

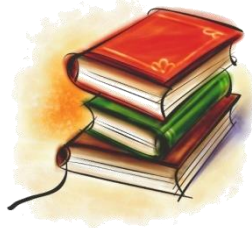
آشنایی با زبان‌شناسی رایانشی

مقدمه و معرفی

هادی ویسی

h.veisi@ut.ac.ir

دانشگاه تهران - دانشکده علوم و فنون نوین



○ 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 ○

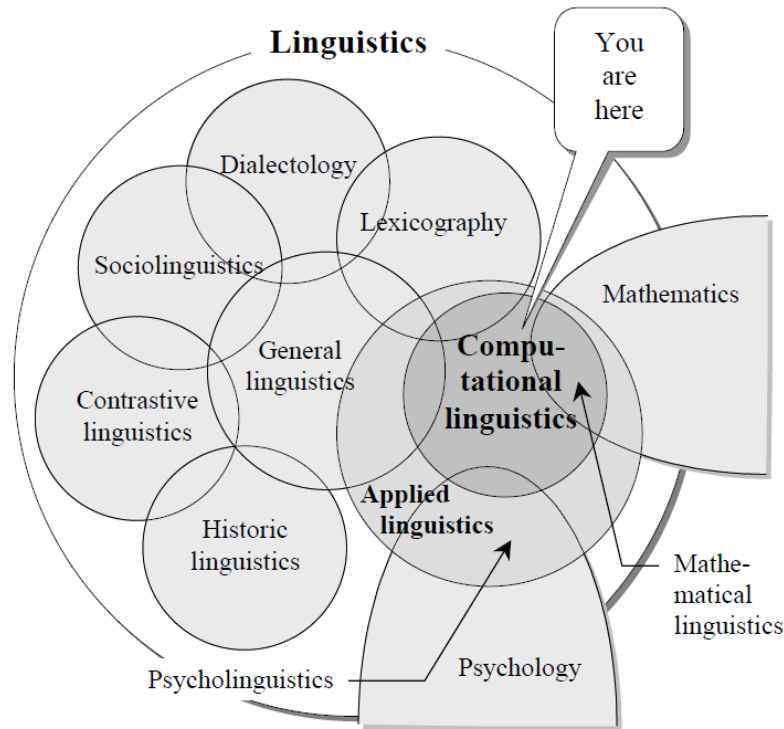
Human Language Technology ○

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)

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



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 برنامه‌ها در رایانه با بیان نام آنها، کنترل لوازم خانگی با صوت

○ فرمان دادن به ربات‌ها و فرمان‌های صوتی در خودرو

○ کیوسک‌های اطلاعات و دستگاه‌های خودپرداز 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 ورود اطلاعات، صحبت کردن برای استفاده از وسایل و ابزارها

• سیستم‌های اطلاع‌رسانی

- تلفن گویاها

• سیستم‌های آموزش

- پرسش و پاسخ

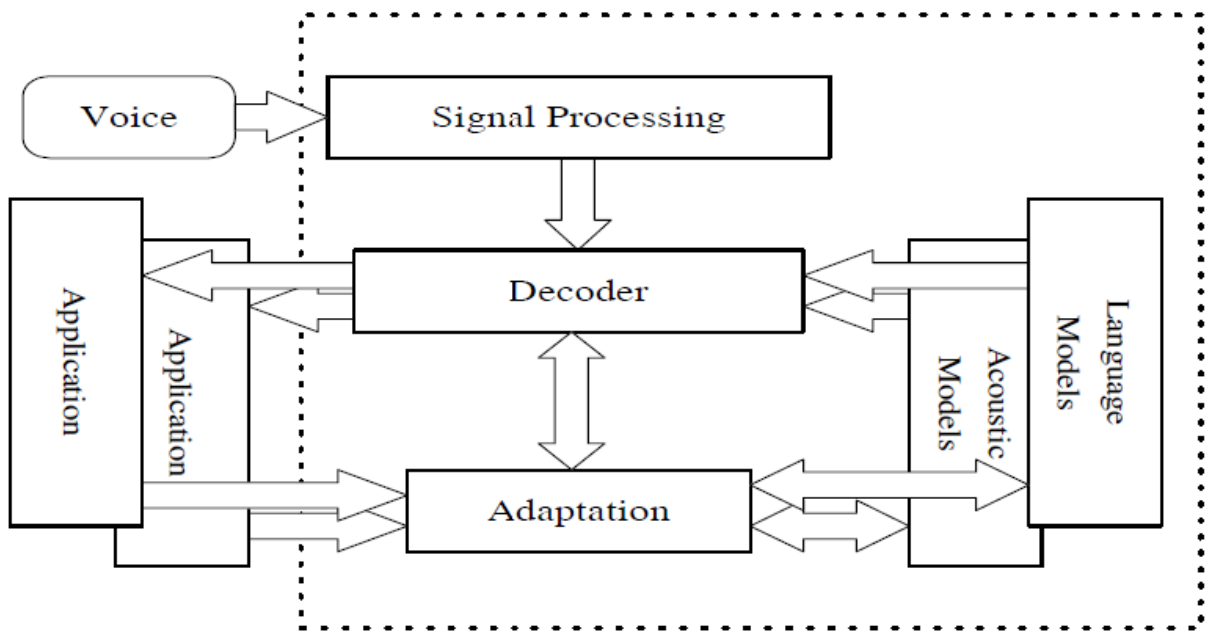
• کاربردهای ترکیبی: ترجمه گفتار به گفتار



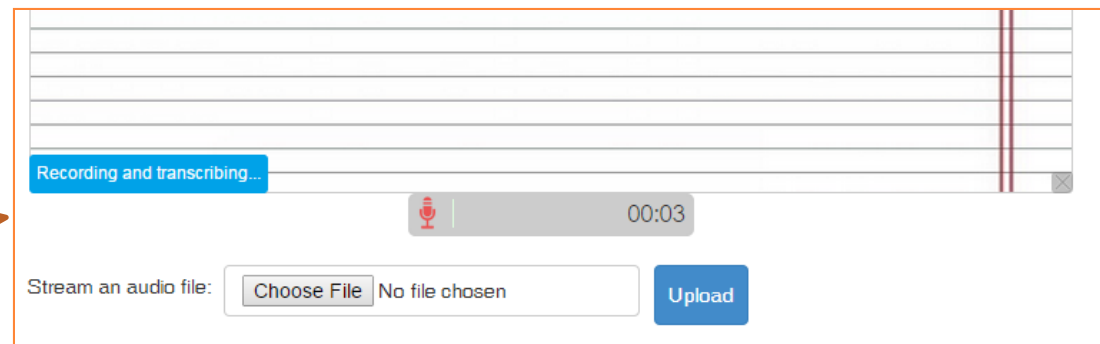


کاربردهای زبان‌شناسی رایانشی: گفتاری ...

