ابزارهای پردازش زبان طبیعی

چهارمین کارگاه سالانه آزمایشگاه فناوری وب

احمد استیری
دانشجوی کارشناسی ارشد
زمستان ۱۳۹۱

مقدمه

- رشد چشمگیر اسناد منتشر شده در وب
- نیاز اساسی به نگهداری، دسته‌بندی، بازیابی و پردازش آنها

✓ **توجه بیش از پیش به پردازش زبان طبیعی توسط رایانه**

در بسیاری از سیستم‌های مرتبط با متن:

- استفاده از تکنیک‌های پردازش زبان طبیعی

✓ **بهبود چشم‌گیر در دقت و صحت نتایج خروجی سیستم**

مقدمه (ادامه)

- مهم‌ترین هدف زبان‌شناسی رایانه‌ای:

ارائه ابزارها و نرم‌افزارهایی جهت تحلیل بر روی متن و یا پاره‌های زبانی.

- لزوم پیش‌پردازش و در اختیار داشتن ابزارهای پایه‌ای پردازش متن قبل از هر گونه انجام عملیات و یا اعمال الگوریتم بر روی متون.
- ابزارهای اولیه‌ی پردازش متن قوی‌تر؛

✓ نتایج حاصل از الگوریتم‌های طراحی شده در گام‌های آتی و سیستم‌های مرتبط، قابل اطمینان‌تر.

مقدمه (ادامه)

- ابزار نرمال‌ساز یا یکسان‌ساز
- ابزار تشخیص‌دهنده جملات
- ابزار تشخیص دهنده لغات
- ابزار ریشه‌یاب
- ابزار برچسب‌زن اجزای واژگانی کلام (POS)
- ابزار پاسر (Parser)
- ابزار برچسب‌زن معنایی کلام (SRL)
- شبکه واژگان
- ...

نرمال‌ساز (Normalizer)

- تفاوت پردازش زبان فارسی از جهات مختلفی با پردازش زبان انگلیسی.
- در زبان فارسی بعضی از حروف به هم چسبیده‌اند، بعضی از حروف جدا از هم نوشته می‌شوند، بعضی از کلمات یکپارچه‌اند، بعضی از کلمات با فاصله یا نیم‌فاصله به دو یا چند بخش تقسیم می‌شوند.
- علاوه براین حروفی مانند "ی" و "ک" در بعضی از نوشته‌ها با نسخه عربی مانند "ی عربی" یا "ک عربی" نوشته می‌شوند.

➢ بوجود آمدن مشکلاتی در مقایسه کلمات.

نرمال‌ساز (Normalizer)

- یکسان‌سازی همه‌ی نویسه‌های (کاراکترهای) متن با جایگزینی با معادل استاندارد آن.
- اصلاح و یکسان‌سازی نویسه‌ی نیم‌فاصله و فاصله در کاربردهای مختلف آن و همچنین حذف نویسه‌ی «ـ» مورد استفاده برای کشش نویسه‌های چسبان.
- اصلاح فاصله‌ها و نیم‌فاصله‌های موجود در متن برای علاماتی نظیر "ها" و "ی" غیرچسبان در انتهای لغات و همچنین پیشوندها و پسوندهای فعل‌ساز نظیر "می"، "ام"، "ایم"، "اید" و موارد مشابه جهت استفاده در فازهای بعدی.

