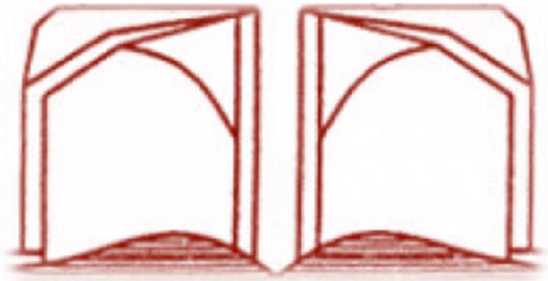


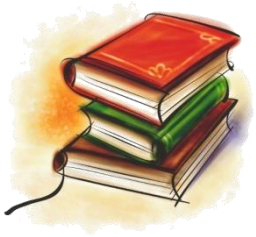
آشنایی با زبان‌شناسی رایانشی

مقدمه و معرفی

هادی ویسی

h.veisi@ut.ac.ir

دانشگاه تهران – دانشکده علوم و فنون نوین



فهرست

- معرفی
- کاربردهای زبان‌شناسی رایانشی: متنی، گفتاری، تصویری و ترکیبی
  - غلط‌یاب املایی و غلط‌یاب گرامری
  - بازیابی متن
  - خلاصه‌سازی
  - عقیده کاوی
  - سیستم‌های پرسش و پاسخ
  - ترجمه ماشینی
  - بازشناسی (تشخیص) گفتار
  - سنتز گفتار
  - دیالوگ و فهم گفتار
  - نویسه‌خوان نوری
  - ترجمه گفتار به گفتار
  - تصویر خوان گفتاری
- روش‌های/دادگان‌های زبان‌شناسی رایانشی

زبان‌شناسی رایانشی . . .

- زبان
  - از مهم‌ترین مشخصه‌های انسان

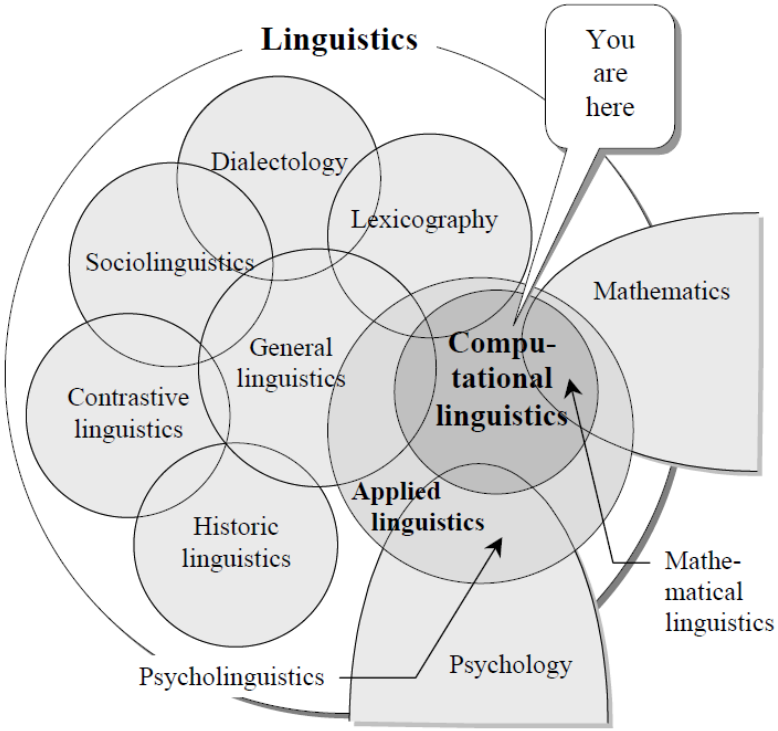
- زبان‌شناسی رایانشی (Computational Linguistics)
  - یک حوزه بین رشته‌ای که به پردازش (مدل‌سازی محاسباتی آماری یا مبتنی بر قاعده) زبان طبیعی می‌پردازد

  - پردازش متن و پردازش گفتار

  - اسامی دیگر
    - Computer Speech and Language Processing
    - Human Language Technology
    - Natural Language Processing (NLP)

زبان‌شناسی رایانشی

- بین‌رشته‌ای است
  - زبان‌شناسی (Linguistics)
  - علوم کامپیوتر و هوش مصنوعی (Computer Science, Artificial Intelligence)
  - روان‌شناسی و علوم شناختی (Psychology, Cognitive Science)



سطوح پردازش زبان

- آواشناسی/واج‌شناسی (Phonetics/Phonology)
- ظاهرشناسی (Graphology)
- ریخت‌شناسی (Morphology)
- نحو/دستور (Syntax/Grammar)
- معناشناسی (Semantics)
- کاربردشناسی (Pragmatics) و گفتمان (Discourse)

سطوح زبان در زبان‌شناسی	توضیح
آواشناسی و واج‌شناسی	صداهای پایه در زبان گفتاری
ظاهرشناسی	شکل نمایش زبان در زبان نوشتاری
ریخت‌شناسی/واژه‌شناسی (صرف)	نحوه شکل‌گیری کلمات
جمله‌شناسی (نحو/دستور)	ترکیب کلمات برای ساخت جمله/عبارت
معناشناسی	معنای کلمات و جملات
کاربردشناسی و گفتمان	معنای سخن در بافت‌های مختلف

کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- بازشناسی (تشخیص) گفتار . . .
  - ASR: Automatic Speech Recognition
  - تبدیل گفتار به متن
  - تایپ گفتاری و سیستم دیکته

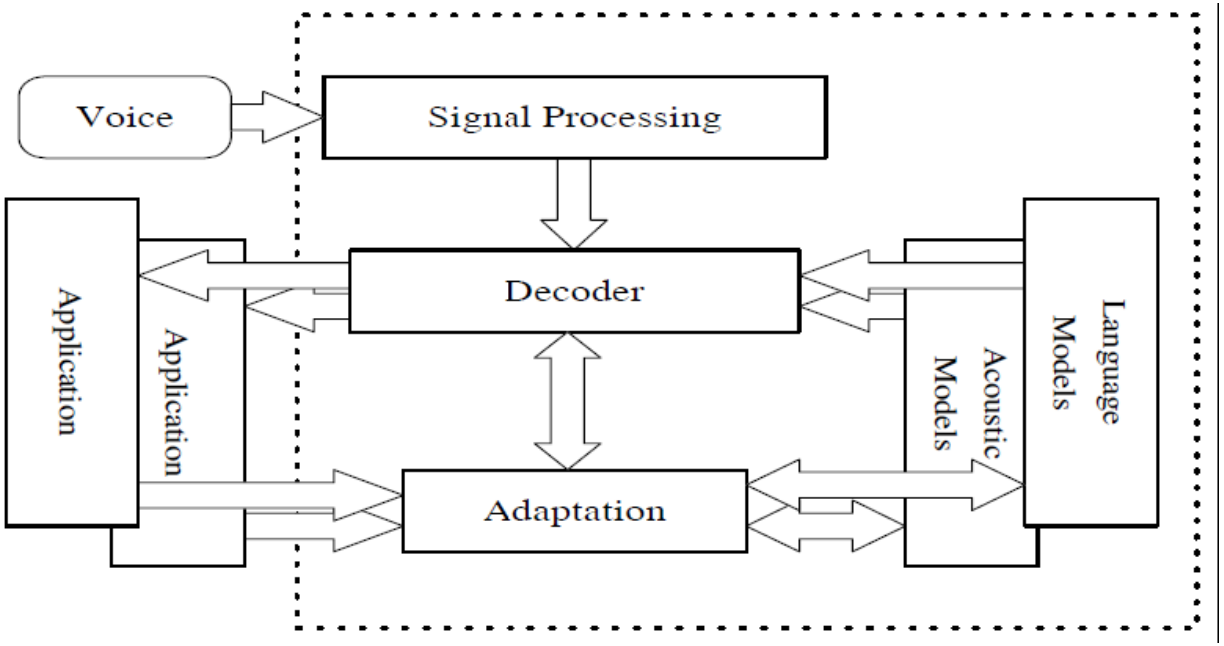
- تشخیص فرامین و دستورات صوتی
  - اجرای برنامه ها در رایانه با بیان نام آنها، کنترل لوازم خانگی با صوت
  - فرمان دادن به ربات‌ها و فرمان‌های صوتی در خودرو
  - کیوسک‌های اطلاعات و دستگاه‌های خود پرداز بانک‌ها
  - استفاده در بازی‌های کامپیوتری (افزایش قابلیت‌ها و جذابیت)
  - و ...

کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

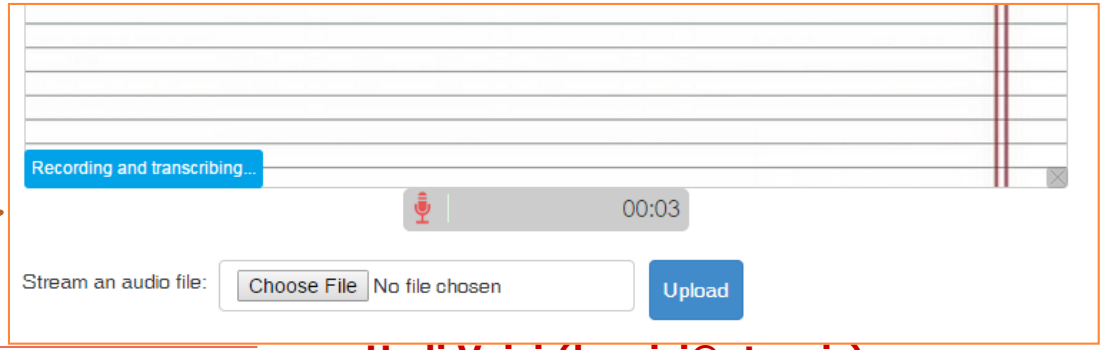
- بازشناسی (تشخیص) گفتار . . .
  - کاربردهای مربوط به معلولین، ناشنوایان و نابینایان
    - معلولین حرکتی: صحبت کردن برای استفاده از وسایل و ابزارها
    - ناشنوایان: تایپ سخنان دیگران، تبدیل گفتار به حرکات ایما و اشاره
    - نابینایان: تایپ گفتاری و رورد اطلاعات، صحبت کردن برای استفاده از وسایل و ابزارها
  - سیستم‌های اطلاع‌رسانی
    - تلفن گویاها
  - سیستم‌های آموزش
    - پرسش و پاسخ
  - کاربردهای ترکیبی: ترجمه گفتار به گفتار

کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- بازشناسی (تشخیص) گفتار . . .



PersianSpeech.com



کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- بازشناسی (تشخیص) گفتار: نیازمندی‌ها
  - پیکره گفتاری برای مدل سازی آوایی
  - پیکره متنی برای مدل سازی زبانی
  - واژگان: حاوی صورت نوشتاری و تلفظی کلمات

  - الگوریتم‌های مدل‌سازی آوایی
  - الگوریتم‌های مدل‌سازی زبانی
  - الگوریتم‌های استخراج ویژگی از گفتار
  - الگوریتم‌های جستجو (رمزگشا)
  - الگوریتم‌های تطبیق (به صدای افراد یا محیط جدید)

کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- سنتز گفتار (Speech Synthesis) . . .
  - تبدیل متن به گفتار (Text-to-Speech)

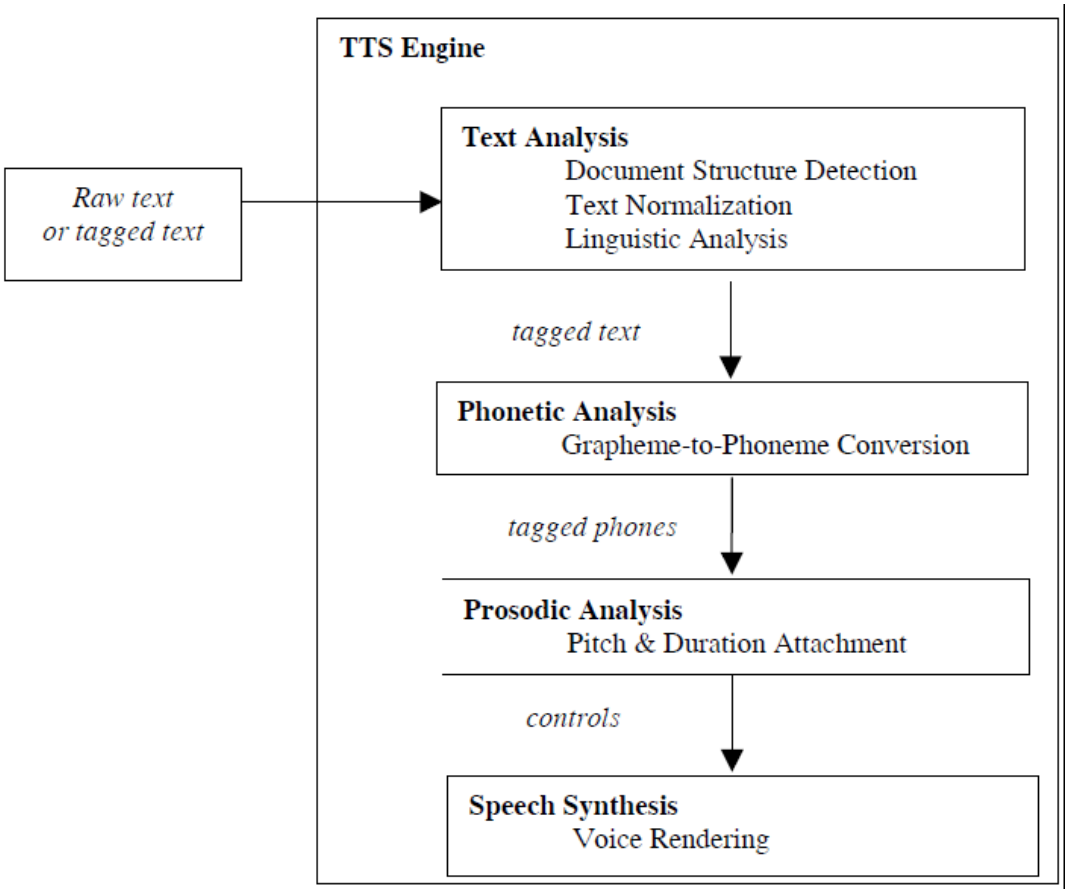
- سخنگو کردن کامپیوتر
- سیستم‌های آموزشی (کتاب‌های الکترونیکی، محتواهای آموزشی، آموزش از راه دور)
- وب سایت ها (اخبار و اطلاعات، آموزش)
- تلفن های همراه (خواندن پیامک، نقشه، کتاب و ....)
- رفع مشکلات و محدودیت‌های افراد ناتوان جسمی در صحبت کردن و خواندن

کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- سنتز گفتار . . .
  - رفع مشکلات و محدودیت‌های افراد ناتوان در خواندن
    - نابینایان، افراد کم‌سواد و بی‌سواد و سالمندان
    - سیستم های صفحه خوان (Screen Reader) ویژه نابینایان
      - JAWS
      - HAL
      - Talks (موبایل)
  - سامانه‌های اطلاع‌رسانی مانند سیستم‌های تلفنی (تلفن گویا)، کیوسک‌ها، نوبت‌دهی بانک‌ها
    - تغییر سریع و آسان پیغام‌های صوتی بدون نیاز به ضبط صدا
  - استفاده در نرم‌افزارهای دیگر: مترجم گفتار به گفتار، OCR
  - فشرده سازی گفتار (کد کردن)

کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- سنتز گفتار . . .



کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- سنتز گفتار

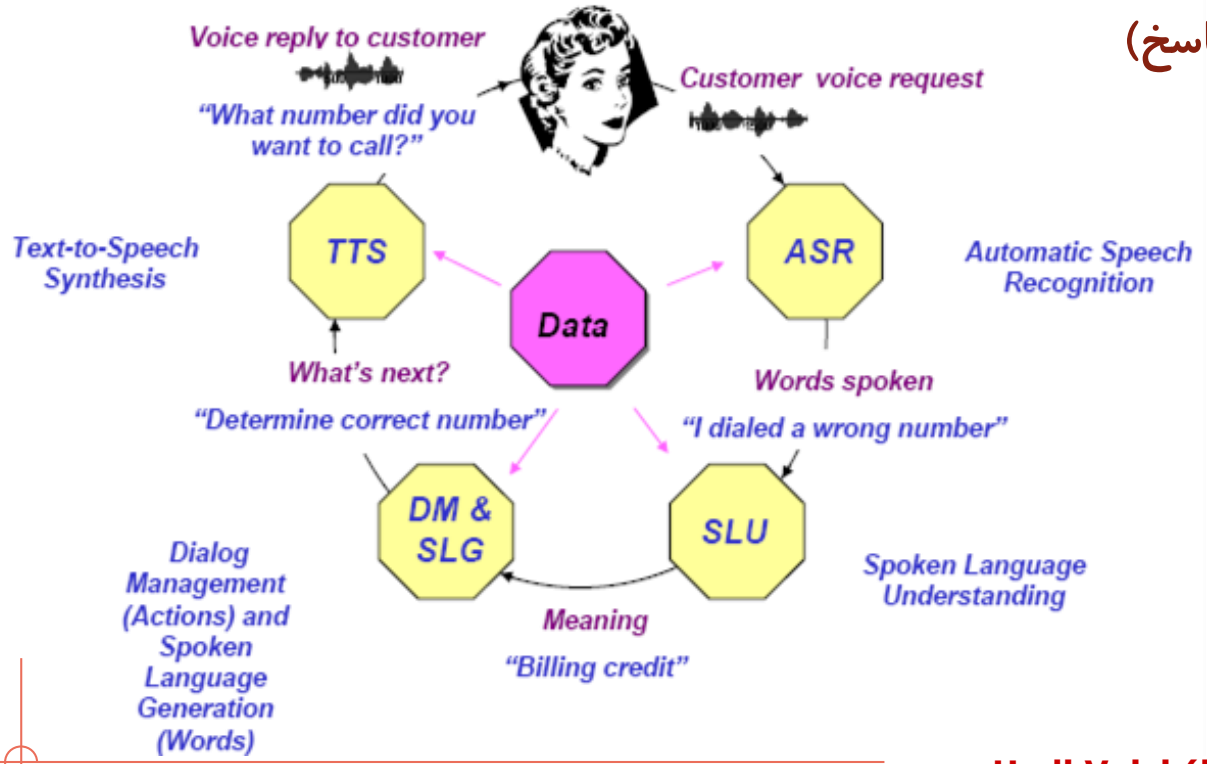
FarsiReader.com

کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- سنتز گفتار: نیازمندی‌ها
  - نرمال‌سازی متن
  - واحدسازی متن
  - تلفظ کردن خودکار کلمات
  - تشخیص تلفظ درست در هم‌نویسه‌ها
    - برچسب زن اجزای کلام
  - تشخیص کسره اضافه
  - تشخیص آوا و نوا در متن
  - روشی برای تولید گفتار
    - نمونه‌هایی از گفتار

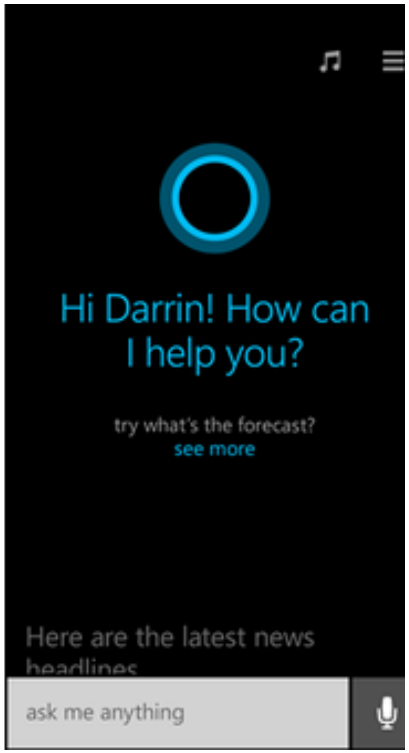
کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- دیالوگ و فهم گفتار (Speech Dialogue/Understanding)
  - تشخیص گفتار (ورود اطلاعات)
  - درک (فهم) گفتار
  - تولید جمله (تولید پاسخ)
  - تولید گفتار (خواندن پاسخ)



کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- دیالوگ و فهم گفتار . . .
  - Apple Siri روی iPhone
  - Microsoft Cortana روی ویندوز موبایل ۸.۱
  - Google Now برای اندروید



کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- دیالوگ و فهم گفتار (سیستم‌های اطلاع‌رسانی تلفنی)
  - Microsoft TellMe
  - Jupiter (اطلاعات آب و هوایی شهرهای دنیا)

Jupiter

A conversational interface for on-line weather information over the phone.

1-888-573-8255

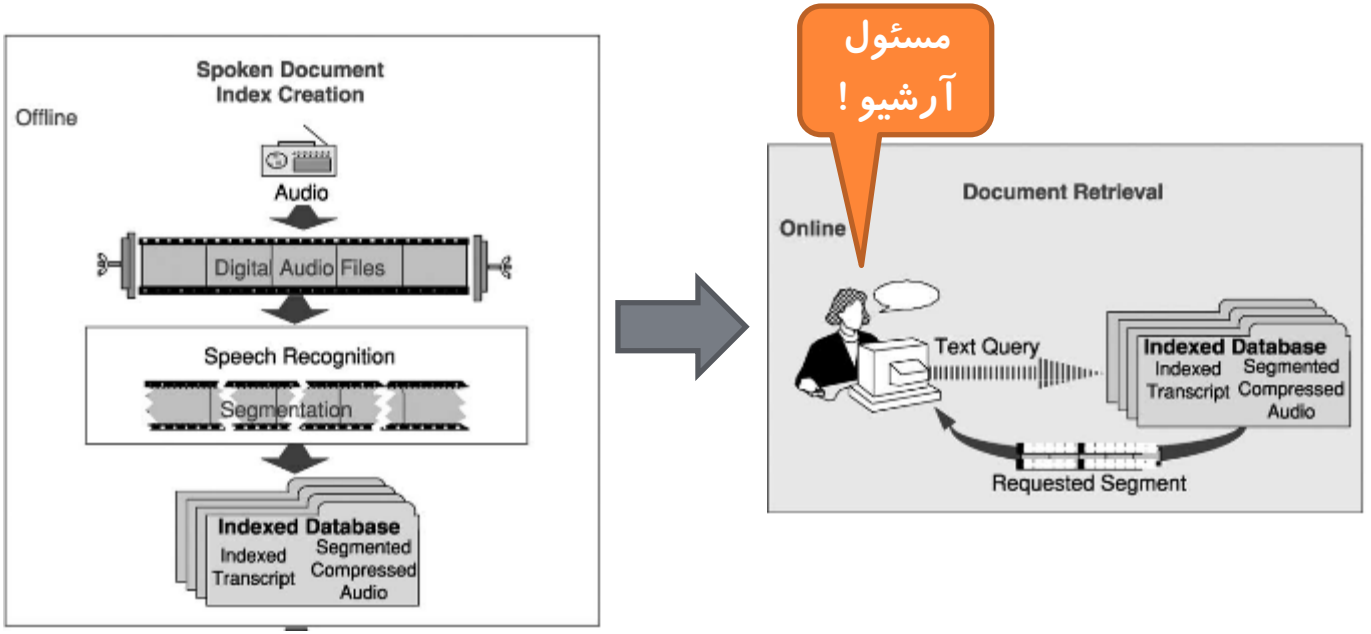
(outside the USA: **1-617-258-0300**)

http://www.sls.lcs.mit.edu/jupiter

Spoken Language Systems Group,
MIT Laboratory for Computer Science

کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

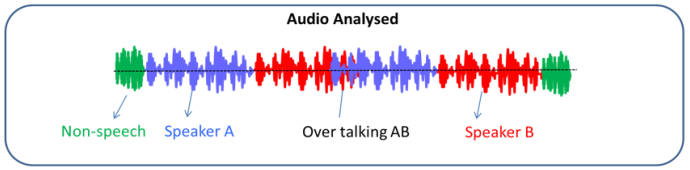
- بازیابی اطلاعات گفتاری (Spoken Document Retrieval)
  - بازیابی (جستجو) کلمه‌ها و عبارت‌ها در یک مجموعه گفتاری
    - مانند آرشیوهای صدا و سیما
    - فایل های صوتی اینترنت



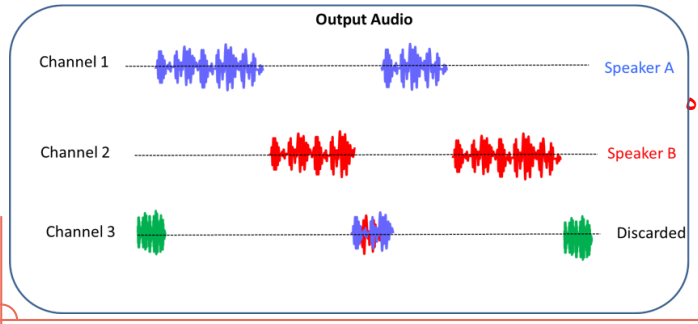
کاربردهای زبان‌شناسی رایانشی: گفتاری . . .

- بازشناسی . . .
  - زبان (Language Identification)
  - جنسیت (Gender Identification)

- بازشناسی گوینده (Speaker Recognition)
  - تشخیص گوینده (Speaker Identification)
    - کسی که صحبت می‌کند، کیست؟
  - حالت دیگر: تایید گوینده (Speaker Verification)
    - آیا او واقعاً حسین است؟



  - جداسازی گوینده (Speaker Diarization)
    - جداسازی بخش‌های یک مکالمه (دو یا چند نفره) به تفکیک گوینده
    - چه کسی، چه زمانی صحبت کرده است؟



  - تایپ گفتاری صورت‌جلسه‌ها (یا مذاکرات صحن علنی مجلس)

کاربردهای زبان‌شناسی رایانشی: متنی . . .

