گزارش نقشه راه پردازش خط و زبان فارسی

محمد بحرانی

مردادماه ۹۵

عناوین مطالب

- مقدمه
- لیست فعالیت‌های حوزه پردازش زبان طبیعی
- وضعیت موجود و وضعیت مطلوب فعالیت‌ها در زبان فارسی
- اولویت‌بندی انجام فعالیت‌ها برای زبان فارسی

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

لیست فعالیت‌های حوزه پردازش زبان طبیعی

- پیکره‌ها و منابع زبانی
  - پیکره‌ها و دادگان‌های متنی
  - دادگان‌های گفتاری
  - دادگان‌های تصاویر متنی
  - واژگان‌ها و هستان‌شناسی‌ها

لیست فعالیت‌های حوزه پردازش زبان طبیعی

- پیکره‌ها و دادگان‌های متنی
  - پیکره متنی خام
  - پیکره متنی با برچسب مقوله نحوی
  - پیکره متنی محاوره‌ای
  - پیکره موازی محاوره‌ای - رسمی
  - پیکره تجزیه‌شده نحوی (treebank) با دستور سازه‌ای
  - پیکره تجزیه‌شده نحوی (treebank) با دستور وابستگی
  - پیکره برچسب‌خورده با نقش‌های معنایی
  - پیکره برچسب‌خورده با عبارات هم‌مرجع
  - پیکره موازی دو یا چندزبانه
  - پیکره متنی تحلیل احساس
  - پیکره با برچسب موجودیت‌های نامدار

لیست فعالیت‌های حوزه پردازش زبان طبیعی

- دادگان‌های گفتاری
  - دادگان گفتاری رسمی-میکروفونی
  - دادگان گفتاری- تلفنی
  - دادگان گفتار محاوره‌ای- میکروفونی
  - دادگان گفتار احساسی
- دادگان‌های تصاویر متنی
  - دادگان متون چاپی برای OCR
  - دادگان متون دست‌نوشته
- هستان‌شناسی‌ها و فرهنگ‌ها
  - وردنت
  - گراف دانش
  - فرهنگ ظرفیت نحوی و معنایی افعال

لیست فعالیت‌های حوزه پردازش زبان طبیعی

- ابزارهای پایه پردازش زبان طبیعی
  - واحدساز (Tokenizer)
  - یکسان‌ساز (Normalizer)
  - ریشه‌یاب (Stemmer)
  - لم‌یاب (Lemmatizer)
  - تحلیل‌گر صرفی (Morphological Analyzer)
  - برچسب‌زن مقوله نحوی (POS Tagger)
  - تجزیه‌گر نحوی (Syntactic Parser)
  - برچسب‌زن نقوش معنایی (Semantic Role Labeler)
  - تجزیه‌گر معنایی (Semantic Parser)

لیست فعالیت‌های حوزه 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

لیست فعالیت‌های حوزه پردازش زبان طبیعی

- سامانه‌های کاربردی پردازش متن
  - ترجمه ماشینی
  - تشخیص و تصحیح خطای املایی و گرامری
  - پرسش و پاسخ خودکار
  - فهم زبان طبیعی
  - بازیابی اطلاعات
  - خلاصه‌سازی متن
  - استخراج اطلاعات
  - تشخیص موجودیت‌های نامدار
  - مرجع‌یابی ضمیر
  - رفع ابهام معنایی کلمات

لیست فعالیت‌های حوزه پردازش زبان طبیعی

- سامانه‌های کاربردی پردازش متن (ادامه)
  - تحلیل نظرات کاربران
  - تشابه‌یابی در متون
  - تشخیص زبان نوشتار
  - قطعه‌بند موضوعی متن
  - ساده‌ساز متن
  - بسط و پیشنهاد عبارت جستجو
  - تولید زبان طبیعی
  - دسته‌بندی و تشخیص موضوع مستندات متنی
  - استخراج کلمات کلیدی از متن
  - تولید خودکار علائم نقطه‌گذاری در متن