جدا کننده جملات (Sentence Splitter)

- تشخیص مرز جملات با استفاده از علامت‌های ".“، ”؛“، ”!“، ”؟“، ”?“، ”:“ و بکارگیری برخی دستورات گرامری زبان فارسی و در نظرگرفتن برخی لغات آغاز کننده‌ی جملات.
- اهمیت تشخیص صحیح جملات با توجه به پایه بودن جمله در بسیاری از پردازش-های زبانی.
- در نظر گرفتن مواردی نظیر حالت مخفف کلمات نظیر ه.ق. ، ... و سایر موارد
- نمونه‌های انگلیسی این ابزار: OpenNLP، Stanford NLP، NLTK و Freeling

جداکننده کلمات (Tokenizer)

- تشخیص لغات با استفاده از علامت‌های فضای خالی، ”،“، ”,“، ”-“ و در نظر گرفتن اصلاحات اعمال شده در مورد پیشوندها و پسوندها در فاز قبلی.
  - کتاب ها
  - می روم
  - دانش آموز
  - ...
- نمونه‌های انگلیسی این ابزار: Flex، JLex، JFLex، ANTLR، Ragel و Quex

حذف‌کننده کلمات ایست (Stop Word Remover)

- حذف ایست‌واژه‌ها یا Stop Wordها از متن.
- ایست واژه‌ها لغاتی هستند که علی‌رغم تکرار فـروان در مـتن، از لحاظ معنایی دارای اهمیـت کمـی هسـتند مثـل "اگـر"، "و"، "ولی"،"که" و غیره.
- حذف ایست‌واژه‌ها:
  - کاهش بار محاسبات
  - افزایش سرعت
  - بهبود نتایج پردازش در اکثر موارد

حذف‌کننده کلمات ایست (Stop Word Remover)

- نمونه‌ای از ایست‌واژه‌ها در زبان فارسی

اکنون	است	زیرا	برای	اینک	اگر
البته	شد	چون	بالاخره	اینطور	بعدا
اما	کرد	باید	اینقدر	بدون	حدودا
از	باشد	حالا	بله	با	خصوصا
که	هست	حتی	زود	حتما	انگار

ریشه‌یاب (Stemmer)

- فرآیند تبدیل کلمات به فرم ریشه‌ای و پایه‌ای آنها.
  - "*دانش‌آموز*" - "*دانشجو*" - "*دانشگاه*" ---»»» "*دان*"

- بدست آوردن ریشه کلمات با حذف پسوندها و پیشوندها.
- منظور از ریشه، دقیقا ریشه‌ی کلمات که در زبان‌شناسی استفاده می‌شود، نیست. بلکه منظور از ریشه، یک نماینده برای کلماتی است که از لحاظ معنایی و نحوی در یک حوزه قرار می‌گیرند.

ریشه‌یاب (Stemmer)

- دسته‌بندی الگوریتم‌های ریشه‌یابی:
  - ریخت‌شناسی و بر پایه قانون
  - استفاده از فرهنگ لغت
  - روش‌های ترکیبی

- رایج‌ترین الگوریتم در زبان انگلیسی: الگوریتم پورتر (Porter)
- نمونه‌های دیگر الگوریتم‌های ریشه‌یابی: الگوریتم *کراوتز* (Krovetz) در انگلیسی و الگوریتم کاظم تقوا در فارسی.

برچسب‌زننده اجزای واژگانی کلام (POS)

- در دستور زبان، اجزای واژگانی کلام یا بخش‌های سخن، طبقه-بندی‌هایی زبانی از کلمات هستند که رفتار نحوی یک قسمت از جمله را بیان می‌دارند.
- مهم‌ترین بخش‌های سخن در زبان فارسی: اسم، ضمیر، صفت، قید و حرف اضافه.
- فرآیند نشانه‌گذاری لغت در یک متن است که این نشانه، بیانگر وجه آن جزء از کلام می‌باشد.
- شکل ساده شده‌ی این موضوع: تشخیص نوع لغت از لحاظ *اسم*، *فعل*، *صفت* و *قید* در مدارس.

