در قالب پروژه «ایجاد و راه‌اندازی سامانه پرسش و پاسخ خودکار قرانی» در این حوزه انجام شده است.

با این وجود هنوز ابزار کاربردی که بتواند در اختیار دیگر توسعه‌دهندگان محصولات و خدمات قرار گیرد وجود ندارد. بنابراین یکی از اقدامات ا سا سی در تهیه ب ستر پردازش معنایی متون فار سی تهیه ابزار برچ سب‌گذاری نقش معنایی متون فارسی است.

تعریف پروژه

به طور کلی برچسب‌گذاری نقش‌های معنایی عبارت است از بازشناسی رویدادهای (افعال) موجود در جمله و سپس تعیین آرگومان‌های معنایی آنها شامل کنشگر، کنش‌پذیر، مکان، زمان، حالت انجام، ابزار، مبدأ، مقصد و غیره. این فعالیت یکی از فعالیت‌های اساسی در پردازش عمیق جملات بوده و دارای پتانسیل فراوان جهت استفاده در بسیاری دیگر از حوزه‌های پردازش زبان طبیعی از جمله پرسش و پاسخ، ترجمه ماشینی، خلاصه‌سازی متون، خطایابی و غیره می‌باشد. این فعالیت به صورت سنتی مرحله بعدی و پیشرفته‌تر از برچسب‌زنی نحوی درنظر گرفته می‌شود.

در پروژه حاضر، پیکره و ابزار برچسب‌زن نقش‌های معنایی (SRL) برای متون فارسی مدنظر است. ورودی سامانه یک سند متنی است و خروجی آن سندی است که متن اصلی به همراه نقش معنایی کلمات می‌باشد.

۲ هدف پروژه

به منظور ایجاد سامانه‌ها و خدمات هوشمند و با کیفیت در حوزه زبان فارسی، به تفسیر معنایی متون فارسی مورد نیاز می‌باشد. هدف از اجرای این پروژه توسعه یک برچسب‌زن معنایی برای زبان فارسی است که برای نیل به آن باید ابتدا آرگومان‌های افعال زبان فارسی تعیین شوند، سپس به تمام جملات دادگان آموزش به صورت دستی برچسب‌های معنایی زده شده و در نهایت با استفاده از این دادگان، یک برچسب‌گذار نقوش معنایی برای زبان فارسی توسعه داده شود.

۳ مراحل اجرا و شرح خدمات پروژه

پیشنهاددهنده می‌بایست مراحل و فازبندی انجام کار را با جزئیات کامل ارائه کند. توجه به موارد زیر در پیشنهاد ارسالی ضروری است:

موارد کلی

- درصورتی‌که پیشنهاددهنده بخشی یا تمام محصول را در اختیار دارد، باید وضعیت محصول به طور دقیق ذکر شود. بیان جزئیات محصول باید شامل مشخصات کلیدی، دقت، کیفیت، سرعت پاسخ‌گویی و معماری سخت‌افزار و نرم‌افزار مورد استفاده باشد.
- پیشنهاد ارسالی محدود به موارد مشخص شده در این مستند نمی‌باشد و در صورت داشتن ایده جدید و یا ویژگی کاربردی در صورت ارائه، ارزیابی و لحاظ خواهد شد.
- در پیشنهاد ارسالی می بایست نحوه بهینه‌سازی و ارتقاء کیفیت سامانه در حین اجرای پروژه ذکر گردد.
- نیازمندی‌های سامانه در تعامل با دیگر سامانه‌ها باید در پیشنهاد ارسالی مشخص شود.
- در پیشنهاد ارسالی می‌بایست حتی‌الامکان از منابع موجود کشور در بخش سخت‌افزار و نرم‌افزار استفاده شود.

ویژگی‌های فنی

هدف اصلی این پروژه، ایجاد سامانه برچسب‌زن نقش‌های معنایی کلمات زبان فارسی می‌باشد. در این راستا پیشنهاد دهنده باید برنامه و استراتژی خود را در زمینه‌های زیر مشخص نماید:

- تهیه دادگان جامع از ساختارهای مختلف معنایی
- معماری سامانهٔ برچسب‌زن نقش معنایی
- ارائه سامانه به صورت منفک، وب‌سرویس و API برای استفاده در سامانه‌های دیگر
- قابلیت توسعه در آینده

در ادامه مشخصه‌ها و الزامات فنی مدنظر برای سامانهٔ برچسب‌زن نقش‌های معنایی مشخص شده است.