لیست فعالیت‌های حوزه پردازش زبان طبیعی

- سامانه‌های کاربردی پردازش گفتار
  - بازشناسی گفتار پیوسته (میکروفونی و تلفنی)
  - بازشناسی گفتار کلمات مجزا (میکروفونی و تلفنی)
  - تبدیل متن به گفتار
  - بازیابی صدا
  - واژه‌یابی در گفتار

لیست فعالیت‌های حوزه پردازش زبان طبیعی

- سامانه‌های کاربردی پردازش تصاویر متنی
  - بازشناسی نویسه‌های نوری (OCR)
  - تشخیص دست‌نوشته (آن‌لاین/آفلاین)
- سامانه‌های ترکیبی
  - ترجمه گفتار به گفتار
  - سامانه محاوره گفتاری/متنی
  - تبدیل تصاویر متنی به گفتار

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- پیکره‌ها و منابع زبانی : پیکره‌ها و دادگان‌های متنی

نوع پیکره یا منبع زبانی	وضعیت موجود در زبان فارسی	وضعیت مطلوب
پیکره متنی خام	- پیکره متنی زبان فارسی (پیکره دکتر بیجن‌خان) با ۱۰۰ میلیون کلمه از منابع مختلف - پیکره همشهری با ۱۷۰ میلیون کلمه از متون خبری - پیکره ایسنا با ۵۵۰ میلیون کلمه از متون خبری	- پیکره متنی خام با ۱۰ میلیارد کلمه از ژانرها و منابع مختلف
پیکره متنی با برچسب مقوله نحوی	- پیکره متنی زبان فارسی (پیکره دکتر بیجن‌خان) با ۱۰ میلیون کلمه برچسب‌خورده	پیکره برچسب‌خورده با ۱۰۰ میلیون کلمه از ژانرها و منابع مختلف
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
پیکره موازی محاوره‌ای - رسمی	- در برنامه آینده موسسه نور می‌باشد.	پیکره موازی محاوره‌ای-رسمی با حداقل ۵۰ میلیون کلمه
پیکره تجزیه‌شده نحوی با دستور سازه‌ای	- 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
پیکره تجزیه‌شده نحوی با دستور وابستگی	- دادگان PerDT (۳۰۰۰۰ جمله، ۵۰۰ هزار کلمه) - دادگان اوپسالا (۶۰۰۰ جمله، ۱۵۰ هزار کلمه) - دادگان DepPerTreebank (۱۰۲۸ جمله)	درخت‌بانک نحوی با ۵ میلیون کلمه و ۱۵۰ هزار جمله با موضوعات مختلف

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- پیکره‌ها و منابع زبانی : پیکره‌ها و دادگان‌های متنی (ادامه)

نوع پیکره یا منبع زبانی	وضعیت موجود در زبان فارسی	وضعیت مطلوب
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 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 موازی دو یا چندزبانه	- پیکره دوزبانه فارسی-انگلیسی امیرکبیر (AFEC) با ۳۸ میلیون کلمه - پیکره موازی انگلیسی-فارسی تهران (TEP) با ۶۰۰ هزار جمله از زیرنویس فیلم‌ها - پیکره موازی انگلیسی-فارسی میزان با ۱ میلیون جمله در حوزه ادبیات کلاسیک	- پیکره موازی فارسی- انگلیسی با حداقل ۲۰۰ میلیون کلمه از حوزه‌های مختلف - پیکره موازی فارسی و حداقل ۱۰ زبان زنده دنیا باحداقل ۱۰۰ میلیون کلمه
پیکره متنی تحلیل احساس	- پیکره سنتی‌پرس (SentiPers) با ۱۱۰۰ جمله برچسب‌خورده، در حوزه نظرات در مورد کالاهای دیجیتال	پیکره با برچسب‌های حاوی بار احساسی با حداقل ۵۰ هزار جمله در حوزه‌های مختلف
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

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- پیکره‌ها و منابع زبانی : دادگان‌های گفتاری