○ 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 مدل سازی آوایی
- الگوریتم‌های مدل سازی زبانی
- الگوریتم‌های استخراج ویژگی از گفتار
- الگوریتم‌های جستجو (رمز گشا)
- الگوریتم‌های تطبیق (به صدای افراد یا محیط جدید)



Requirements

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 سایت‌ها (اخبار و اطلاعات، آموزش)

• تلفن‌های 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

- سیستم‌های صفحه‌خوان (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

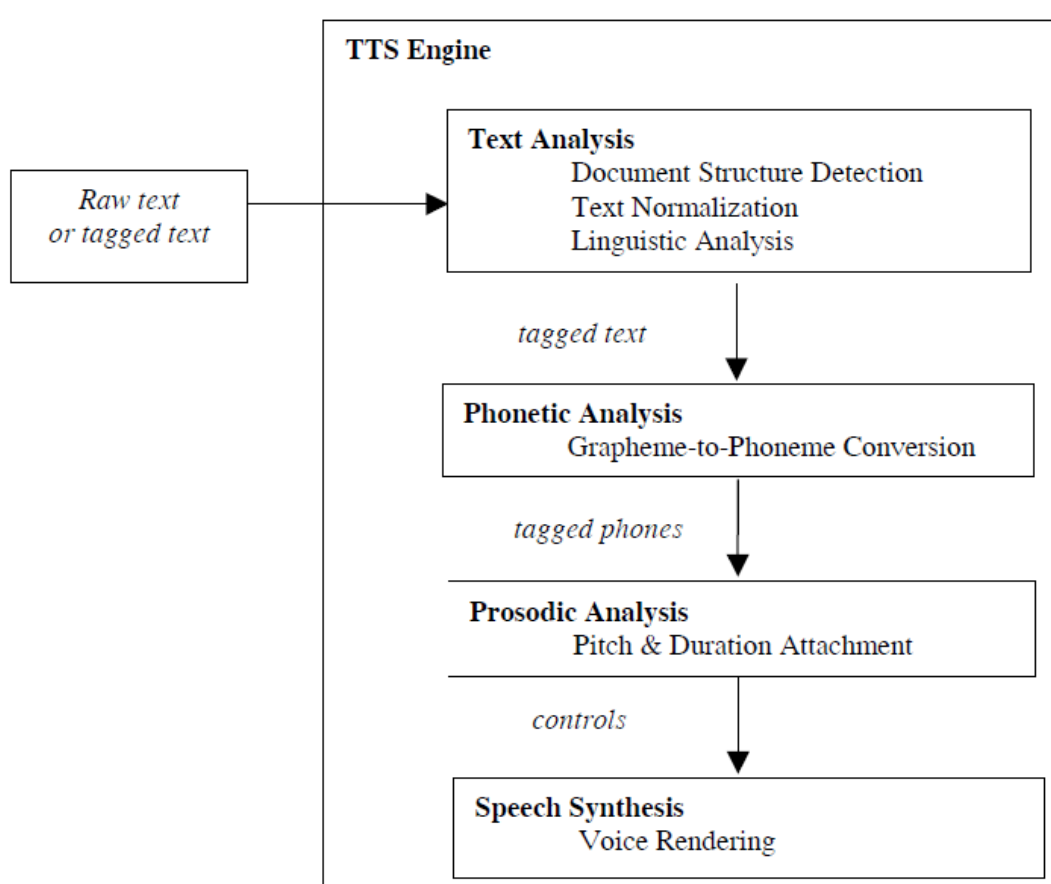
- استفاده در نرم‌افزارهای دیگر: مترجم گفتار به گفتار، OCR

- فشردن سازی گفتار (کد کردن)



کاربردهای زبان‌شناسی رایانشی: گفتاری ...

○ سنتز گفتار ...





کاربردهای زبان‌شناسی رایانشی: گفتاری ...

○ سنتز گفتار

FarsiReader.com

The screenshot shows the FarsiReader.com website interface. At the top, there is a navigation bar with a search bar and a 'درباره ما' (About Us) link. Below the navigation bar, there is a main content area with a large text input field and a 'بخوان' (Read) button. The interface includes various controls for text-to-speech synthesis, such as a 'دموی متن به گفتار فارسی آریانا' (Ariana Persian text-to-speech demo) section, a 'مرد' (Male) voice selection, and a 'خواندن علامت نگارشی' (Read punctuation) checkbox. There is also a progress bar and a 'تنظیمات پیش فرض' (Default settings) button. On the right side, there is a sidebar with a logo for 'عصر گویش پرداز' (Eصر گویش پرداز) and a list of services including 'وبسایت های ما' (Our websites), 'نرم افزار تایپ گفتاری نویسا' (Noyisa speech-to-text software), 'مجله دیجیتال عصر گویش' (Eصر گویش digital magazine), and 'فروشگاه آنلاین فارسی ویب' (Farsi web online store).