- الزامات فنی:
  - تهیه پیکرهٔ برچسب‌خورده با حجم حداقل ۲ میلیون کلمه
    - قاب افعال به ازای تمام معانی هر فعل فارسی به طور جداگانه مشخص شود و برای هر قاب حداقل ۲۰ نمونه جمله تهیه شده و برچسب‌گذاری گردد.
    - شیوه‌نامهٔ برچسب‌زنی با توافق کارفرما تهیه شود.
    - در انتخاب جملات پیکره تنوع در طول جملات، سبک نو شتاری، د سته‌های مو ضوعی مدنظر و ... قرار گیرد.
    - در انتخاب جملات پیکره باید تنوع در تجزیه وابستگی حتی‌الامکان وجود داشته باشد.

- تعیین نقش معنایی حتی‌الامکان با در نظر گرفتن معنی واژه‌ها انجام گیرد (به عنوان مثال اگر فعل جمله شکستن است کنش‌پذیر آن باید از جنس شکستنی باشد).

- حداقل دقت ۸۵٪ برای سامانه برچسب‌زن نقش‌های معنایی
- حداقل معیار کاپا معادل ۷۵٪ برای صحت برچسب‌دهی پیکره
- طراحی و پیاده‌سازی وب سرویس و API برای کلیه بخش‌های سیستم
- پوشش نیازهای غیرکارکردی نظیر سرعت پاسخ‌گویی، زمان پاسخ‌گویی به کاربران همزمان، سرویس‌دهی همزمان به کاربران، دسترس‌پذیری، تحمل‌پذیری در برابر خطا، واسط کاربری، امنیت سیستم در برابر نفوذهای احتمالی
- سازگاری با آخرین نسخه استانداردهای بین المللی
- طراحی ورودی و خروجی واسط‌های کاربری ابزار برچسب‌زن نقش معنایی، مطابق استاندارد و کاربرد

- الزامات نگهداری:
  - ارائه خدمات پشتیبانی رایگان به مدت دو سال
  - نگهداری از سامانه به صورت رایگان به مدت دو سال

- واسط کاربری:
  - طراحی واسط کاربری مناسب

مراحل اجرای پروژه

اجرای پروژه به مدت دوازده ماه در سه مرحله انجام می‌گردد.

- فاز اول: تهیهٔ شیوه‌نامه برچسب‌زنی، انتخاب افعال و انتخاب جملات دادگان، تعیین آرگومان افعال زبان فارسی
- فاز دوم: برچسب‌زنی جملات دادگان مطابق شیوه‌نامه
- فاز سوم: توسعهٔ ابزار برچسب‌زن نقش معنایی برای متون فارسی

خروجی‌های پروژه

- ابزار، وب‌سرویس و API مربوط به برچسب‌زن نقش‌های معنایی با دقت حداقل ۸۵٪ و شرایط ذکرشده در بند ویژگی‌های فنی.
- پیکره برچسب‌خورده با نقش‌های معنایی در حجم حداقل ۲ میلیون کلمه و با شرایط ذکرشده در بند ویژگی‌های فنی.
- ابزارها و داده‌های جانبی تهیه شده در حین انجام پروژه مانند:
  - لیست افعال انتخابی به همراه ساخت ظرفیت معنایی آنها

- واسط کاربری برای برچسب‌زنی جملات پیکره
- ابزار ارزیابی دادگان و برچسب‌زن نقش معنایی

- گزارش مربوط به سامانه برچسب‌زن نقش‌های معنایی زبان فارسی (مشتمل بر سیاق طراحی، معماری حاصله و اعتبارسنجی سامانه مربوطه)
- گزارش حاوی قاب‌های افعال زبان فارسی و ارتباط معنایی آنها مشتمل بر قواعد و دستورات برچسب‌گذاری معنایی
- گزارش حاوی شیوه‌نامه برچسب‌زنی معنایی متون فارسی
- گزارش بررسی ابزارهای موجود.