نوع پیکره	وضعیت موجود در زبان فارسی	وضعیت مطلوب
دادگان گفتاری رسمی-میکروفونی	- دادگان فارس‌دات بزرگ، با حدود ۱۴۰ ساعت گفتار میکروفونی تقطیع‌شده از ۱۰۰ گوینده با برچسب واجی - دادگان فارس‌دات با حدود ۵ ساعت گفتار میکروفونی تقطیع‌شده از ۳۰۰ گوینده با برچسب واجی و آوایی	دادگان گفتار رسمی میکروفونی با حداقل ۱۰۰۰ ساعت گفتار از حداقل ۵۰۰ گوینده و با میکروفون‌های متنوع در شرایط محیطی مختلف
دادگان گفتاری- تلفنی	- دادگان فارس‌دات بزرگ تلفنی با ۷۰ ساعت محاورهٔ دوطرفه تلفنی که بخشی از آن تقطیع و برچسب‌دهی شده است. - دادگان فارس‌دات تلفنی با حدود ۵ ساعت گفتار محاوره‌ای یک‌طرفه تلفنی با تقطیع و برچسب‌دهی - دادگان تلفنی شرکت عصرگویش‌پرداز با حدود ۷ ساعت گفتار رسمی تلفنی	- دادگان تلفنی محاوره‌ای یک‌طرفه با حجم حداقل ۵۰۰ ساعت با تنوع گوشی تلفن و موبایل و از حدود ۲۰۰ گوینده مختلف - دادگان تلفنی از گفتار رسمی با حجم حداقل ۵۰۰ ساعت با تنوع گوشی تلفن و موبایل و از حدود ۲۰۰ گوینده مختلف
دادگان گفتار محاوره‌ای- میکروفونی	کارهای پراکنده و اندک دانشگاهی	دادگان گفتار محاوره‌ای میکروفونی با حداقل ۵۰۰ ساعت گفتار از حداقل ۲۰۰ گوینده و با میکروفون‌های متنوع در شرایط محیطی مختلف
دادگان گفتار احساسی	- پایگاه داده گفتار احساسی Persian ESD با حدود ۵۰۰ جمله در ۵ حالت احساسی - دادگان گفتار احساسی سهند با ۵۰ جمله در ۵ حالت احساسی	دادگان گفتار احساسی با حداقل ۵۰۰۰ جمله در انواع حالات احساسی

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- پیکره‌ها و منابع زبانی : دادگان‌های تصاویر متنی

نوع پیکره	وضعیت موجود در زبان فارسی	وضعیت مطلوب
دادگان متون چاپی برای OCR	دادگان farsiocr با ۲۰۰ صفحه اسکن شده با رزولوشن ۳۰۰ نقطه بر اینچ	دادگان متون چاپی با حداقل ۱۰۰۰۰ صفحه از فونت‌های رایج
دادگان متون دست‌نوشته	- بانک اطلاعات حروف گسسته دست‌نویس فارسی با ۱۰ میلیون تصویر از حروف گسسته - مجموعه ارقام دستنویس هدی با حدود ۱۰۰ هزار تصویر دست‌نوشته	دادگان متون دست‌نوشته با حداقل ۱۰۰۰۰ صفحه دست‌نوشته از حداقل ۱۰۰۰ نویسنده

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- پیکره‌ها و منابع زبانی : هستان‌شناسی‌ها و فرهنگ‌ها

نوع پیکره	وضعیت موجود در زبان فارسی	وضعیت مطلوب
وردنت	- فارس‌نت ورژن ۲ با ۳۰ هزار مدخل واژگانی و با دقت ۹۰٪ - فارس‌نت ورژن ۳ با ۱۰۰ هزار مدخل واژگانی در حال توسعه است. - وردنت فارسی حوزه فاوا با ۳۰ هزار مدخل واژگانی	وردنت فارسی در حوزه عمومی با حداقل ۱۰۰ هزار مدخل واژگانی با دقت ۹۵٪ و شامل انواع روابط
گراف دانش	یک گراف دانش اولیه با ۲۰۰ هزار موجودیت نامدار و ۵۰۰ هزار رابطه و بدون دارا بودن هستان‌شناسی در حال توسعه است.	گراف دانش با ۲۰۰ هزار موجودیت نامدار دارای ویژگی‌های هستان‌شناسی
فرهنگ ظرفیت نحوی و معنایی افعال	فرهنگ ظرفیت نحوی و معنایی افعال موسسه نور با ۴۵۰۰ فعل ساده و مرکب به همراه ظرفیت نحوی آنها	فرهنگ ظرفیت نحوی و معنایی افعال موسسه نور با حداقل ۱۰۰۰۰ فعل ساده و مرکب به همراه ظرفیت نحوی آنها

