گزارش نقشه راه پردازش خط و زبان فارسی

محمد بحرانی

مهر ۹۵

مقدمه

- پردازش زبان طبیعی (NLP) یکی از نیازهای عصر فناوری جهت استفاده بهینه از منابع اطلاعاتی است.
- به دلیل اهمیت حفظ و نگهداری از زبان و خط فارسی در محیط رایانه‌ای نیاز به فعالیت‌های حوزه پردازش زبان طبیعی بیش از پیش احساس می‌شود.
- به رغم تلاش‌های صورت گرفته برروی پردازش رایانه‌ای زبان فارسی، هنوز در این حوزه فاصله زیادی نسبت به زبان‌های دیگر (مانند انگلیسی) وجود دارد.

مقدمه

- دسته‌بندی کلی فعالیت‌ها در حوزه پردازش زبان طبیعی
  - پیکره‌ها و منابع زبانی
  - ابزارهای پایه پردازش زبان طبیعی
  - ابزارها و سامانه‌های کاربردی پردازش زبان طبیعی
    - سامانه‌های کاربردی پردازش متن
    - سامانه‌های کاربردی پردازش گفتار
    - سامانه‌های کاربردی پردازش تصاویر متنی
    - سامانه‌های ترکیبی

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- پیکره‌ها و منابع زبانی

نوع پیکره یا منبع زبانی	وضعیت موجود در زبان فارسی	وضعیت مطلوب
پیکره با برچسب موجودیت‌های نامدار	- پیکره واحدهای اسمی شرکت آرمان رایان شریف در حال توسعه به ۴ میلیون کلمه با ۶ نوع موجودیت اسمی - پیکره موسسه نور با ۵۰۰ هزار کلمه با ۳ نوع موجودیت اسمی	پیکره با برچسب موجودیت‌های نامدار در انواع مختلف با ۱۰ میلیون کلمه
پیکره برچسب‌خورده با عبارات هم‌مرجع	- پیکره PCAC-2008 با حدود ۱۰۰هزار کلمه (۳۱ متن) دارای برچسب مرجع برای حدود ۲۰۰۰ ضمیر - پیکره با حدود ۱۵۰۰۰ جمله در پژوهشگاه خواجه‌نصیر در حال تهیه است.	پیکره برچسب‌خورده با عبارات هم‌مرجع با حداقل ۳۰۰ هزار جمله از متن‌های مختلف
پیکره برچسب‌خورده با نقش‌های معنایی	- دادگان موسسه نور با ۳۰۰۰۰ جمله و حدود ۵۰۰ هزار کلمه	پیکره برچسب‌خورده با انواع نقش‌های معنایی از متون مختلف با حداقل ۱۰ میلیون کلمه
پیکره متنی محاوره‌ای	- بخشی از پیکره دکتر بیجن‌خان (حدود ۷ میلیون کلمه) - بخشی از پیکره irBlogs (تقریباً ۵۰۰ میلیون کلمه از وبلاگ‌های فارسی)	پیکره محاوره‌ای با ۲ میلیارد کلمه از ژانرها و منابع مختلف مختلف
پیکره تجزیه‌شده نحوی (treebank)	- درخت‌بانک نحوی شریف (۳۰۰۰۰ جمله، ۵۰۰ هزار کلمه) - پیکره PerTreebank (۱۰۲۸ جمله) - در دانشگاه تهران (با ۵۰۰ هزار کلمه) و پژوهشگاه خواجه‌نصیر (با ۱ میلیون کلمه) در حال تهیه است.	درخت‌بانک نحوی با ۵ میلیون کلمه و ۱۵۰ هزار جمله با موضوعات مختلف

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- ابزارهای پایه پردازش زبان طبیعی

نام ابزار یا سامانه	وضعیت موجود در زبان فارسی	وضعیت مطلوب
- واحدساز و یکسان‌ساز - ریشه‌یاب و لم‌یاب - برچسب‌زن مقوله نحوی	- جعبه ابزار مرکز تحقیقات مخابرات - جعبه ابزار آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی - جعبه ابزار شرکت آرمان رایان شریف - جعبه ابزار گروه سبحه	جعبه ابزار با دقت ۹۸٪ برای هر یک از ابزارها
تجزیه‌گر نحوی	- تجزیه‌گر وابستگی موسسه نور - تجزیه‌گر نحوی آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی (تجزیه‌گر سازه‌ای)	تجزیه‌گر نحوی سازه‌ای با دقت ۸۵٪ تجزیه‌گر نحوی وابستگی با دقت ۹۵٪
برچسب‌زن نقوش معنایی	کارهای پراکنده دانشگاهی با کیفیت متوسط و قابلیت کار برروی جملات ساده	سامانه برچسب‌زن نقش معنایی با دقت ۹۵٪ برروی جملات حوزه خبری و با دقت ۸۵٪ برروی جملات تخصصی
تجزیه‌گر معنایی	کارهای پراکنده و اندک دانشگاهی	تجزیه‌گر معنایی با قابلیت کار برروی متون حوزه‌های موردنظر و با دقت ۸۰٪