- حدود ۸۰٪ از اطلاعات به صورت متن است

کاربردهای زبان‌شناسی رایانشی: متنی . . .

- غلط‌یاب املایی (Spell Checker) . . .
  - تشخیص غلط‌های املایی واژه‌ها
  - اصلاح غلط‌های املایی واژه‌ها
    - ارائه لیستی از واژه‌های صحیح پیشنهادی
  - اصلاح غلط‌های فاصله‌گذاری و علائم سجاوندی
    - اصلاح کاربرد نابجای فاصله به جای شبه‌فاصله

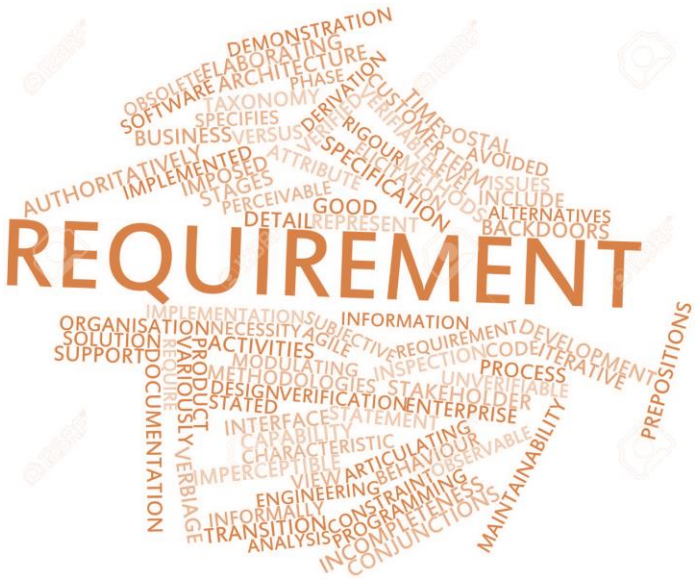
املایی

در Word

One of the essential requirements of artifisial intelligence applied systems such as speech recognition are the incorporation of language models and language information. AGP uses the latest methods in Natural Language Processing to extract and apply language information to various systems.

کاربردهای زبان‌شناسی رایانشی: متنی . . .

- غلط‌یاب املایی (Spell Checker): نیازمندی‌ها
  - لیستی از واژه‌های صحیح: لغت‌نامه
    - تحلیل ساخت واژی
  - الگوریتمی برای مقایسه واژه‌های متن با واژه‌های لغت‌نامه
  - الگوریتمی برای ارائه پیشنهاد جایگزین برای واژه‌های نادرست متن
  - الگوریتمی برای رتبه‌بندی پیشنهادها
    - مدل‌سازی آماری زبان (احتمالاتی)



کاربردهای زبان‌شناسی رایانشی: متنی . . .

- غلط‌یاب گرامری (Grammar Checker)
  - تشخیص غلط‌های گرامری
  - اصلاح غلط‌های گرامری
    - پیشنهاد برای اصلاح گرامری
  - نیازمندی‌ها
    - گرامر محاسباتی زبان
    - الگوریتمی برای مقایسه کلمات متن با گرامر درست
      - روش‌های تجزیه نحوی
    - الگوریتمی برای ارائه پیشنهاد گرامر درست

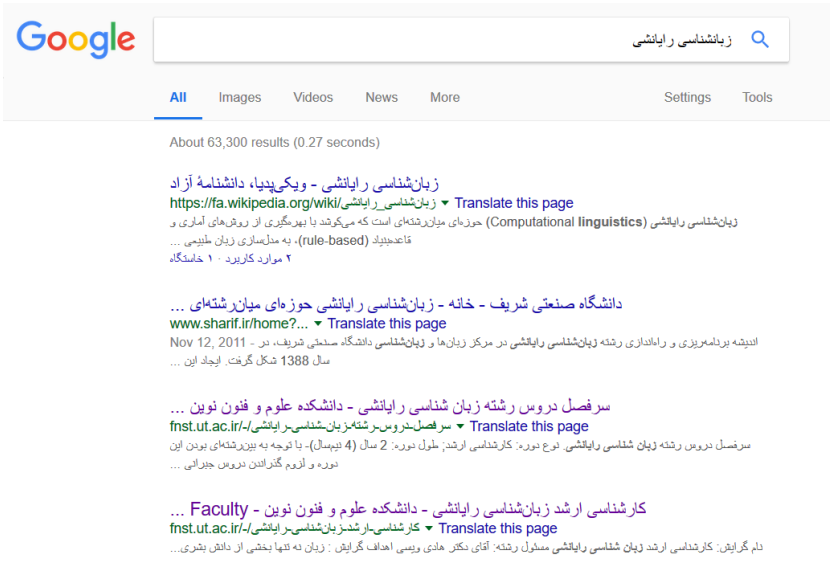
در Word

گرامری

One of the essential requirements of artifisial intelligence applied systems such as speech recognition are the incorporation of language models and language information. AGP uses the latest methods in Natural Language Processing to extract and apply language information to various systems.

کاربردهای زبان‌شناسی رایانشی: متنی . . .

- بازیابی متن (Text Retrieval) . . .
  - جستجو در متن (غیرساختار یافته)
    - جستجوهای اینترنتی مانند گوگل
  - مقایسه شباهت متن/کلمات جستجو با متن همه صفحات
    - جمع‌آوری مخزن داده‌ها و نمایه کردن آن
  - رتبه‌بندی نتایج

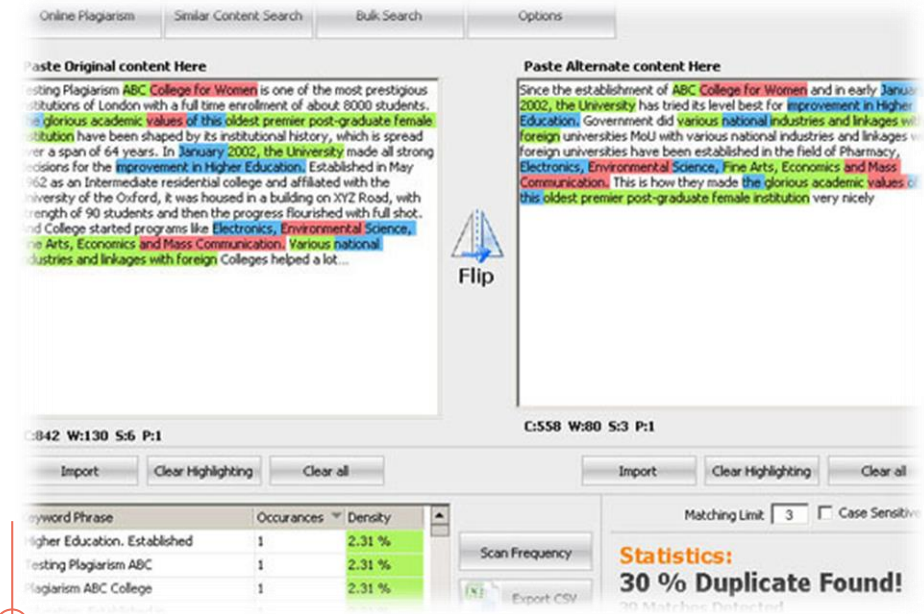


کاربردهای زبان‌شناسی رایانشی: متنی . . .

- بازیابی متن (Text Retrieval): نیازمندی‌ها
  - جمع‌آوری داده‌های مورد نیاز به عنوان منبع و نمایه کردن آنها
    - ذخیره ریشه کلمات به همراه/به جای خود کلمات
      - روش‌های ریشه‌یابی
  - الگوریتمی برای مقایسه شباهت متن/کلمات جستجو با متن همه صفحات
    - روش‌های تشابه‌یابی متن
  - الگوریتمی برای رتبه‌بندی نتایج

کاربردهای زبان‌شناسی رایانشی: متنی . . .

- تشابه‌یابی متن (Document Similarity)
  - یافتن دو متن مشابه
  - کشف مقاله‌ها و پایان‌نامه‌های مشابه (تقلب علمی)
    - تشابه ظاهری
    - تشابه مفهومی
    - تشابه ایده
    - ترجمه
  - بازیابی اطلاعات
  - کشف ایمیل‌های هرز (Spam)
  - مستندات کپی و یکسان در اینترنت
    - روند انتشار در شبکه های اجتماعی



کاربردهای زبان‌شناسی رایانشی: متنی . . .