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- ابزارهای پایه پردازش زبان طبیعی

نام ابزار یا سامانه	وضعیت موجود در زبان فارسی	وضعیت مطلوب
واحدساز	- جعبه ابزار مرکز تحقیقات مخابرات - جعبه ابزار آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی - جعبه ابزار شرکت آرمان رایان شریف - جعبه ابزار گروه سبحه	واحدساز با دقت ٪۹۸
یکسان‌ساز	- جعبه ابزار مرکز تحقیقات مخابرات - جعبه ابزار آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی - جعبه ابزار شرکت آرمان رایان شریف - جعبه ابزار گروه سبحه	یکسان‌ساز با دقت ٪۹۸
- ریشه‌یاب - لم‌یاب - تحلیل‌گر صرفی	- جعبه ابزار مرکز تحقیقات مخابرات (ریشه‌یاب) - جعبه ابزار آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی (ریشه‌یاب، تحلیل‌گر صرفی) - پارس‌مورف، تحلیل‌گر صرفی دانشگاه شریف	- ریشه‌یاب با دقت ٪۹۸ - لم‌یاب با دقت ٪۹۸ - تحلیل‌گر صرفی با دقت ٪۹۵
برچسب‌زن مقوله نحوی	- جعبه ابزار گروه سبحه - جعبه‌ابزار مرکز تحقیقات مخابرات - جعبه ابزار شرکت آرمان رایان شریف - جعبه ابزار آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی	برچسب‌زن مقوله نحوی با دقت ٪۹۸

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- ابزارهای پایه پردازش زبان طبیعی (ادامه)

نام ابزار یا سامانه	وضعیت موجود در زبان فارسی	وضعیت مطلوب
تجزیه‌گر نحوی	- تجزیه‌گر وابستگی موسسه نور - تجزیه‌گر نحوی دانشکده برق و کامپیوتر دانشگاه تهران (تجزیه‌گر سازه‌ای) - 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 خطای املایی و گرامری	- سامانه ویرایشگر وفا، قابلیت تشخیص و پیشنهاد برای تصحیح خطاهای املایی و گرامری - 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 با کیفیت بالا
فهم زبان طبیعی	کارهای پراکنده دانشگاهی با کیفیت متوسط و در حوزه‌های خاص	سامانه درک زبان طبیعی در حوزه‌های مختلف و با دقت ۹۰٪
بازیابی اطلاعات	- سامانه‌های بازیابی اطلاعات موجود در موتورهای جستجوی بومی پارسی‌جو، یوز و ...	سامانه بازیابی اطلاعات با معیار F معادل ۸۵٪
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 متن (ادامه)

نام ابزار یا سامانه	وضعیت موجود در زبان فارسی	وضعیت مطلوب
قطعه‌بند موضوعی متن	کارهای بسیار اندک دانشگاهی برروی متونی از حوزه‌های خاص	سامانه قطعه‌بند موضوعی متن با قابلیت کار برروی انواع متون و با دقت ۹۰٪
ساده‌ساز متن	کارهای بسیار اندک دانشگاهی برروی متونی از حوزه‌های خاص	سامانه ساده‌ساز متن با قابلیت کار برروی انواع متون و با دقت ۸۵٪
بسط و پیشنهاد عبارت جستجو	سامانه‌های موجود در موتورهای جستجوی بومی	سامانه بسط و پیشنهاد با کیفیت مشابه گوگل
تولید زبان طبیعی	کارهای بسیار اندک دانشگاهی در حوزه‌های خاص	سامانه تولید زبان طبیعی با قابلیت تولید متون طبیعی و روان در زمینه‌های مختلف
دسته‌بندی و تشخیص موضوع مستندات متنی	کارهای پراکنده دانشگاهی با کیفیت متوسط برروی حدوداً ۱۰ – ۲۰ موضوع	سامانه دسته‌بندی و تشخیص موضوع متون با دقت ۹۰٪ برروی حداقل ۱۰۰ موضوع مختلف
استخراج کلمات کلیدی از متن	کارهای پراکنده و اندک دانشگاهی با کیفیت متوسط و برروی حوزه‌های خاص	سامانه استخراج کلمات کلیدی از متون حوزه‌های مختلف با معیار F معادل ۸۵٪
تولید خودکار علائم نقطه‌گذاری در متن	کارهای پراکنده و اندک دانشگاهی با کیفیت متوسط	سامانه تولید خودکار انواع علائم نقطه‌گذاری برای متون حوزه‌های مختلف و با معیار F معادل ۹۵٪