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

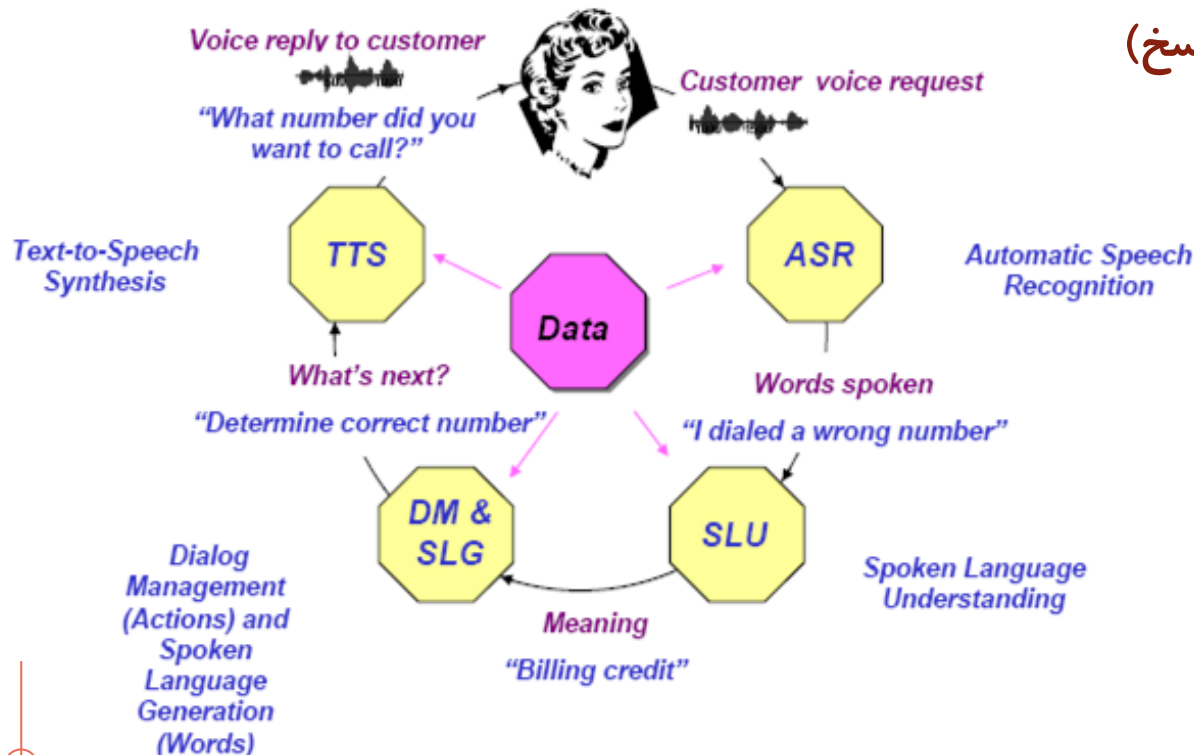




کاربردهای زبان‌شناسی رایانشی: گفتاری ...

○ دیالوگ و فهم گفتار (Speech Dialogue/Understanding)

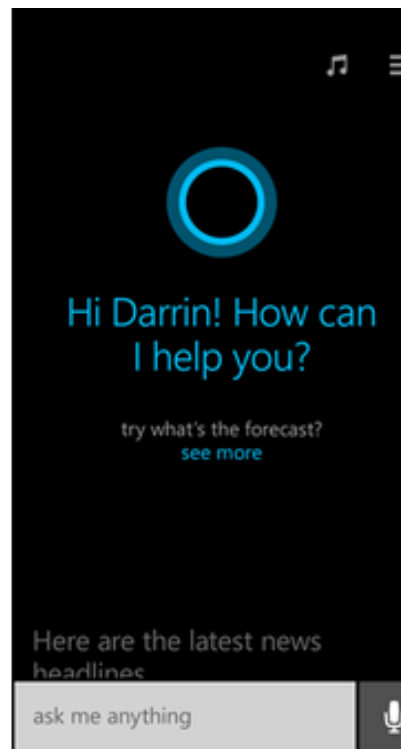
- تشخیص گفتار (ورود اطلاعات)
- درک (فهم) گفتار
- تولید جمله (تولید پاسخ)
- تولید گفتار (خواندن پاسخ)



کاربردهای زبان‌شناسی رایانشی: گفتاری ...

○ دیالوگ و فهم گفتار ...

- iPhone روی Apple Siri
- 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



Microsoft
Tell me.

Say it. Get it.

or call Tellme at
1-800-555-TELL

+1-800-555-8355

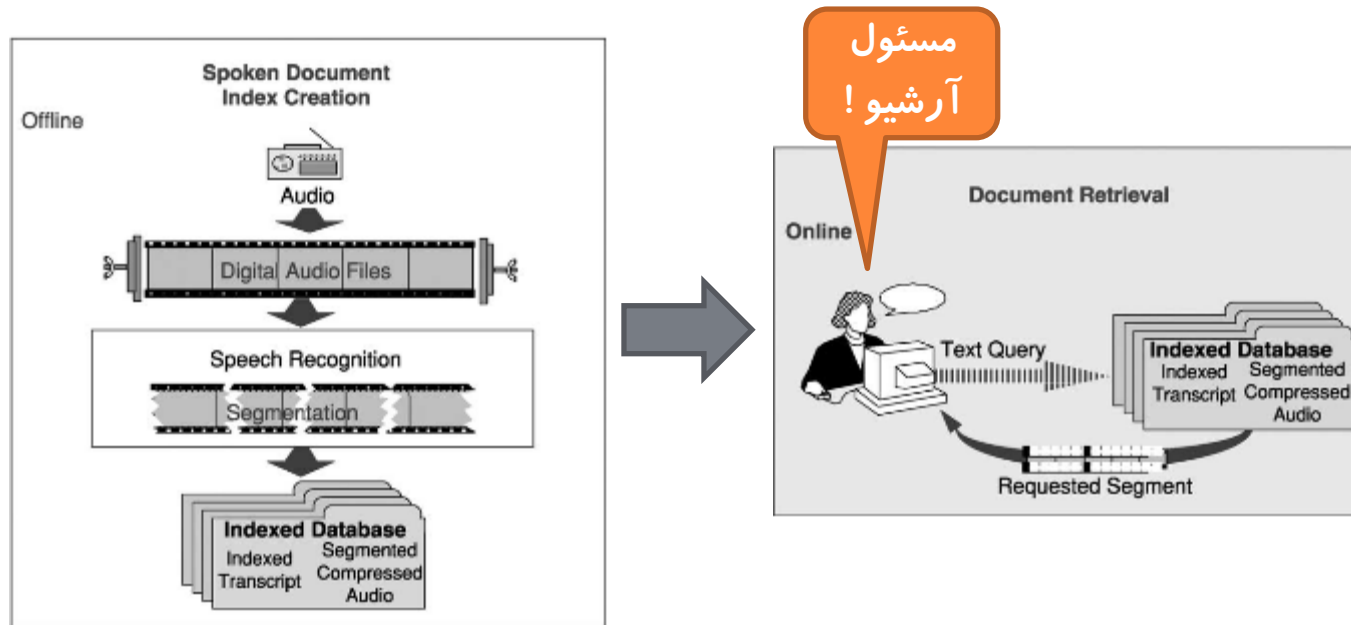
کاربردهای زبان‌شناسی رایانشی: گفتاری ...