- خلاصه‌سازی (Summarization) . . .
  - تولید چکیده‌های مقالات علمی، اخبار و ...
  - هدف: حفظ حداکثری نکات اصلی و کاهش بیشترین میزان افزونگی
  - انواع خلاصه‌سازی
    - استخراجی (Extractive): شناسایی جملاتی از متن اصلی حاوی مفاهیم اصلی
    - چکیده‌ای (Abstractive): چکیده‌ای از مفاهیم مهم متن به صورت خلاصه با کلمات و ساختار متفاوت از متن اصلی
    - تک سندی/ چندسندی
    - تک زبانه/چند زبانه

کاربردهای زبان‌شناسی رایانشی: متنی . . .

- خلاصه‌سازی (Summarization): نیازمندی‌ها
  - تشخیص موجودیت‌های اسمی (Named Entity Recognition)
    - تشخیص اسامی خاص
  - تشخیص مرجع ضمیر (Co-reference Resolution)
  - استخراج واحدهای هم‌معنا
  - نمایش متن (کلمه/جمله)

کاربردهای زبان‌شناسی رایانشی: متنی . . .

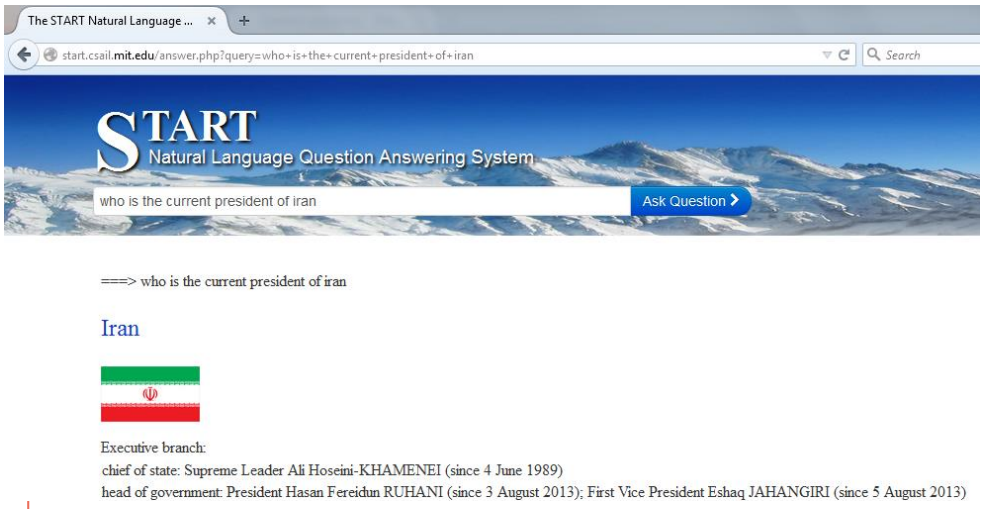
- **عقیده کاوی (Opinion Mining) . . .**
  - **بررسی نظرات کاربران در مورد محصولات/اخبار/شبکه‌های اجتماعی/ ...**
    - حدود ۸۱٪ کاربران اینترنتی قبل از خرید آن را در اینترنت بررسی می‌کنند
    - حدود ٤٠٪ از خریدهای محصولات توسط افراد بر اساس نظرات کاربران دیگر است
    - مثال: دسته‌بندی نظرات کاربران بر اساس موافق یا مخالف (Polarity Detection)
  - **اسم دیگر: تحلیل احساس (Sentiment Analysis)**

کاربردهای زبان‌شناسی رایانشی: متنی . . .

- عقیده کاوی (Opinion Mining): نیازمندی‌ها
  - شناسایی کلمات دارای بار احساسی
    - صفت‌های/فعل‌های مثبت و منفی
  - شناسایی قوانین نحوی تغییر دهنده
    - مثال «بد نیست» دو کلمه با بار منفی اما در مجموع دارای بار مثبت
  - دادگان برچسب خورده
    - متن (نظر) حاوی عقیده و برچسب آن
  - الگوریتم‌های یادگیری از روی داده

کاربردهای زبان‌شناسی رایانشی: متنی . . .

- سیستم‌های پرسش و پاسخ (Question Answering Systems) . . .
  - بازیابی اطلاعات متنی
  - وردی: سوالات به صورتی که انسان می‌پرسد نه در قالب کلمات کلیدی
    - مثال: شرکت مایکروسافت چقدر هزینه تحقیق و توسعه در سال ۲۰۱۴ پرداخت کرده است؟
  - نسل جدید موتورهای جستجو
  - انواع: دامنه محدود (Restrict Domain) و دامنه باز (Open Domain)



IBM Watson

کاربردهای زبان‌شناسی رایانشی: متنی . . .

- سیستم‌های پرسش و پاسخ (QA Systems): نیازمندی‌ها
  - واحدسازی و نرمال‌سازی پرسش و متون پایگاه داده
  - ریشه‌یابی
  - تحلیل نحوی
  - تحلیل معنایی
    - شبکه‌ی واژگان یا هستان‌شناسی (وردنت)
  - الگوریتم‌های تشابه‌یابی
  - پایگاه داده

کاربردهای زبان‌شناسی رایانشی: متنی . . .

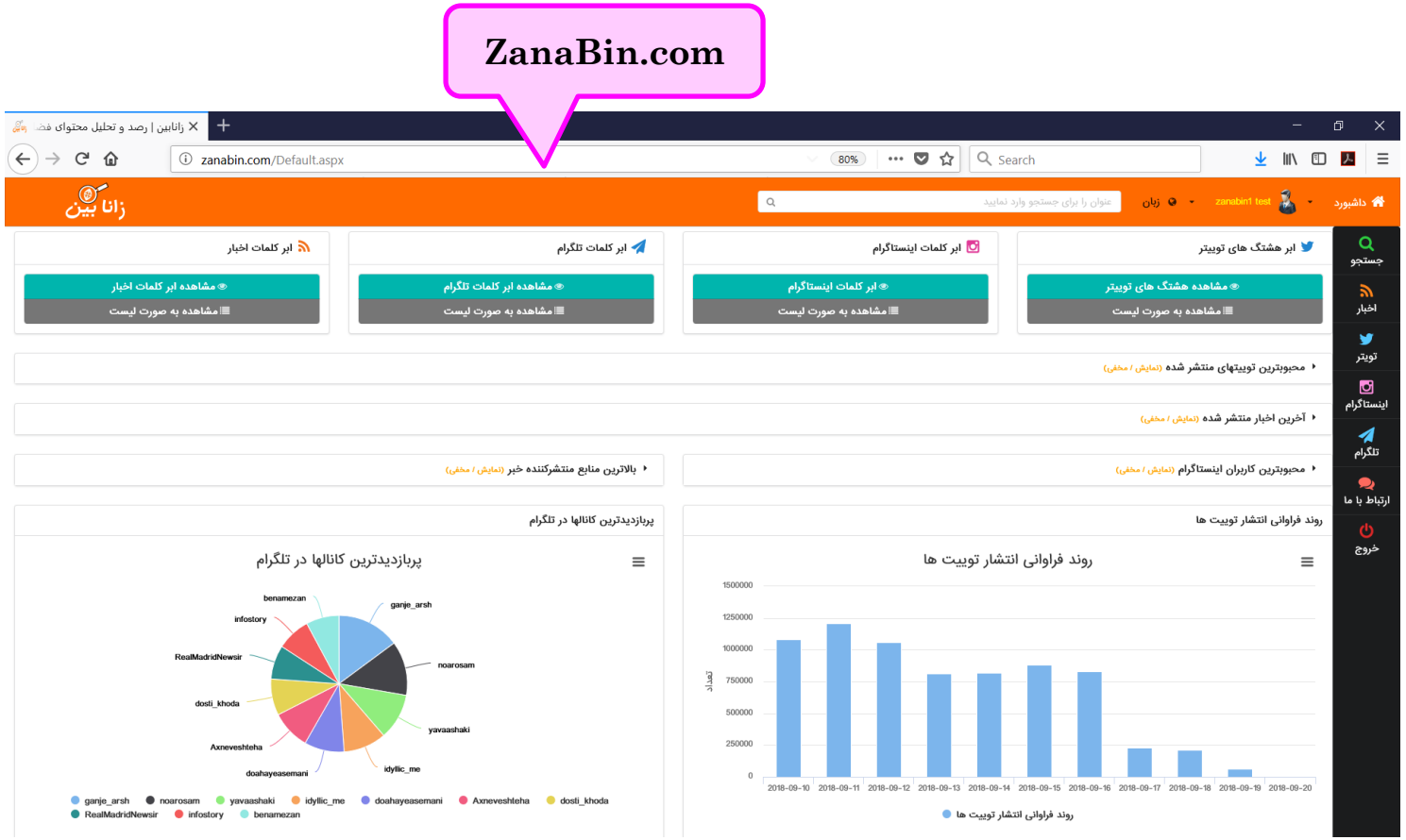
- پایش و تحلیل فضای مجازی . . .
  - جمع‌آوری اطلاعات از بسترها (سایت‌های خبری، شبکه‌های اجتماعی) و منابع (سایت‌ها و اکانت‌ها) مختلف
  - پردازش اطلاعات و تحلیل آنها
    - ارائه آمارها و تحلیل های مختلف
      - داغ ترین مطالب
      - نحوه نشر اطلاعات
      - نظرات کاربران در مورد پست‌ها