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	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٪

وضعیت موجود و وضعیت مطلوب فعالیت‌ها

- سامانه‌های ترکیبی

نام ابزار یا سامانه	وضعیت موجود در زبان فارسی	وضعیت مطلوب
ترجمه گفتار به گفتار	- مترجم هوشمند همراه شرکت سروش مهر با قابلیت ترجمه ۲۰۰۰ جمله از فارسی به ۹ زبان مختلف و بالعکس - مترجم پارسیا محصول شرکت عصرگویش‌پرداز با قابلیت ترجمه ۵۰۰ جمله از فارسی به انگلیسی و عربی (هر دو محصول ترجمه را بر اساس translation memory انجام می‌دهند).	سامانه ترجمه گفتار به گفتار با قابلیت ترجمه جملات مختلف زبان‌های گوناگون به فارسی و بالعکس
سامانه محاوره گفتاری/متنی	- سامانه دیالوگ گفتاری نیوشا محصول شرکت عصرگویش‌پرداز برای کاربردهای تلفن‌گویا و تلفن‌بانک - کارهای پراکنده دانشگاهی با کیفیت متوسط و خاص‌منظوره	سامانه محاوره گفتاری مشابه با Apple SIRI
تبدیل تصاویر متنی به گفتار	- یک سامانه اولیه با قابلیت خواندن یک فونت خاص در دانشگاه تهران در حال توسعه است.	سامانه تبدیل تصاویر متنی به گفتار مشابه با MIT finger reader

اولویت‌بندی فعالیت‌ها: پیکره‌ها و منابع زبانی

- اولویت ۱:
  - پیکره تجزیه‌شده نحوی (treebank) با دستور سازه‌ای در حجم ۱ میلیون کلمه
  - پیکره متنی محاوره‌ای در حجم ۵۰۰ میلیون کلمه از منابع مختلف
  - پیکره موازی محاوره‌ای- رسمی با ۱۰ میلیون کلمه
  - دادگان گفتار محاوره‌ای با حجم ۱۰۰ ساعت گفتار از منابع مختلف
  - دادگان متون دست‌نوشته با ۱۰۰۰ صفحه
  - گراف دانش

اولویت‌بندی فعالیت‌ها: پیکره‌ها و منابع زبانی

- اولویت ۲:
  - ادامه فعالیت‌های مذکور در اولویت ۱
  - پیکره موازی دوزبانه انگلیسی-فارسی در حجم ۱۰۰ میلیون کلمه
  - پیکره موازی چندزبانه در حجم ۴۰ میلیون کلمه و با ۵ زبان زنده دنیا (به غیر از انگلیسی)
  - پیکره برچسب‌خورده با نقش‌های معنایی در حجم ۲ میلیون کلمه
  - پیکره با برچسب موجودیت‌های نامدار با ۵ میلیون کلمه
  - پیکره برچسب‌خورده با عبارات هم‌مرجع با ۲.۵ میلیون کلمه
  - وردنت با ۱۰۰ هزار مدخل واژگانی
  - دادگان متون چاپی برای OCR با ۲۰۰۰ صفحه از ۵ فونت مختلف

