پردازش زبان طبیعی
NLP

تهیه و تنظیم : مصطفی کریمی منش

زبان طبیعی

- **زبان طبیعی** زبانی است هر انسانی از محیط آموخته و در تعاملات خود با دیگران بکار می برد (می نویسد و صحبت می کند)
- به عبارت رسمی تر **زبان** یک سیستم قراردادی منظم از آواها یا نشانه‌های کلامی یا نوشتاری بوده که توسط انسانهای متعلق به یک گروه اجتماعی یا فرهنگی خاص برای نمایش و فهم ارتباطات و اندیشه‌ها به کار برده می‌شود.

- زبان های طبیعی مختلف و زیادی وجود دارند

زبان طبیعی

- ممکن است که فرم گفتاری و نوشتاری زبان ها متفات باشند و همچنین از هم مستقل باشند.
- در مقابل زبان طبیعی، **زبانهای مصنوعی**، زبانهایی هستند که بوسیله انسانها به منظور تعامل با فناوریهای خود ساخته اند. مانند زبانهای برنامه نویسی
- ویژگی خاص زبان این است که می‌توان آن را دو بار تجزیه کرد. در تجزیه اول کلام را می‌توان به واحدهای معنایی کوچک‌تر تجزیه کرد. به این واحدها تکواژ می‌گویند.

زبان طبیعی

- در مرحله دوم هر تک‌واژ را می‌توان به واحدهای کوچک‌تر آوایی تقسیم کرد که از لحاظ کاربرد آوایی بسیطاند و از نظر معنایی فاقد معنا هستند . به این جزءهای کوچک‌تر واج می‌گویند.

زبان شناسی

- **زبان‌شناسی** (Linguistics) علمی است که به مطالعه و بررسی روشمند زبان می‌پردازد.
- زبان‌شناسی می‌کوشد تا به پرسش‌هایی همچون
  - زبان چیست؟
  - زبان چگونه عمل می‌کند و از چه ساخت‌هایی تشکیل شده‌است؟
  - انسان‌ها چگونه با یکدیگر ارتباط برقرار می‌کنند؟
  - زبان بشر چگونه تکامل یافته‌است؟
  - ویژگی‌های مشترک زبان‌های جهان کدامند؟
  - انسان چگونه می‌نویسد و از چه راهی زبان نانوشتاری را واکاوی (تحلیل) می‌کند؟
  - و...

زبان شناسی

- در زبان‌شناسی، ابعاد مختلف زبان در قالب حوزه‌های:
  - صرف
  - نحو،
  - آواشناسی
  - واج‌شناسی
  - معناشناسی
  - کاربردشناسی
  - تحلیل گفتمان

  بررسی می شود.

زبان شناسی

- **نحو** یا **جمله‌شناسی** به دانش مطالعهٔ قواعد مربوط به نحوهٔ ترکیب و در کنار هم آمدن واژه‌ها به منظور ایجاد و درک جملات در یک زبان اطلاق می‌شود.
- **صرف** مطالعه در رابطه با ساختار کلمات است. در این بخش از ریشه واحد برای ساختن معانی متعدد استفاده می شود.
- **واج‌شناسی** به بررسی نظام آوایی زبان می‌پردازد و جایگاه عناصر آوایی را در زبان مشخص می‌کند. در این حوزه، مسائلی مانند آوا، واج، تکیه، آهنگ، وزن شعر مورد بررسی قرار می‌گیرند.

زبان شناسی

– در دانش زبان‌شناسی به دو آوا که جانشین کردن یکی با دیگری در معنی واژه تغییر بدهد واج می‌گویند.

- **آواشناسی** به مطالعه اصوات گفتار انسان می پردازد و با خواص فیزیکی اصوات گفتاری (آواها) ارتباط دارد
- **معناشناسی** (Semantics)، دانش بررسی و مطالعه‌ی معانی در زبان‌های انسانی است. بطورکلی، بررسی ارتباط میان واژه و معنا را معناشناسی می‌گویند.
- **زبان‌شناسی تاریخی** (Historical linguistics) به دانش مطالعه تغییرات زبان‌ها در گذر زمان اطلاق می‌شود.

پردازش زبان طبیعی

- پردازش زبان‌های طبیعی یکی از زیرشاخه‌های بااهمیت در حوزه هوش مصنوعی ، و مرتبط با دانش زبانشناسی است.

- تلاش عمده در این زمینه ماشینی کردن فرایند درک و برداشت مفاهیم یک زبان طبیعی انسانی‌ست.

پردازش زبان طبیعی

- به تعریف دقیق‌تر پردازش زبان‌های طبیعی عبارت است از استفاده از رایانه برای پردازش زبان گفتاری و نوشتاری.

- منظور از پردازش زبان طبیعی این است که کامپیوتری داشته باشیم تا قادر باشد زبان انسان را تحلیل کند، بفهمد و حتی بتواند زبان طبیعی تولید کند.

پردازش زبان طبیعی

- هدف اصلی در پردازش زبان طبیعی، ایجاد تئوری هایی محاسباتی از زبان، با استفاده از الگوریتم‌ها و ساختارهای داده ای موجود در علوم کامپیوتر است.

- در راستای تحقق این هدف، نیاز به دانشی وسیع از زبان است و علاوه بر محققان علوم کامپیوتر، نیاز به دانش زبان شناسان نیز در این حوزه می باشد.