- نیازمندی‌ها
  - نرمال سازی و واحدسازی
  - تشابه یابی اسناد
  - بازشناسی موجودیت های اسمی (NER)
  - قطعه‌بندی (Chunking)

Hadi Veisi (h.veisi@ut.ac.ir)

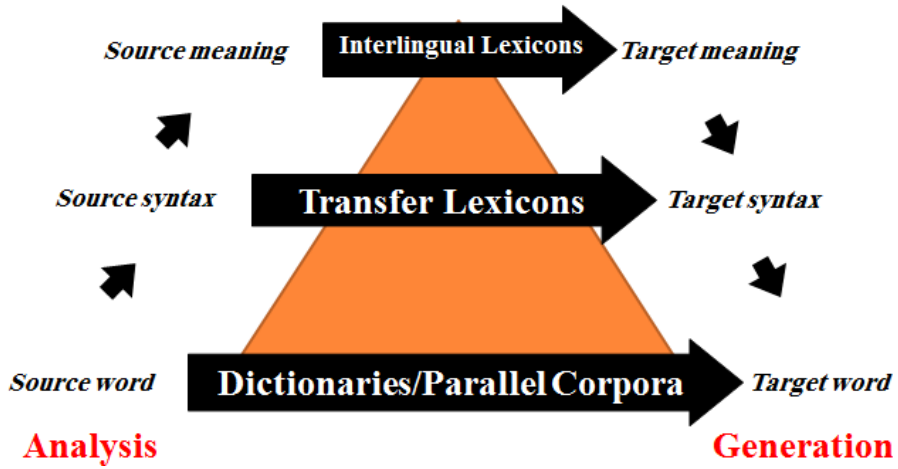
کاربردهای زبان‌شناسی رایانشی: متنی . . .

- پایش و تحلیل فضای مجازی



کاربردهای زبان‌شناسی رایانشی: متنی . . .

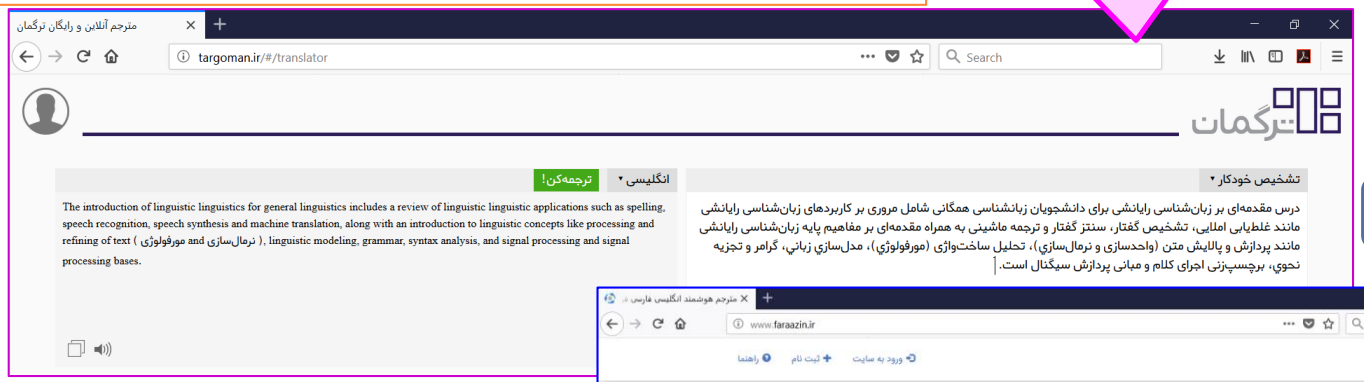
- ترجمه ماشینی (Machine Translation). . .
  - ترجمه متن از زبانی به یک یا چند زبان دیگر

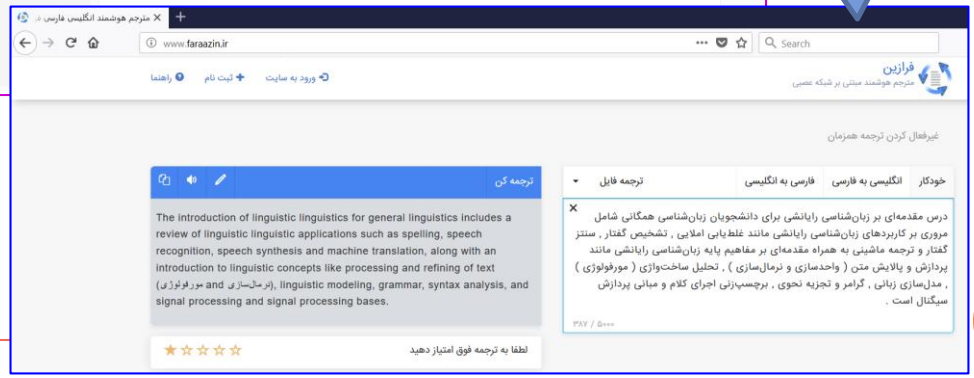


- روش‌ها
  - مبتنی بر قانون: پردازش ساخت‌واژی، نحوی، معنایی
  - آماری: یادگیری از روی داده

کاربردهای زبان‌شناسی رایانشی: متنی . . .

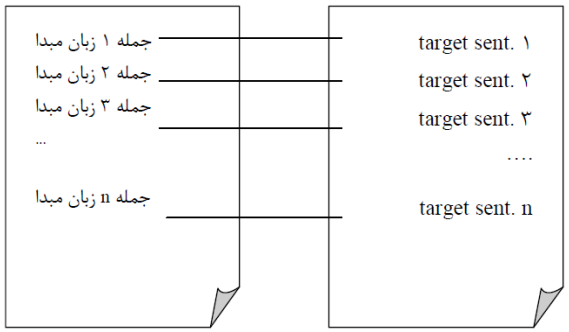
- ترجمه ماشینی (Machine Translation). . .





کاربردهای زبان‌شناسی رایانشی: متنی . . .

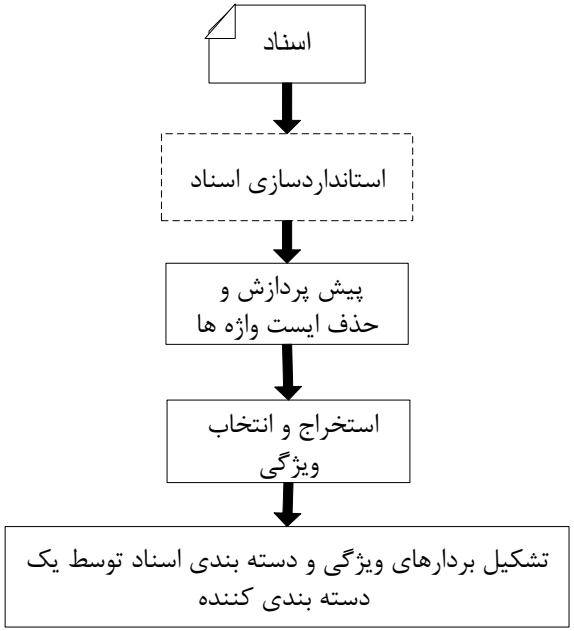
- ترجمه ماشینی (Machine Translation): نیازمندی‌ها
  - دادگان متنی موازی
    - فارسی – انگلیسی



- الگوریتم‌های یادگیری
- بازشناسی موجودیت‌های اسمی (NER)
- تحلیل نحوی
- تحلیل معنایی
- رفع ابهام معنایی (WSD: Word Sense Disambiguation)
- ارزیابی سیستم‌های ترجمه
  - BLUE

کاربردهای زبان‌شناسی رایانشی: متنی . . .