○ بازیابی اطلاعات گفتاری (Spoken Document Retrieval)

• بازیابی (جستجو) کلمه‌ها و عبارات در یک مجموعه گفتاری

○ مانند آرشیوهای صدا و سیما

○ فایل‌های صوتی اینترنت



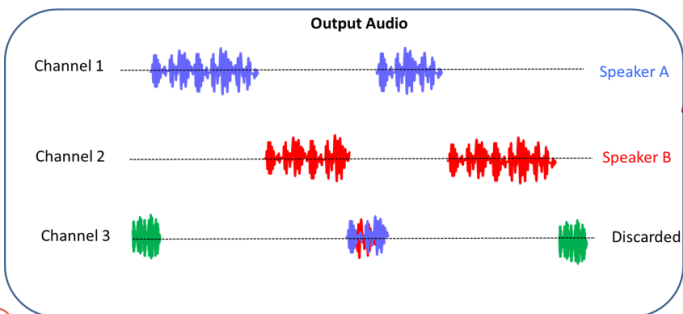
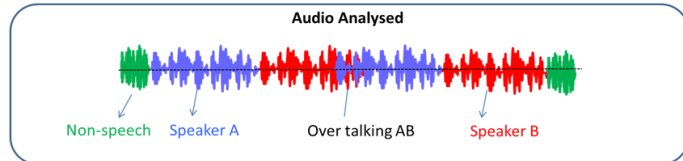
کاربردهای زبان‌شناسی رایانشی: گفتاری ...

○ 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 املائی (Spell Checker) ...

- تشخیص غلط‌های املائی واژه‌ها

- اصلاح غلط‌های املائی واژه‌ها

- ارائه لیستی از واژه‌های صحیح پیشنهادی

- اصلاح غلط‌های فاصله‌گذاری و علائم سجاوندی

- اصلاح کاربرد نابجای فاصله به جای شبه‌فاصله



املائی

Word در

One of the essential requirements of artificial 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 املائی (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

گرامری

در Word

One of the essential requirements of artificial 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

Google

All Images Videos News More Settings Tools

About 63,300 results (0.27 seconds)

زبان‌شناسی رایانشی - ویکی‌پدیا، دانشنامهٔ آزاد
<https://fa.wikipedia.org/wiki/رایانشی> Translate this page
زبان‌شناسی رایانشی (Computational linguistics) حوزه‌ای میان‌رشته‌ای است که می‌کشد با بهره‌گیری از روش‌های آماری و قاعده‌بنیاد (rule-based) به مدل‌سازی زبان طبیعی ...
۲ مورد کاربرد - ۱ خانسگاه

دانشگاه صنعتی شریف - خانه - زبان‌شناسی رایانشی حوزه‌ای میان‌رشته‌ای ...
www.sharif.ir/home?... Translate this page
نویسنده بردنام‌ریزی و راه‌اندازی رشته زبان‌شناسی رایانشی در مرکز زبان‌ها و زبان‌شناسی دانشگاه صنعتی شریف، در - Nov 12, 2011
سال 1388 شکل گرفت. ایجاد این ...

سرفصل دروس رشته زبان‌شناسی رایانشی - دانشکده علوم و فنون نوین ...
fnst.ut.ac.ir/ Translate this page
سرفصل دروس رشته زبان‌شناسی رایانشی. نوع دوره: کارشناسی ارشد؛ طول دوره: 2 سال (4 ترم)؛ با توجه به بین‌رشته‌ای بودن این دوره و لزوم گذراندن دروس جزئی ...

کارشناسی ارشد زبان‌شناسی رایانشی - دانشکده علوم و فنون نوین - Faculty ...
fnst.ut.ac.ir/ Translate this page
دام گرایش: کارشناسی ارشد زبان‌شناسی رایانشی مسئول رشته: آقای دکتر هادی ویسی اهداف گرایش: زبان نه تنها بخشی از دانش بشری ...



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



Requirements





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 شبکه‌های اجتماعی

The screenshot shows a plagiarism detection tool with two text boxes for comparison. The left box contains text about 'ABC College for Women' and the right box contains a similar text. A 'Flip' button is between them. Below the text boxes, a table shows keyword analysis results:

Keyword Phrase	Occurrences	Density
Higher Education. Established	1	2.31 %
Testing Plagiarism ABC	1	2.31 %
Plagiarism ABC College	1	2.31 %

At the bottom, a 'Statistics' section indicates '30 % Duplicate Found!' and '30 Matches Detected'.