اولویت‌بندی فعالیت‌ها: پیکره‌ها و منابع زبانی

- اولویت ۳:
  - پیکره تجزیه‌شده نحوی با دستور وابستگی با ۵ میلیون کلمه
  - پیکره متنی خام
  - پیکره متنی با برچسب مقوله نحوی
  - پیکره تحلیل احساس با ۲۰۰۰۰ جمله

اولویت‌بندی فعالیت‌ها: ابزارهای پایه

- اولویت ۱:
  - برچسب‌زن نقوش معنایی با دقت ۸۰٪
  - تجزیه‌گر معنایی برای حوزه‌های خاص با دقت ۶۵٪
- اولویت ۲:
  - ادامه فعالیت‌های اولویت ۱
  - تجزیه‌گر نحوی سازه‌ای با دقت ۷۵٪
  - ریشه‌یاب با دقت ۹۵٪
- اولویت ۳:
  - ادامه فعالیت‌های اولویت ۲
  - واحدساز
  - یکسان‌ساز
  - برچسب‌زن مقوله نحوی

اولویت‌بندی فعالیت‌ها: سامانه‌های کاربردی

- اولویت ۱:
  - سامانه محاوره گفتاری/متنی برای چند حوزه خاص با دقت ۸۰٪ در فهم زبان طبیعی
  - سامانه استخراج اطلاعات و استخراج روابط از متن
  - سامانه خودکار مرجع‌یاب ضمیر با دقت ۷۵٪
  - سامانه تشخیص موجودیت‌های نامدار با دقت ۸۰٪ برای ۶ نوع موجودیت
  - سامانه بازیابی صدا برای مستندات گفتاری با دقت ۶۵٪
  - سامانه بازشناسی گفتار پیوسته محاوره‌ای با دقت ۸۵٪
  - سامانه تشخیص دست‌نوشته آنلاین/آفلاین با دقت حدود ۸۰٪

اولویت‌بندی فعالیت‌ها: سامانه‌های کاربردی

- اولویت ۲:
  - ادامه فعالیت‌های اولویت ۱
  - سامانه مترجم ماشینی برای ترجمه زبان‌های مختلف به فارسی و برعکس
  - توسعه سامانه پرسش و پاسخ خودکار برای کاربرد در حوزه عمومی
  - سامانه خلاصه‌ساز متن به صورت استخراجی و با دقت ۶۵٪
  - توسعه سامانه تشخیص و تصحیح خطای املایی و گرامری
  - توسعه سامانه‌های تشابه‌یاب متون
  - سامانه بازشناسی گفتار کلمات مجزا برای ۵۰۰۰ کلمه مجزا با دقت ۹۵٪
  - سامانه واژه‌یابی در گفتار با دقت ۸۰٪ برروی ۲۵ کلمه کلیدی
  - افزایش کیفیت سامانه تبدیل متن به گفتار (TTS) فارسی
  - توسعه سامانه بازیابی اطلاعات جویشگرهای بومی
  - توسعه سامانه بسط و پیشنهاد عبارت جستجو در جویشگرهای بومی
  - توسعه سامانه‌های OCRبرای فونت‌های مختلف

اولویت‌بندی فعالیت‌ها: سامانه‌های کاربردی

- اولویت ۳:
  - ادامه فعالیت‌های اولویت ۲
  - سامانه تحلیل نظرات کاربران در چند حوزه کاربردی
  - سامانه تولید خودکار علائم نقطه‌گذاری در متن
  - سامانه خلاصه‌ساز متن به صورت چکیده‌ای با دقت ۵۵٪
  - سامانه استخراج کلمات کلیدی از متن
  - سامانه تولید زبان طبیعی
  - سامانه ساده‌ساز متن
  - سامانه تشخیص زبان نوشتار
  - سامانه قطعه‌بند موضوعی متن
  - سامانه دسته‌بندی و تشخیص موضوع مستندات متنی
  - سامانه ترجمه گفتار به گفتار با قابلیت ترجمه جملات انگلیسی به فارسی و برعکس
  - سامانه تبدیل تصاویر متنی به گفتار

با سپاس

؟