- تشخیص موضوع/عنوان (Topic Identification)
  - برای دسته‌بندی و ساختار دادن به متون (دسته‌بندی اخبار)



- نیازمندی‌ها
  - ایست واژه‌ها
  - نمایش سند
  - داده دارای برچسب دسته (نوع/موضوع سند)
  - الگوریتم های یادگیری

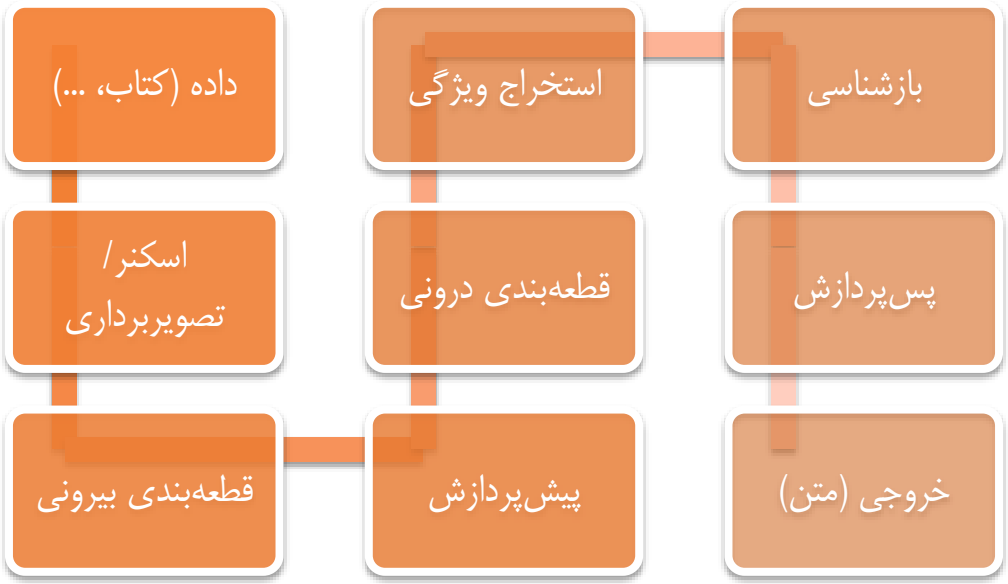
کاربردهای زبان‌شناسی رایانشی: متنی . . .

- تشخیص نویسنده (جنسیت)
  - برای شناسایی هویت افراد و یا تصدیق هویت
    - ردیابی هویت بزهکاران
  - تشخیص جنسیت برای تبلیغات
  - برای ترجمه و در زبان‌هایی که ضمیر دارای جنسیت است

- نیازمندی‌ها
  - ویژگی‌ها: غنای واژگان، طول کلمات، تعداد علائم سجاوندی، طول جمله، تعداد صفات/ضمایر و ...
  - داده دارای برچسب دسته (نوع/موضوع سند)
  - الگوریتم های یادگیری

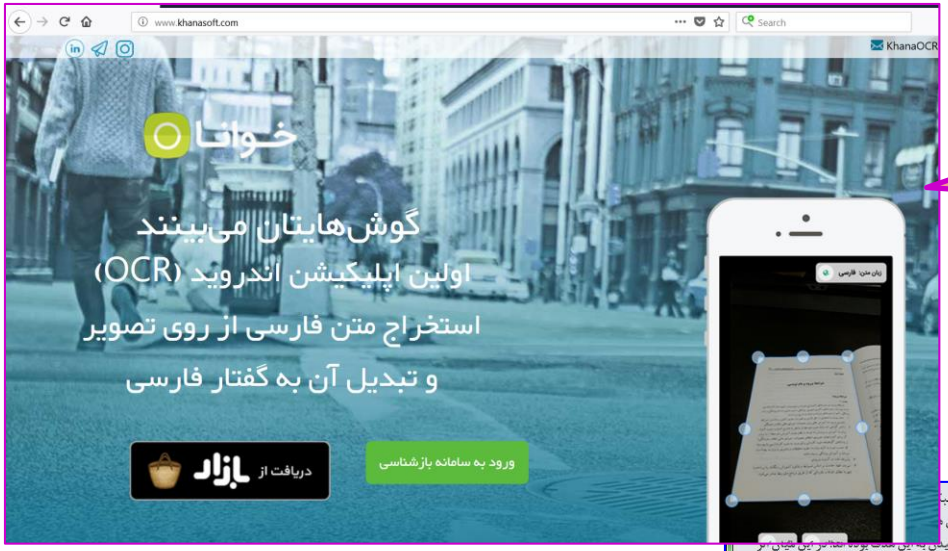
کاربردهای زبان‌شناسی رایانشی: تصویری . . .

- نویسه‌خوان نوری (OCR: Optical Character Recognition) . . .
  - تبدیل عکس (حاوی تصویر متن) به متن الکترونیکی
  - نوع دیگر: بازشناسی دست‌خط (Handwritten Recognition)



کاربردهای زبان‌شناسی رایانشی: تصویری . . .

- نویسه‌خوان نوری (OCR) . . .



khanasoft.com

Farsiocr.ir

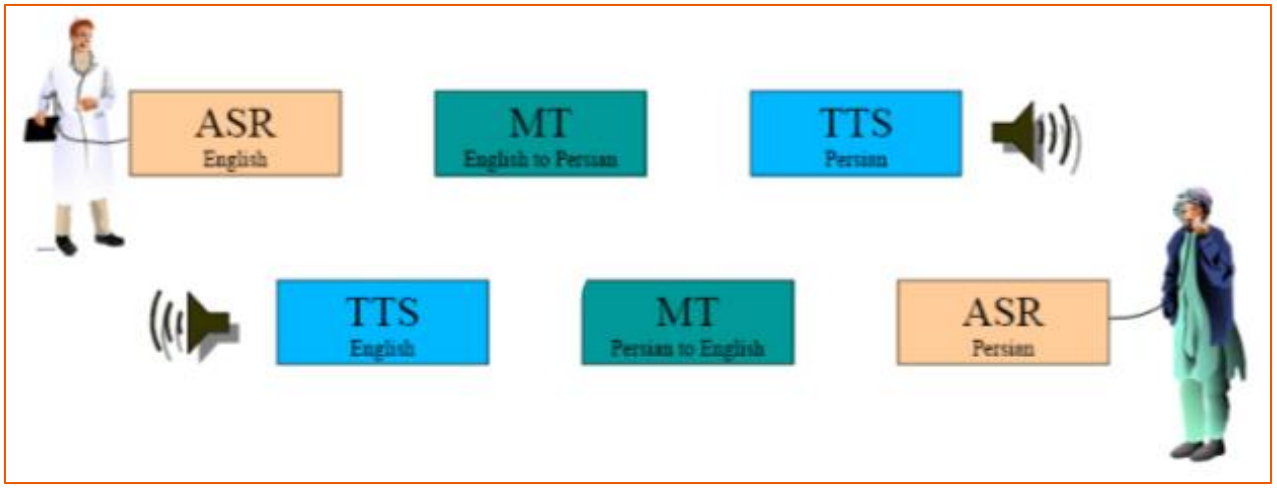
کاربردهای زبان‌شناسی رایانشی: تصویری . . .

- نویسه‌خوان نوری (OCR): نیازمندی‌ها
  - دادگان تصویری دارای برچسب متنی معادل
  - لغت‌نامه (برای پس پردازش)

  - الگوریتم‌های مدل‌سازی نویسه‌ها/زیرکلمات
  - الگوریتم‌های تحلیل ساختار تصویر
  - الگوریتم‌های استخراج ویژگی از تصویر

کاربردهای زبان‌شناسی رایانشی: ترکیبی . . .

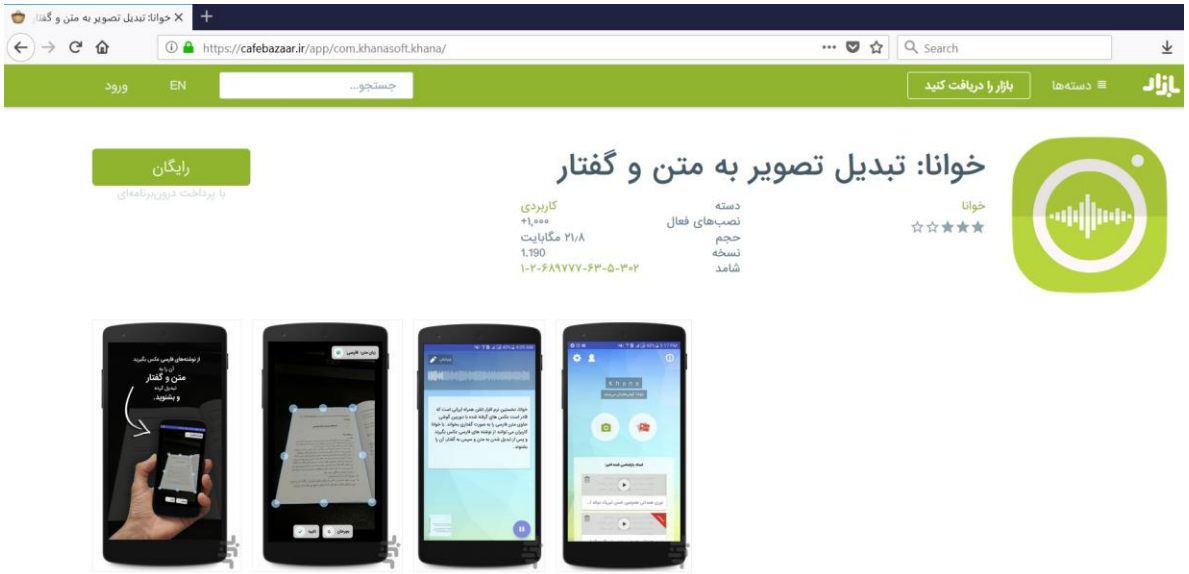
- ترجمه گفتار به گفتار
  - ترجمه از روی گفتار یک زبان به معادل گفتاری زبانی دیگر
  - ترکیب ASR، MT و TTS