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): نیامندی‌ها

- 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 ۴۰٪ 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)

نظرات کاربران در مورد یکی از محصولات در سایت دیجی کالا

محمدرضا دریانورد (1393/10/27):
شانس ما رو . دیشب که اومدم خرید رو انجام بدم ناموجود شد.
دیجی کالا لطفا دوباره این کالا رو با متعلقاش بیار .منتظرم

همهان (1393/10/25):
من این گوشی رو تازه خریدم، چنتا از برنامه هاشو به روز کردم، اما بعد از به روز کردن، صدای قابل های صوتی و تصویرک گوشیم قطع شد! بعد وقتی امر تماسی با کسی می گیرم یا کسی باهام تماس می گیره صداها نمی شنوه و منم صداشو نمی شنوم، اما وقتی بلندگو رو روشن می شنوم مشکل صدا حل میشه. میشه بگید مشکل از کجاست؟

حواد کاظمی (1393/10/25):
من این گوشی رو دو ماهه خریدم، متاسفانه مشکلی برام پیش اومده، ک شب که زدم تو شارژ صبح بیدار شدم دیدم گوشی خاموش شده و دیگه هم روشن نشد. کسی میدونه مشکل چیه؟ تو رو خدا باید الان چیکار کنم؟

رها سنایی (1393/10/23):
من الان حدود 1ساله این گوشیه دارم . ازشم راضیم .
از عکسش چون خردم تنظیم میکنم اما فیلمش بخاطر فوکوس اتوماتیک باید گوشی خیلی دقیق و آروم حرکت داده بشه اگه نه وسط فیلم لنز دنبال چهره ها میگردد و تار میشه .
الان اندروید 4.4.4دارم اما از این ورژن اندروید راضی نیستم چون وقتی یک هفته گوشی خاموش نشه سرعت عملش میاد پایین (قبلا اینو تو گوشی های سامسونگ دیده بودم) امیدوارم آپدیت بعدی مشکلش حل شه .
در مورد صفحه نمایش خیلی مراقب باشی چیز نوک تیز اطرافش نکشید یا روی لپه ها ضربه وارد نکند چون من تا 2تا از فامیل هامون صفحهشون شکست و خیلی خرجشون شد .



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)



====> who is the current president of iran

Iran



Executive branch:

chief of state: Supreme Leader Ali Hoseini-KHAMENEI (since 4 June 1989)

head of government: President Hasan Fereidun RUHANI (since 3 August 2013); First Vice President Eshaq JAHANGIRI (since 5 August 2013)



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 تحلیل‌های مختلف

- داغ‌ترین مطالب
- نحوه نشر اطلاعات
- نظرات کاربران در مورد پست‌ها



• نیازمندی‌ها

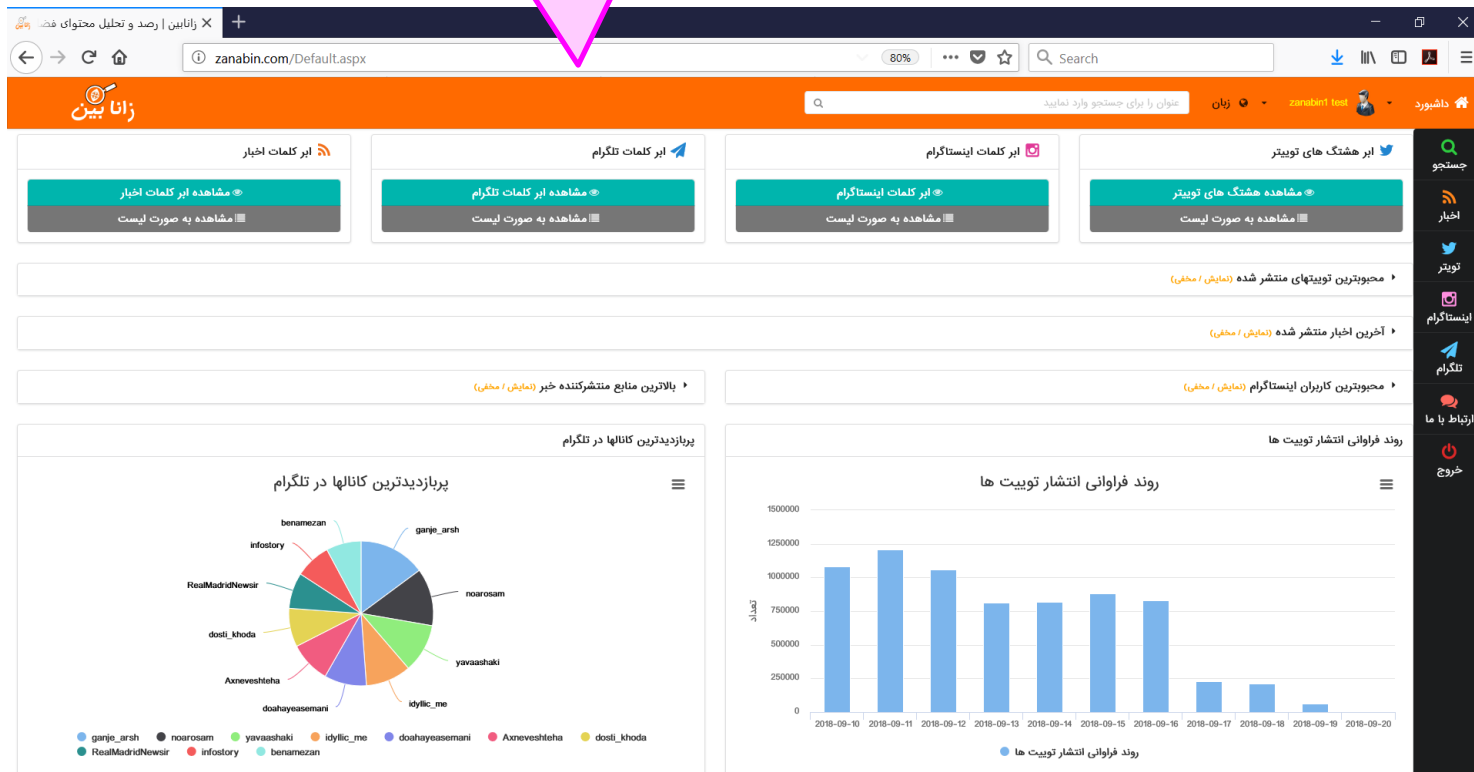
- نرمال سازی و واحدسازی
- تشابه یابی اسناد
- بازشناسی موجودیت‌های اسمی (NER)
- قطعه‌بندی (Chunking)



کاربردهای زبان‌شناسی رایانشی: متنی ...