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- سامانه‌های کاربردی پردازش متن

نام ابزار یا سامانه	وضعیت موجود در زبان فارسی	وضعیت مطلوب
ترجمه ماشینی	- سامانه مترجم ماشینی ترگمان با معیار بلو معادل ۲۰ در ترجمه انگلیسی به فارسی و ۳۴ در ترجمه فارسی به انگلیسی برای متون خبری - سامانه مترجمه ماشینی فرازین با معیار بلو معادل ۲۴ در ترجمه انگلیسی به فارسی برای متون خبری - سامانه مترجم ماشینی پارس با قابلیت ترجمه انگلیسی به فارسی در حوزه عمومی و ۳۸ حوزه تخصصی	سامانه مترجم ماشینی با قابلیت ترجمه زبان‌های زنده دنیا به فارسی و بالعکس با معیار بلو معادل ۵۰ (یا حداقل معادل بلوی گوگل) برای متون عمومی و تخصصی
تشخیص و تصحیح خطاهای املایی و گرامری و معنایی	- سامانه ویرایشگر وفا، قابلیت تشخیص و پیشنهاد برای تصحیح خطاهای املایی و گرامری و معنایی - سامانه ویراستیار، قابلیت تشخیص و پیشنهاد برای تصحیح خطاهای املایی	سامانه ویرایشگر با قابلیت تشخیص و پیشنهاد برای تصحیح خطاهای املایی، دستوری و معنایی با دقت میانگین ۸۰٪ و قابل استفاده به صورت افزونه در واژه‌پردازهای رایج
پرسش و پاسخ خودکار	- سامانه پرسش و پاسخ قرآن‌جوی در حوزهٔ علوم قرآنی - کارهای پراکنده دانشگاهی با کیفیت متوسط و در حوزه‌های خاص	سامانه پرسش و پاسخ خودکار در حوزه‌های مختلف و مشابه با IBM Watson
فهم زبان طبیعی	کارهای پراکنده دانشگاهی با کیفیت متوسط و در حوزه‌های خاص	سامانه درک زبان طبیعی در حوزه‌های مختلف و با دقت ۹۰٪ (مشابه با Apple Siri)
بازیابی اطلاعات	- سامانه‌های بازیابی اطلاعات موجود در موتورهای جستجوی بومی پارسی‌جو، یوز و ...	سامانه بازیابی اطلاعات در حد موتور جستجوی گوگل
خلاصه‌سازی متن	- سامانه خلاصه‌ساز متنی ایجاز (آزمایشگاه فناوری وب دانشگاه فردوسی مشهد) با دقت ۴۵٪ به صورت استخراجی	- خلاصه‌ساز استخراجی با دقت ۸۵٪ - خلاصه‌ساز چکیده‌ای با دقت ۷۵٪

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- سامانه‌های کاربردی پردازش متن (ادامه)

نام ابزار یا سامانه	وضعیت موجود در زبان فارسی	وضعیت مطلوب
استخراج اطلاعات	کارهای اندک و پراکنده دانشگاهی با کیفیت پایین	سامانه استخراج اطلاعات، قابل کار برروی متون عمومی و تخصصی با کیفیت بالا و معیار F معادل ۸۵٪
تشخیص موجودیت‌های نامدار	- سامانه NER شرکت آرمان رایان شریف، با قابلیت ۶ موجودیت نامدار با دقت ۸۰٪ برروی متون خبری - سامانه NER موسسه نور، با قابلیت ۳ موجودیت نامدار با دقت ۸۰٪ برروی متون خبری	سامانه NER با قابلیت تشخیص انواع موجودیت‌های نامدار پرکاربرد با دقت ۹۵٪ برروی انواع متون
مرجع‌یابی ضمیر	کارهای پراکنده دانشگاهی با کیفیت متوسط برای گروهی از ضمایر (عمدتاً ضمایر متصل) و معمولاً به صورت نیمه‌خودکار	مرجع‌یاب خودکار ضمیر با دقت ۹۰٪ برای انواع ضمایر متصل و منفصل
رفع ابهام معنایی کلمات	کارهای پراکنده دانشگاهی با کیفیت متوسط و تعداد اندک کلمات هدف	سامانه رفع ابهام معنایی با دقت ۹۰٪ با حداقل ۲۰۰ کلمه هدف
تحلیل نظرات کاربران	کارهای پراکنده دانشگاهی با کیفیت متوسط و برای حوزه‌های خاص (مانند نظرات کاربران در مورد کالاها و محصولات)	سامانه تحلیل نظر کاربران با قابلیت کار برروی انواع متون و نظرات و با دقت ۸۰٪
تشابه‌یابی در متون	- سامانه مشابهت‌یابی مؤسسه نور - سامانه مشابهت‌یابی جهاد دانشگاهی	سامانه مشابهت‌یابی با دقت ۹۵٪ با قابلیت کار برروی انواع متون