برچسب‌زننده اجزای واژگانی کلام (POS)

- نمونه‌ای فرضی از یک مجموعه برچسب (Tagset) برای زبان انگلیسی

- # - Pound sign
- $ - Dollar sign
- '' - Close double quote
- `` - Open double quote
- ' - Close single quote
- ` - Open single quote
- , - Comma
- . - Final punctuation
- : - Colon, semi-colon
- -LRB- - Left bracket
- -RRB- - Right bracket
- CC - Coordinating conjunction
- CD - Cardinal number
- DT - Determiner
- EX - Existential there
- FW - Foreign word
- IN - Preposition
- JJ - Adjective
- JJR - Comparative adjective
- JJS - Superlative adjective
- LS - List Item Marker
- MD - Modal
- NN - Singular noun
- NNS - Plural noun
- NNP - Proper singular noun
- NNPS - Proper plural noun
- PDT - Predeterminer
- POS - Possesive ending
- PRP - Personal pronoun
- PP$ - Possesive pronoun
- RB - Adverb
- RBR - Comparative adverb
- RBS - Superlative Adverb
- RP - Particle
- SYM - Symbol
- TO - to
- UH - Interjection
- VB - Verb, base form
- VBD - Verb, past tense
- VBG - Verb, gerund/present participle
- VBN - Verb, past participle
- VBP - Verb, non 3rd ps. sing. present
- VBZ - Verb, 3rd ps. sing. present
- WDT - wh-determiner
- WP - wh-pronoun
- WP$ - Possesive wh-pronoun
- WRB - wh-adverb

برچسب‌زننده اجزای واژگانی کلام (POS)

- نمونه‌ای از یک مجموعه برچسب (Tagset) برای زبان فارسی

ADV	Adverb	قید
AJ	Adjective	صفت
CL	Classifier	شاخص
CONJ	Conjunction	حرف ربط
DET	Determiner	حرف تعریف
INT	Interjection	حرف صوت
N	Noun	اسم
NUM	Number	عدد
P	Preposition	حرف اضافه
POSTP	Postposition (را) پسین	حرف اضافه
PRO	Pronoun	ضمیر
PUNC	Punctuation	جدا کننده
RES	Residual:Arabic and Latin words, etc	متفرقه
V	Verb	فعل

برچسب‌زننده اجزای واژگانی کلام (POS)

- نمونه‌ی فارسی این ابزار:
  - ابزار آزمایشگاه آقای دکتر بیجن‌خان

- نمونه‌های انگلیسی آن:
  - Illinois Part Of Speech Tagger
  - Stanford POS Tagger