○ پایش و تحلیل فضای مجازی

ZanaBin.com

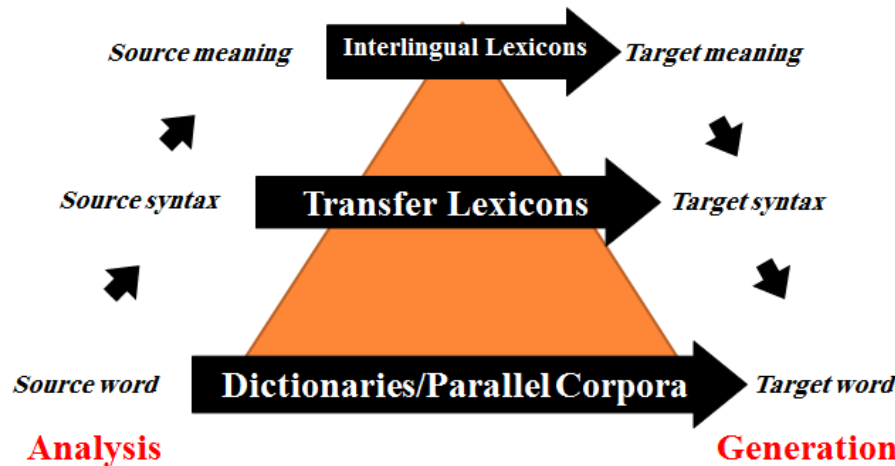




کاربردهای زبان‌شناسی رایانشی: متنی ...

○ ترجمه ماشینی (Machine Translation) ...

- ترجمه متن از زبانی به یک یا چند زبان دیگر



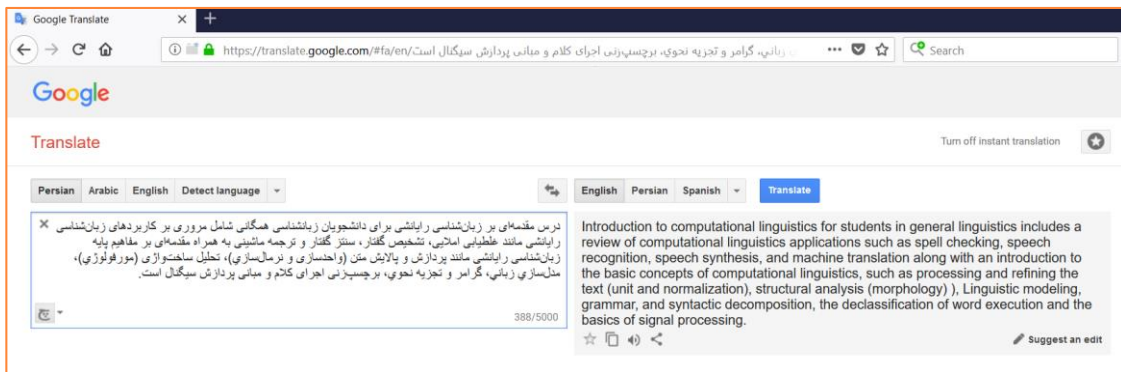
• روش‌ها

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



کاربردهای زبان‌شناسی رایانشی: متنی ...

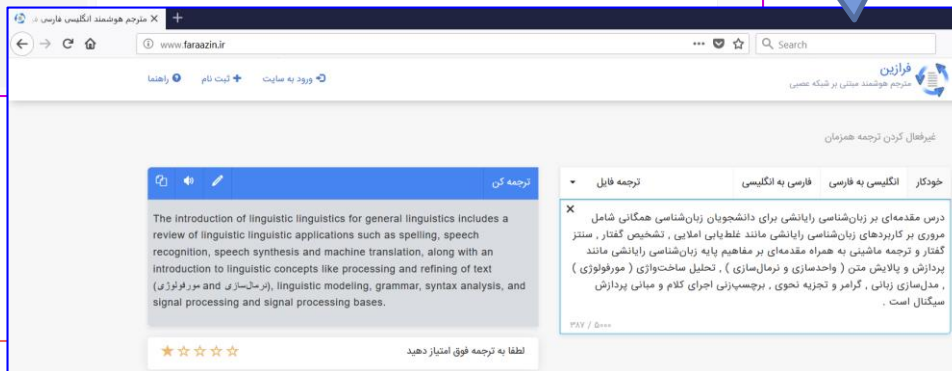
ترجمه ماشینی (Machine Translation) ...



targoman.com



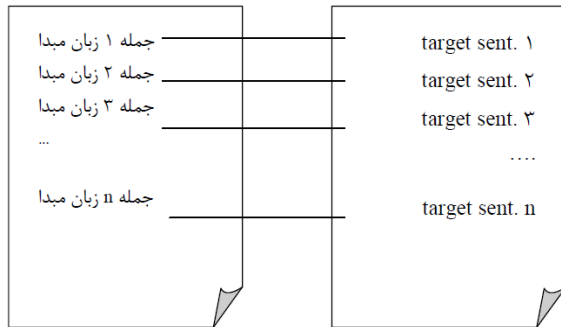
faraazin.ir





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

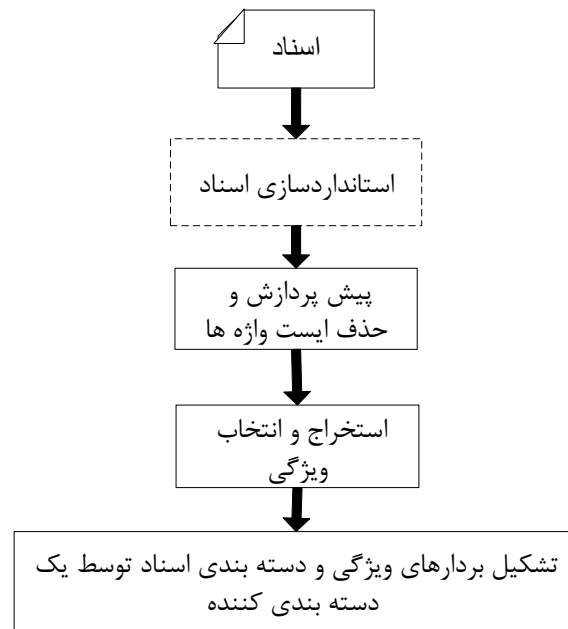
REQUIRED



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)
- الگوریتم‌های 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