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- سامانه‌های کاربردی پردازش گفتار

نام ابزار یا سامانه	وضعیت موجود در زبان فارسی	وضعیت مطلوب
بازشناسی گفتار پیوسته (میکروفونی و تلفنی)	- موتور و سرویس بازشناسی گفتار پیوسته نویسا با واژگان بزرگ ۱۳۰ هزار کلمه‌ای با دقت ۹۵٪ برای گفتار رسمی و میکروفون اختصاصی و ۹۰٪ با سایر میکروفون‌ها در محیط بدون نویز - موتور بازشناسی گفتار پیوسته شنوا با واژگان بزرگ ۶۵ هزار کلمه‌ای با دقت ۹۰٪ برای گفتار رسمی میکروفونی در محیط بدون نویز	- موتور و سرویس بازشناسی گفتار پیوسته با واژگان بزرگ، با دقت ۹۵٪ برای گفتار رسمی و محاوره‌ای با انواع میکروفون‌ها و ۸۵٪ برای گفتار تلفنی و بدون افت کارایی در محیط‌های نویزی
بازشناسی گفتار کلمات مجزا (میکروفونی و تلفنی)	موتور بازشناسی گفتار کلمات مجزای میکروفونی و تلفنی شرکت عصرگویش با دقت ۹۸٪ برروی واژگان ۴۰۰ کلمه‌ای در محیط بدون نویز	موتور بازشناسی گفتار کلمات مجزای میکروفونی و تلفنی شرکت عصرگویش با دقت ۹۸٪ برروی واژگان ۱۰۰ هزار کلمه‌ای و بدون افت کارایی در محیط‌های نویزی
تبدیل متن به گفتار	- موتور تبدیل متن به گفتار آریانا با واژگان ۱۰۰ هزار کلمه‌ای با معیار MOS معادل ۴.۵ و معیار DRT معادل ۹۰٪ برای خوشایندی و طبیعی بودن با صدای مختلف زن و مرد. دقت ۷۰٪ در خوانش کسره اضافه و دقت ۸۰٪ در خوانش همنگارها - موتور تبدیل متن به گفتار رسا محصول شرکت گاتا با معیار MOS معادل ۴.۵ و معیار DRT معادل ۹۰٪ برای خوشایندی و طبیعی بودن با صدای مختلف زن و مرد. دقت ۸۰٪ در خوانش کسره اضافه و همنگارها	موتور تبدیل متن به گفتار بدون محدودیت واژگان با خوشایندی و طبیعی بودن بالا (MOS معادل ۵ و DRT معادل ۹۵٪) با صداهای مختلف زن و مرد و خوانش کاملاً صحیح کسره اضافه و همنگارها
بازیابی صدا - بازیابی موسیقی - بازیابی گفتار (جستجو در گفتار)	کارهای پراکنده دانشگاهی با کیفیت پایین در زمینه بازیابی موسیقی و بازیابی مستندات گفتاری	- سامانه بازیابی مستندات گفتاری با معیار F معادل ۸۰٪ به صورت query by keyword - سامانه بازیابی موسیقی با معیار F معادل ۹۰٪ به صورت query by humming و query by example
واژه‌یابی در گفتار	- سامانه واژه‌یاب شرکت عصرگویش‌پرداز با دقت ۸۰٪ برروی ۱۰ کلمه کلیدی - سامانه جویا مربوط به پژوهشگاه خواجه‌نصیر	سامانه واژه‌یابی در گفتار با دقت ۹۵٪ با حداقل ۵۰ کلمه کلیدی

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- سامانه‌های کاربردی پردازش تصاویر متنی