ویژگی محصولات و خروجی‌های پروژه

۱. نرم‌افزارهای تولیدشده حتی الامکان به platform خاصی وابسته نباشد.
۲. از متدولوژی‌های معتبر در تولید نرم‌افزار استفاده شود. متدولوژی‌های موردنظر در تولید نرم‌افزار، متدولوژی‌های XP و یا RUP هستند. کلیه مستندات باید در حین فرایند تولید نرم‌افزار و بر اساس دستورالعمل‌های متدولوژی تولید و تأیید گردند.
۳. مستندات تحویلی به صورت الکترونیکی در قالب‌های PDF و HTML و قابل ویرایش مانند TXT و DOC مطابق قالبی که توسط کارفرما ارائه خواهد شد، به صورت فارسی تهیه شود.
۴. سازگاری با آخرین نسخه استانداردهای بین‌المللی، ملی و مرسوم تا حد ممکن که در آخرین نسخه از نرم‌افزار استفاده شده‌اند یا قرار است در پروژه از آن استفاده شود مانند استاندارد یونی‌کد.
۵. واگذاری کلیه حقوق (غیر انحصاری) کُدها و مستندات تولیدشده توسط مجری به دارنده حق تکثیر نرم‌افزار اصلی و/یا هر نهاد یا سازمان دیگری که کارفرما لازم بداند و/یا انتشار آنها با مالکیت عمومی (public domain) در صورت لزوم

شاخص‌های ارزیابی دادگان

۱- معیار کاپا[۱]:

معیاری آماری است که در زبان‌شناسی پیکره‌ای برای بررسی میزان تفاهم میان دو برچسب‌گذاری بر روی پیکره به کار می‌رود. در این معیار تلاش شده است که برچسب‌گذاری‌هایی که به طور اتفاقی شبیه به یکدیگر شده‌اند تاثیری در نتیجه نداشته باشند. به این منظور کاپا به صورت فرمول زیر محاسبه می‌شود. در این فرمول Pr(a) میزان تشابه صحیح میان دو برچسب‌گذاری است و Pr(e) میزان تشابه اتفاقی میان دو برچسب‌گذاری است. برای استفاده از معیار کاپا در مواردی که تعداد برچسب‌گذارها بیشتر از ۲ نفر است، این مقدار به ازای هر دو برچسب‌گذار محاسبه شده و از مقادیر آن میانگین گرفته می‌شود.

$$k = \frac{\Pr(a) - \Pr(e)}{۱ - \Pr(e)}$$

برای تضمین یک‌دست بودن برچسب‌گذاری‌ها، باید بخش ثابتی از پیکره را تمامی افراد برچسب‌گذار، برچسب‌گذاری کنند و سپس میزان تفاوت میان این برچسب‌گذاری‌ها بر اساس معیار کاپا سنجیده شود (inter annotator agreement). همچنین باید بخشی از فعالیت هر برچسب‌گذار در طول زمان به‌صورت تصادفی انتخاب شود و به وسیله یک فرد خبره پایش و بررسی گردد (intra annotator agreement).

۲- پوشش یا تنوع (کمی و کیفی)

- حجم دادگان

مهمترین شاخص در تعیین حجم دادگان، تعداد جملات تشکیل‌دهندهٔ آن است. تعداد کلمات نیز می‌تواند شاخصی برای تعیین حجم دادگان باشد.

- تنوع سیاق، حوزه، ....
  - میانگین طول جملات در کل دادگان
  - تعداد جملات کوتاه، متوسط، و بلند پیکره
  - تعداد جملات کوتاه، متوسط، و بلند پیکره در هر یک از سیاق‌ها و حوزه‌ها

شاخص‌های ارزیابی ابزار برچسب‌زن نقش‌های معنایی

شاخص‌های کارکردی شامل میزان دقت ابزار برچسب‌زن و شاخص‌های غیرکارکردی نیز شامل سرعت برچسب‌زنی جملات، داشتن یک رابط کاربری مناسب، تحمل‌پذیری در برابر خطا و ... می‌باشد.

۱- دقت ابزار برچسب‌زن نقش معنایی کلمات

[۱] Kappa

میزان دقت ابزار برچسب‌زن نقش‌های معنایی بر اساس نسبت تعداد برچسب‌های صحیح به کل برچسب‌های جمله اندازه‌گیری می‌شود. برای اندازه‌گیری این معیار باید برچسب‌های حاصل از خروجی ابزار برچسب‌زن نقش‌های معنایی به ازای هر جمله، با برچسب‌های استاندارد جمله در دادگان آزمون مقایسه شود. دقت نهایی با میانگین‌گیری از دقت برچسب‌زنی کل جملات دادگان آزمون به دست می‌آید.

۲- معیارهای ارزیابی نیازهای غیرکارکردی

- دسترس‌پذیری
- کارایی
  - زمان پاسخ‌گویی
  - توانایی در پاسخ‌گویی به درخواست‌های همزمان
  - زمان پاسخ‌گویی به درخواست‌های همزمان
- تحمل‌پذیری در برابر خطا
- امنیت