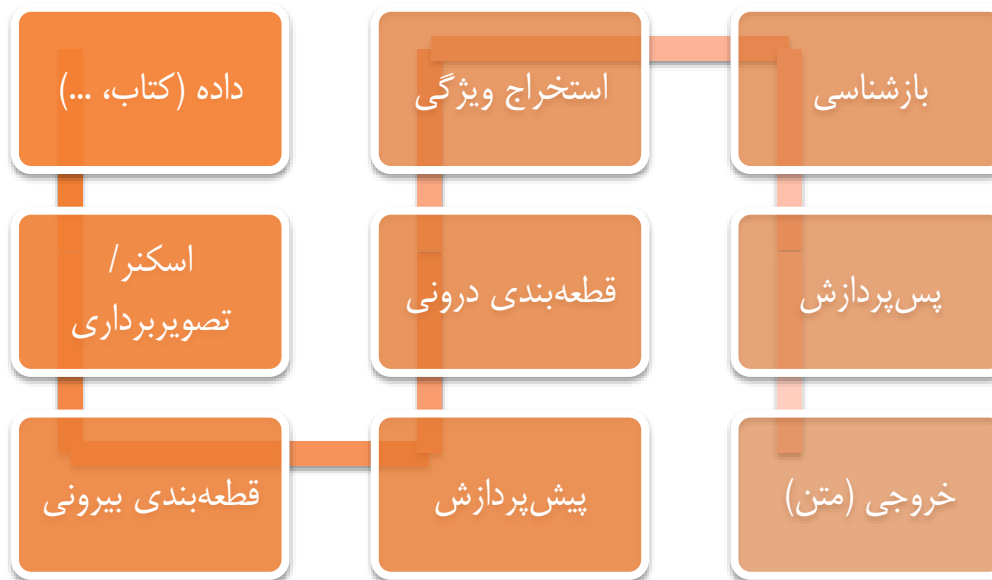
- ویژگی‌ها: غنای واژگان، طول کلمات، تعداد علائم سجاوندی، طول جمله، تعداد صفات/ضمایر و ...
- داده دارای برچسب دسته (نوع/موضوع سند)
- الگوریتم‌های یادگیری



کاربردهای زبان‌شناسی رایانشی: تصویری ...

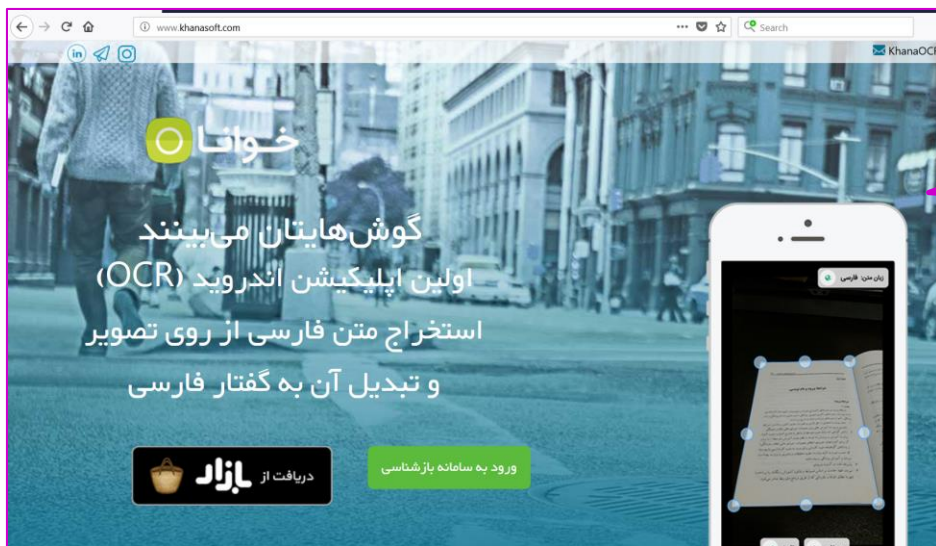
○ نویسه‌خوان نوری (OCR: Optical Character Recognition) ...

- 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

کلمات کلیدی: اثر انگشت، طبقه بندی، شباهت
 با توجه به کاربردهای شناسایی و تشخیصی و
 روشهایی با دقت بالا و کم هزینه جهت رسیدن به بین‌المللی شدن و افزایش اعتبار
 انگشت به علت دارا بودن ویژگیهایی همچون تغییر ناپذیری، منحصر بفرد بودن، قابلیت
 طبقه بندی و غیر تهجمی بودن، جایگاه ویژه ای دارد. در یک سیستم خودکار
 شناسایی افراد که بر اساس اثر انگشت کار می کند، لازم است که تصویر اثر انگشت
 ورودی با تک تک تصاویر موجود در مجموعه تصاویر تطابق داده شود. با توجه به اینکه
 معمولاً تعداد این تصاویر خیلی زیاد است زمان جستجو و شناسایی تصویر ورودی به
 طور قابل ملاحظه ای بالا می رود. در این راستا برای کاهش زمان جستجو، طبقه بندی
 تصاویر اثر انگشت در کلاسهای مختلف ضروری به نظر می رسد. واضح است که هر چه
 تصاویر در کلاسهای بیشتری طبقه بندی گردند، زمان جستجو و تطابق به مراتب کاهش
 می یابد. University of Tehran, Tehran, Iran Feb. ۲۱ ۲۰۰۵