نام ابزار یا سامانه	وضعیت موجود در زبان فارسی	وضعیت مطلوب
بازشناسی نویسه‌های نوری	- نویسه‌خوان نوری پرشیانگار با دقت ۹۵٪ برای فونت‌های رایج - نویسه‌خوان نوری آراکس با دقت ۹۵٪ برای فونت‌های رایج - نویسه‌خوان نوری گوگل	نویسه‌خوان نوری بومی با دقت ۹۸٪ برای انواع فونت‌ها و تصاویر با رزولوشن پایین
تشخیص دست‌نوشته	- کارهای پراکنده دانشگاهی در زمینه تشخیص دست‌نوشته آنلاین و آفلاین با کیفیت متوسط در تست‌های آزمایشگاهی - سیستم تشخیص دست‌نوشته آنلاین سامسونگ با دقت بالا دارای کاربرد در گوشی‌های هوشمند	- سیستم تشخیص دست‌نوشته آنلاین با دقت ۹۸٪ - سیستم تشخیص دست‌نوشته آفلاین با دقت ۹۰٪

اولویت‌بندی فعالیت‌ها (ارزیابی کیفی)

- پیکره ها و منابع زبانی

	وضعیت موجود	پیچیدگی	کسب و کار	پیش نیاز	محصولات منتج
پیکره تجزیه‌شده نحوی (treebank)	در حال توسعه در طرح جویشگر	بالا	پایین	- پیکره خام - دادگان وابستگی	تجزیه‌گر نحوی
پیکره متنی محاوره‌ای	نامطلوب	بالا	پایین		بازشناسی گفتار محاوره ای
پیکره برچسب‌خورده با نقش‌های معنایی	نیمه مطلوب	بالا	پایین	پیکره خام	برچسب زن نقوش معنایی
پیکره برچسب خورده با موجودیت‌های نامدار	در حال توسعه در طرح جویشگر و سایر طرح ها	بالا	پایین	پیکره خام	شناسایی موجودیت های نامدار
پیکره برچسب خورده با عبارات هم‌مرجع	در حال توسعه در طرح جویشگر و سایر طرح ها	متوسط	پایین	پیکره خام	مرجع‌یابی ضمیر

اولویت‌بندی فعالیت‌ها (ارزیابی کیفی)

- ابزارهای پایه پردازش زبان طبیعی

	وضعیت موجود	پیچیدگی	کسب و کار	پیش نیاز	محصولات منتج
واحدساز/ یکسان ساز	نیمه مطلوب	پایین	پایین		کلیه پردازش های متنی
ریشه‌یاب / لم‌یاب	نیمه مطلوب	متوسط	پایین	واژگان	اکثر پردازش های متنی
برچسب‌زن مقوله نحوی	مطلوب	پایین	پایین	پیکره متنی با برچسب مقوله نحوی	- تجزیه گر نحوی - ترجمه ماشینی- تشخیص و تصحیح خطای املایی و گرامری و معنایی- پرسش و پاسخ خودکار- خلاصه سازی متن- استخراج اطلاعات- تشخیص موجودیت های نامدار- مرجع یابی ضمیر- ...
تجزیه‌گر نحوی	در حال توسعه در طرح جویشگر	بالا	پایین	- پیکره تجزیه‌شده نحوی با دستور زایشی - برچسب زن مقوله نحوی	- ترجمه ماشینی- تشخیص و تصحیح خطای املایی و گرامری و معنایی- پرسش و پاسخ خودکار- خلاصه‌سازی متن- تشخیص موجودیت های نامدار- مرجع یابی ضمیر- استخراج اطلاعات- تحلیل نظرات کاربران - ...
برچسب‌زن نقوش معنایی	نامطلوب	متوسط	پایین	پیکره برچسب‌خورده با نقش‌های معنایی	- تجزیه گر معنایی - فهم زبان طبیعی - استخراج اطلاعات
تجزیه‌گر معنایی	نامطلوب	بالا	پایین	- پیکره تجزیه‌شده معنایی - برچسب‌زن نقوش معنایی	- فهم زبان طبیعی - استخراج اطلاعات

اولویت‌بندی فعالیت‌ها (ارزیابی کیفی)