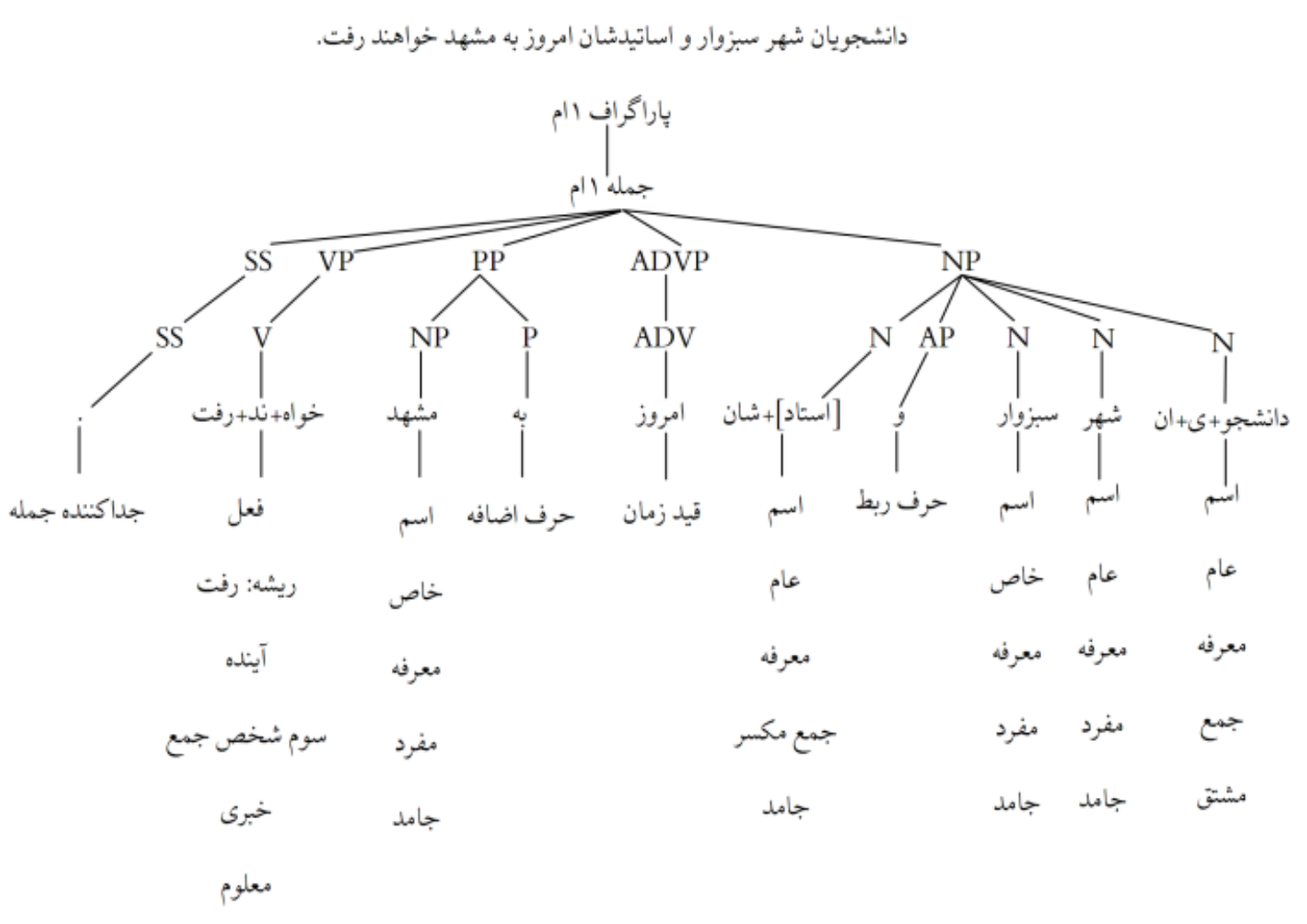
پارسر (Parser)

- بهره‌گیری از دستورات گرامری زبان.
- تشخیص گروه‌های تشکیل دهنده جملات متن، تجزیه‌ی نحوی و تشکیل درخت تجزیه‌ی جملات.
- تجزیه و تحلیل جمله و شکستن آن به اجزای تشکیل دهنده مانند *گروه‌های اسمی*، *فعلی*، *قیدی* و غیره.
- نقش اساسی در طراحی و یا افزایش دقت سایر ابزارهای پردازش متن.

پارسر (Parser)

- تقسیم‌بندی اجزای هر جمله در قالب گروه‌های اسمی، فعلی، حرف اضافه‌ای و ...
- روابط بین گروه‌ها
- امکان وجود زیرگروه‌ها در هر گروه
- تقسیم‌بندی‌های سلسله مراتبی
  - ❖درخت تجزیه
- نمایان‌سازی ساختار نحوی یک جمله بر اساس برخی روابط گرامری موجود در آن به شکلی ساده و قابل فهم برای کسانی که دانش عمیق زبان‌شناسی ندارند.