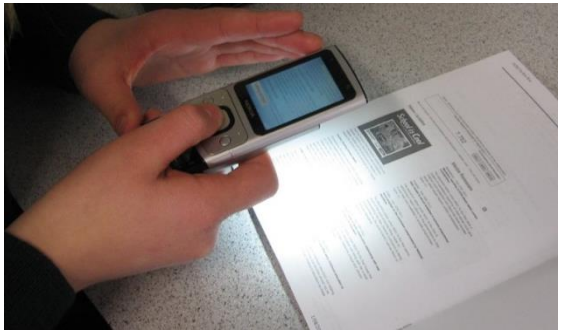


- نرم‌افزار Skype

کاربردهای زبان‌شناسی رایانشی: ترکیبی . . .

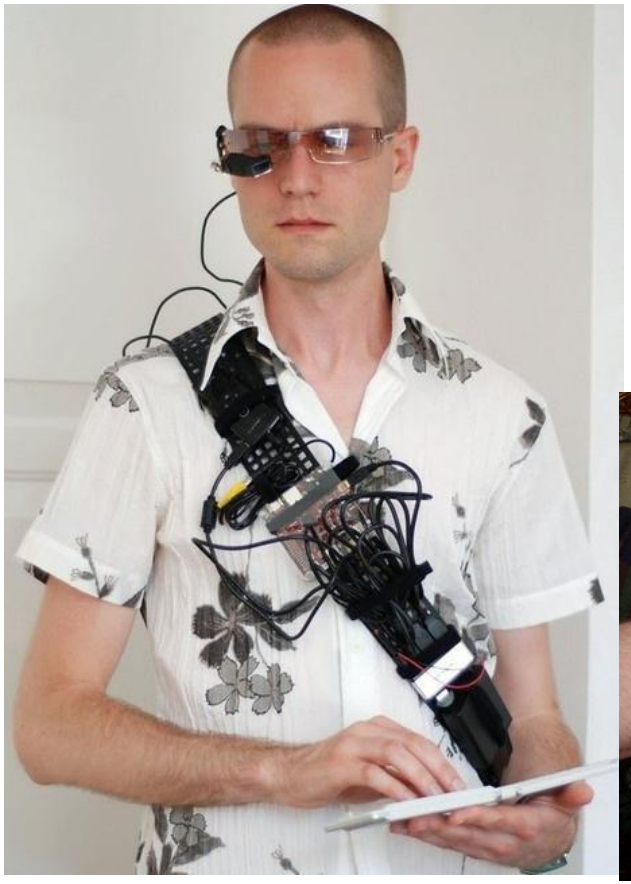
- تصویر خوان گفتاری
  - گرفتن عکس از متن و خواندن آن
  - ترکیب OCR و TTS





کاربردهای زبان‌شناسی رایانشی: ترکیبی

- کامپیوترهای پوشیدنی (Wearable Computers)
  - کاربردهای نظامی
  - عینک گوگل



روش‌های زبان‌شناسی رایانشی . . .

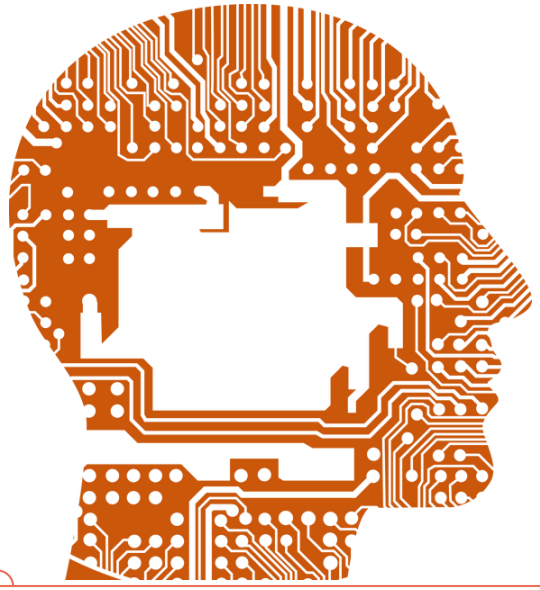
- برای پردازش متن
  - نرمال‌سازی و واحدسازی (Text Tokenization and Normalization)
  - تحلیل ساخت واژی (Morphological Analysis)
  - برچسب‌زنی اجزای کلام (Part-of-Speech Tagging)
  - تجزیه/تحلیل نحوی (Syntactic Parsing/Analysis)
  - تحلیل معنایی (Semantic Analysis)
  - رفع ابهام معنایی (Word Sense Disambiguation)
  - تشخیص مرجع ضمیر (Co-reference Resolution)
  - تشخیص موجودیت‌های اسمی (Named Entity Recognition)
  - ریشه‌یابی/بن‌واژه‌سازی (Stemming/Lemmatization)
  - مدل‌سازی زبانی (Language Modeling)
  - تلفظ خودکار کلمات
    - تشخیص تلفظ درست در هم‌نویسه‌ها
  - تشخیص کسره اضافه

روش‌های زبان‌شناسی رایانشی . . .

- برای پردازش گفتار
  - مدل‌سازی آوایی (Acoustic Modeling)
  - تولید گفتار
  - تحلیل نوا و آوا
  - استخرج ویژگی از گفتار
  - تبدیل صدا
    - تبدیل صدای یک نفر به صدای فردی دیگر یا صدای ناشناس
  - بهسازی گفتار و حذف نویز
  - فشرده‌سازی و کدکردن گفتار

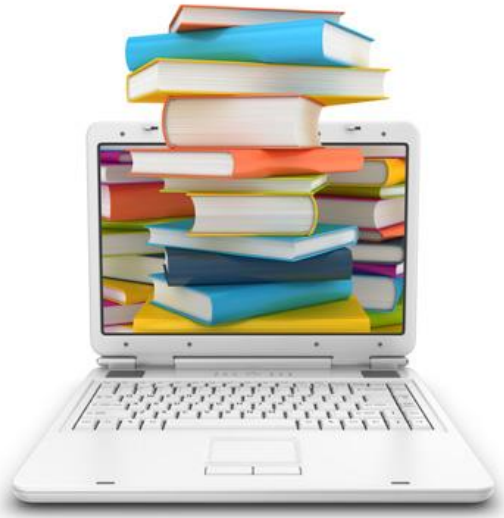
روش‌های زبان‌شناسی رایانشی

- روش‌های مشترک
  - یادگیری ماشین (Machine Learning) برای مدل‌سازی و بازشناسی
    - مدل مخفی مارکوف (HMM: Hidden Markov Model)
    - شبکه‌های عصبی مصنوعی (ANN: Artificial Neural Network)
    - یادگیری بیز(Bayesian Learning)
    - درخت تصمیم (Decision Tree)
    - ماشین بردار پشتیبان (SVM: Support Vector Machine)
    - میدان های تصادفی شرطی (CRF: Conditional Random Fields)



دادگان‌های زبان‌شناسی رایانشی . . .

- متنی
  - پیکره متن خام: ساخت مدل زبانی
  - پیکره موازی: ترجمه ماشینی
  - جملات با برچسب نحوی: برچسب زنی اجزای کلام
  - بانک درخت: تحلیل نحوی
  - شبکه واژگان (وردنت): تحلیل معنایی
  - اسناد برچسب خورده: موضوع، احساس
  - لغت‌نامه: غلط‌یابی املایی
  - واژگان محاسباتی: نوشتار+تلفظ+ریشه+برچسب‌ها + ...
  - گرامر محاسباتی



دادگان‌های زبان‌شناسی رایانشی

گفتاری

- پیکره گفتاری دارای برچسب متنی: بازشناسی گفتار
- پیکره گفتاری تقطیع شده: بازشناسی گفتار + تحلیل‌های آوایی/نوایی

تصویری

- تصاویر حاوی نوشته به همراه متن معادل: نویسه خوان نوری