- ابزارها و سامانه‌های کاربردی پردازش زبان طبیعی

	وضعیت موجود	پیچیدگی	کسب و کار	پیش نیاز	محصولات منتج
تشخیص موجودیت‌های نامدار	نامطلوب/در حال توسعه در طرح جویشگر	متوسط	پایین	- پیکره برچسب خورده با موجودیت‌های نامدار - ابزارهای پایه (واحدساز، یکسان‌ساز، چانکر، لم یاب، برچسب زن مقوله نحوی)	- استخراج اطلاعات - ترجمه ماشینی - پرسش و پاسخ خودکار - بازیابی اطلاعات
ترجمه ماشینی	نیمه مطلوب/ در حال توسعه در طرح جویشگر	بالا	بالا	- پیکره موازی دوزبانه انگلیسی-فارسی - پیکره موازی چندزبانه - ابزارهای پایه	ترجمه گفتار به گفتار
تشخیص و تصحیح خطای املایی و گرامری و معنایی	نیمه مطلوب	بالا	بالا	- ابزارهای پایه - واژگان	
تحلیل نظرات کاربران	نامطلوب	متوسط	متوسط	- پیکره متنی تحلیل احساس - ابزارهای پایه	
بازشناسی گفتار پیوسته (میکروفونی و تلفنی)	نیمه مطلوب / در حال توسعه در سایر طرح ها	بالا	بالا	- دادگان گفتاری - پیکره متنی خام - واژگان	- سامانه محاوره متنی/گفتاری - ترجمه گفتار به گفتار - جستجوی گفتاری - بازیابی گفتار

اولویت‌بندی فعالیت‌ها (ارزیابی کیفی)

- ابزارها و سامانه‌های کاربردی پردازش زبان طبیعی (ادامه)

	وضعیت موجود	پیچیدگی	کسب و کار	پیش نیاز	محصولات منتج
بازیابی گفتار (جستجو در گفتار)	نامطلوب	بالا	بالا	- بازشناسی گفتار پیوسته - بازیابی اطلاعات	
بازشناسی نویسه‌های نوری	نیمه مطلوب	بالا	بالا	دادگان متون چاپی	تبدیل تصاویر متنی به گفتار
مرجع‌یابی ضمیر	نامطلوب/در حال توسعه در طرح جویشگر	بالا	پایین	- پیکره برچسب خورده با عبارات هم‌مرجع - ابزارهای پایه (واحدساز، یکسان‌ساز، چانکر، لم یاب، برچسب زن مقوله نحوی)	- استخراج اطلاعات - ترجمه ماشینی - پرسش و پاسخ خودکار - خلاصه سازی متن
تشابه‌یابی در متون	نیمه مطلوب	بالا	متوسط/بالا	- ابزارهای پایه - بازیابی اطلاعات	
تشخیص دست‌نوشته	نامطلوب	بالا	بالا	دادگان متون دست‌نوشته	تبدیل تصاویر متنی به گفتار
سامانه محاوره گفتاری/متنی	نامطلوب	بالا	بالا	- بازشناسی گفتار پیوسته - فهم زبان طبیعی - تولیدکننده زبان طبیعی - تبدیل متن به گفتار	

اولویت‌بندی فعالیت‌ها (ارزیابی کمّی)

- کمّی‌سازی آیتم‌های جدول
  - آیتم «وضعیت موجود»
    - مطلوب/در حال توسعه: ۱
    - نیمه مطلوب: ۲
    - نامطلوب: ۳
  - آیتم «پیچیدگی»
    - بالا: ۱
    - متوسط: ۲
    - پایین: ۳

اولویت‌بندی فعالیت‌ها (ارزیابی کمّی)

- کمّی‌سازی آیتم‌های جدول
  - آیتم «کسب و کار»
    - پایین: ۱
    - متوسط: ۲
    - بالا: ۳
  - آیتم «پیش‌نیاز»: مجموع امتیازهای آیتم «وضعیت موجود» هر پیش‌نیاز (با علامت منفی)
  - آیتم «محصولات»: تعداد محصولات منتج از فعالیت

اولویت‌بندی فعالیت‌ها (ارزیابی کمّی)

- ضرایب وزنی هر آیتم
  - وضعیت موجود: ۲
  - پیچیدگی: ۱
  - کسب و کار: ۳
  - پیش‌نیاز: ۱-
  - محصولات: ۱

اولویت‌بندی فعالیت‌ها (نتایج)

- ده فعالیت با امتیاز بالاتر (امتیاز ۱۲ به بالا)
  - *RFP* برای موارد قرمزرنگ در حال تهیه است.

واحدساز/ یکسان‌ساز	۳۰
ریشه‌یاب	۲۳
برچسب‌زن مقوله نحوی	۱۵
تجزیه‌گر نحوی	۱۴
بازشناسی نویسه‌های نوری	۱۴
تشخیص دست‌نوشته	۱۴
بازیابی گفتار	۱۴
واژه‌یابی در گفتار	۱۴
ترجمه گفتار به گفتار	۱۳
برچسب‌زن نقوش معنایی	۱۲
بازشناسی گفتار کلمات مجزا	۱۲

با سپاس

؟