مجموعه مقالات علمی-پژوهشی نشریه علمی-پژوهشی دانش‌های بنیادین
 شماره ۱۴، بهار ۱۳۹۵، صفحات ۴۸-۳۵
 © ۱۳۹۵، بهار ۱۳۹۵، شماره ۱ (۱۴)، اضافه شد. (۰.۹۷ تاییه)
 © ۱۳۹۵، بهار ۱۳۹۵، شماره ۱ (۱۴)، اضافه شد. (۱.۱۲ تاییه)
 © ۱۳۹۵، بهار ۱۳۹۵، شماره ۱ (۱۴)، اضافه شد. (۲۲.۸۱ تاییه)
 © ۱۴۰۰، بهار ۱۴۰۰، شماره ۱ (۱۴)، اضافه شد. (۸.۴۲ تاییه)
 © ۱۴۰۰، بهار ۱۴۰۰، شماره ۱ (۱۴)، اضافه شد. (۰.۹۷ تاییه)
 © ۱۴۰۰، بهار ۱۴۰۰، شماره ۱ (۱۴)، اضافه شد. (۱۹.۴۵ تاییه)

فهرست بندی تصاویر اثر انگشت با شبکه های عصبی
 سید الهه ارجمن الرحیم
 دانشیار گروه مهندسی رایانشی، دانشکده مهندسی کامپیوتر، دانشگاه تهران
 ghavami@ut.ac.ir
 ۱۴
 University of Tehran, Tehran, Iran
 Feb. 2005, Vol. 14



کاربردهای زبان‌شناسی رایانشی: تصویری ...

○ نویسه‌خوان نوری (OCR): نیازمندی‌ها

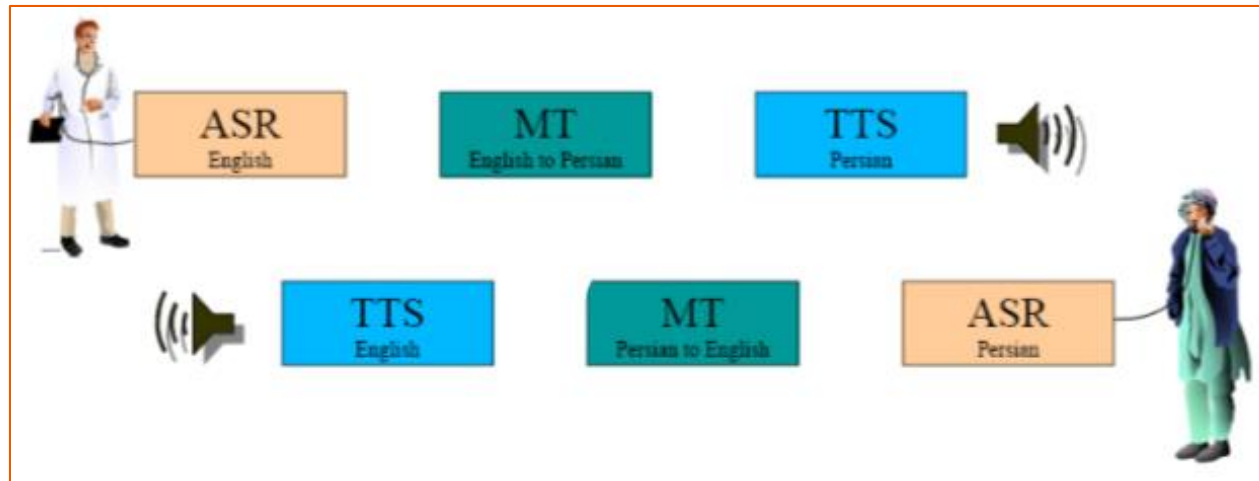
- دادگان تصویری دارای برچسب متنی معادل
- لغت‌نامه (برای پس پردازش)

- 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



رایگان
با پرداخت درون‌برنامه‌ای

خوانا: تبدیل تصویر به متن و گفتار

کاربردی
۴۱,۰۰۰

۲۱۸ مگابایت
1,190

۱۰۲-۶۸۱۷۷۷-۶۳-۵-۳۰۲

دسته
نصب‌های فعال
حجم
نسخه
شامد



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 واژگی (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

- استخراج 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

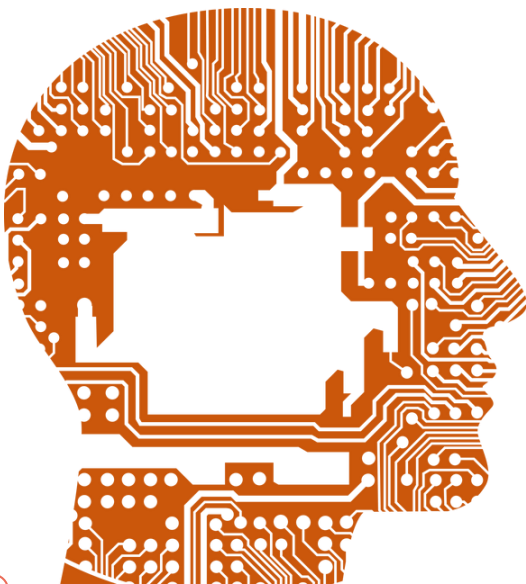


روش‌های زبان‌شناسی رایانشی

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)
- میدان‌های تصادفی شرطی (CRF: Conditional Random Fields)

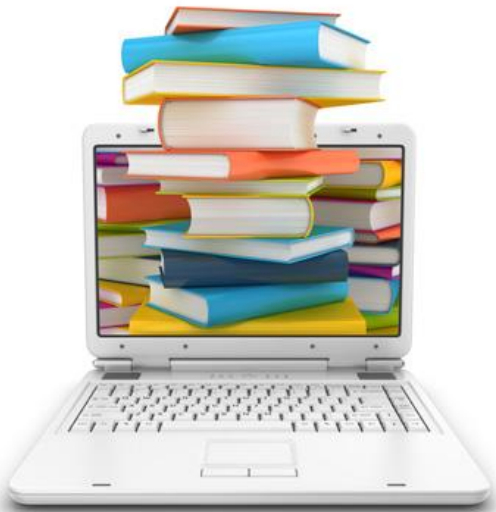




دادگان‌های زبان‌شناسی رایانشی ...

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 املائی
- 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

speech

sentence