پارسر (Parser)



برچسب‌زنی نقش معنایی کلمات (SRL)

- ابزاری برای تشخیص نقش گرامری کلمه در جمله.
- استخراج نقش‌های معنایی جملات نظیر فاعل، مفعول مستقیم، مفعول غیرمستقیم، فعل و ...
- نقش اساسی در پردازش‌های زبانی.
- کاربرد: بسیاری از حوزه‌های دیگر پردازش زبان طبیعی (NLP) از قبیل ترجمه ماشینی، خطایاب و شباهت معنایی و ...
- نمونه‌های انگلیسی این ابزار: OpenNlP، Illinois SRL، Swirl و LTHSRL

Named entity recognition

- ابزاری برای تشخیص اسامی و نوع آنها اعم از اسامی افراد، اماکن، مقادیر عددی و ... .
- روش‌های تشخیص اسم:
  - مراجعه به لغت‌نامه
  - مراجعه به شبکه واژگان
  - در نظر گرفتن ریشه‌ی کلمه
  - استفاده از قواعد نحوی ساخت‌واژه
  - ...

Named entity recognition

- پس از تشخیص اسم‌ها:
  - تشخیص نوع اسم با استفاده یک لغت‌نامه از اسامی افـراد، مکـان‌هـا، مقادیر عددی و ...

- نمونه های انگلیسی این ابزار:
  - Stanford NER
  - Illinois NER

شبکه واژگان

- شبکه‌ای متشکل از هزاران مفهومی که بوسیله روابط معنایی به هم مرتبطند.
- هر مفهوم، نشان‌دهنده‌ی مجموعه‌ای انتزاعی از عناصری می‌باشد که بر اساس مختصه‌های مشترکشان، یک گروه را تشکیل می‌دهند.
- در شبکه واژگان، ابتدا لغات در یکی از دسته‌های اسم، فعل، صفت، و قید قرار گرفته و سپس لغات هر یک از این دسته‌ها در گروه‌های هم‌خانواده‌ی خود قرار می‌گیرند.

شبکه واژگان

- هر یک از این گروه‌های هم‌خانواده از یک یا چند لغت تشکیل می‌شود، که یک مفهوم مشخص را عنوان می‌کنند و لغات تشکیل‌دهنده این گروه می‌توانند به جای یکدیگر در یک متن استفاده شوند و توسط یکسری روابط معنایی با سایر گروه‌ها مرتبط می‌شوند.
- روابط معنایی بین گروه‌های هم‌خانواده بر حسب نوع گروه (اسم، فعل، صفت و قید) متفاوت است.
- در واقع شبکه واژگان دارای سه پایگاه داده می‌باشد: یکی برای اسامی، یکی برای افعال و یکی نیز مشترکاً برای صفات و قیود.

شبکه واژگان

- شبکه واژگان شامل مجموعه‌ی مترادف‌های کلمات می‌باشد که از آن به عنوان "Synsets" یاد می‌شود.
- هر Synset یک مفهوم و یا یک معنی از گروهی از کلمات، را شامل می‌شود.
- Synsetها روابط معنایی متفاوتی چون مترادف، متضاد، ابرمفهوم، زیرمفهوم (IS-A)، جزئیت (Part of)، شمول (Has-A)را دربر می‌گیرند.
- شبکه واژگان هم‌چنین تعاریف متنی از مفاهیم را فراهم می‌سازد (Glossary) که شامل تعاریف و مثال‌ها می‌باشد.

شبکه واژگان

- بخشی از روابط معنایی موجود در شبکه واژگان

رابطه‌ی معنایی	دسته‌ی نحوی	مثال
ترادف (Synonymy)	فعل اسم صفت قید	Rise, ascend Pipe, tube Sad, unhappy Rapidly, speedly
تضاد (Antonymy)	فعل اسم صفت قید	Rise, fall Top, bottom Wet, dry Rapidly, slowly
زیرمفهوم (Hyponymy)	اسم	Sugar, maple, maple maple, tree tree, plant
جزئیت (Meronymy)	اسم	Brim, hat gin, martini ship, fleet

شبکه واژگان

- از نمونه‌های فارسی آن:
  - شبکه واژگان فارس‌نت
  - فردوس‌نت

- از نمونه‌های انگلیسی آن:
  - Princeton Wordnet
  - EuroWordnet

با سپاس از صبر، حوصله و توجه شما