پردازش زبان طبیعی

پردازش زبان طبیعی به پنج رده تقسیم‌بندی می‌شود:

- آوا شناسی و صداشناسی (phonetics and phonology)که به تشخیص آواها و صداها و بازشناسی گفتار می‌پردازد.
- ریخت‌شناسی (morphology) که به ساختارهای کلمات و ریشه‌یابی واژگان می‌پردازد.
- نحو (syntax)که به ارتباط کلمات به همدیگر و مباحث دستوری آنها در گروه‌ها و جملات می‌پردازد.

پردازش زبان طبیعی

- معناشناسی (semantics)که به ارتباطات معنایی کلمات می‌پردازد.
- عمل‌گرایی (pragmatics)که کاربردهای زبان برای رساندن یک مطلب به مخاطب یا مخاطبان، در حالت عملی یا در نوشتار و گفتار مطرح می کند.

پردازش زبان طبیعی

دسته بندی کاربردهای پردازش زبان طبیعی:

- کاربردهای بر پایه متن.
- فهم زبان طبیعی.
- سیستم های مکالمه.
- چند بعدی

پردازش زبان طبیعی

کاربردهای متنی :

- خلاصه سازی خودکار
- استخراج مهمترین جمله یک متن
- استخراج کلیدواژه
- تولید چکیده
- ویرایش ادبی

پردازش زبان طبیعی

- غلط یابی املایی
- ترجمه ماشینی
- بازیابی اطلاعات
- سیستم پرسش و پاسخ هوشمند
- موتورهای جستجو

پردازش زبان طبیعی

کاربردهای گفتاری:

- سیستم های پرسش و پاسخ انسان با رایانه
- سرویس های اتوماتیک ارتباط با مشتری از طریق تلفن
- سیستم های آموزش به دانش آموزان
- سیستم های کنترلی توسط صدا.

پردازش زبان طبیعی

زبان طبیعی فهم :

- یک سطح عمیق از آنالیز هستند:
- مثال: پیدا کردن متونی که در رابطه با فلسفه اسلامی و تمدن غرب است.
- سیستم باید اطلاعات کافی را برای مشخص کردن اینکه آیا مقاله ها ملاک تعریف شده ای به وسیله پرسش معرفی می کنند، استخراج کند.

پردازش زبان طبیعی

کاربردهای مبتنی بر مکالمه : همان ارتباط بین ماشین وانسان را شامل می شود

- سیستم پردازش پایگاه داده
- سرویس های مشتری خودکار مثل سرویس های بانکی
- سیستم حل مسئله

پردازش زبان طبیعی

چند بعدی :

شامل دو یا بیش از دو بعد ارتباطی است:

- متن
- گفتار
- اشاره
- تصویر

پیکره ها

(croup)

پیکره ها

- برای هر نوع پژوهش در حیطه‌ی زبان‌شناسی و پردازش زبان طبیعی با هدف استخراج قواعد زبان طبیعی و نیل به اهداف آموزشی، حجم زیادی از داده‌های زبانی را بر اساس معیارهای مشخص و در بیشتر موارد به صورت الکترونیکی جمع‌آوری و ذخیره می‌کنند.

پیکره ها

- از این پیکره‌ها به منظور به دلایل زیر استفاده می شود:
  - تحلیل‌های آماری
  - توصیف زبان (توصیف ویژگی‌های صرفی، لغوی، املایی و آوایی)
  - بررسی فرضیه‌ها یا رخدادها
  - مقایسه‌ی الگوریتم‌ها
  - بررسی صحت قواعد زبانی

پیکره ها

- در بیشتر موارد پیکره‌ی متنی، مجموعه‌ای از متون نوشتاری است که از منابع واقعی مثل پایان‌نامه‌ها، روزنامه‌ها، مجلات، کتاب‌ها و صفحات وب در موضوعات مختلف جمع‌آوری، تصحیح و بر حسب نیاز برچسب‌گذاری و طبقه‌بندی موضوعی شده است.

پیکره ها

- از این پیکره‌ها به منظور به دلایل زیر استفاده می شود:
  - تحلیل‌های آماری
  - توصیف زبان (توصیف ویژگی‌های صرفی، لغوی، املایی و آوایی)
  - بررسی فرضیه‌ها یا رخدادها
  - مقایسه‌ی الگوریتم‌ها
  - بررسی صحت قواعد زبانی

پیکره ها

- تاکنون پیکره‌هایی برای زبان انگلیسی، فرانسوی، آلمانی، ایتالیایی، اسپانیایی، سوئدی، نروژی، هلندی، عربی، عبری، ارمنی، لاتینی، یونانی، ژاپنی تولید شده است.

- در زبان فارسی نیز می‌توان به پیکره‌هایی مانند پیکره‌ی تطبیقی فارسی- انگلیسی دانشگاه تهران، پیکره‌ی موازی انگلیسی - فارسی میزان، پیکره‌ی متنی زبان فارسی، فارس‌نت، مجموعه‌ی همشهری، فرهنگ ظرفیت نحوی افعال فارسی و فرهنگ جامع واژگان مترادف و متضاد زبان فارسی اشاره کرد.

پیکره ها

- ویژگی‌ها و شرایط استفاده از پیکره‌های زبان فارسی در وب‌سایت دادگان که به عنوان مرجع دادگان زبان فارسی محسوب می‌شود، قابل مشاهده است.
- در این وب‌سایت ۲۷ پیکره مربوط به زبان فارسی معرفی شده‌اند.

www.dadehgan.ir

پیکره ها

- برخی از پیکره های مطرح در زبان انگلیسی که اساس ساخت پیکره ها در سایر زبانها هستند به قرار زیر است:

  - Word Net
  - Frame Net
  - Verb Net
  - Prop bank

پیکره ها – Word Net

- این پیکره به عنوان مطرح‌ترین پیکره‌ی زبان‌شناسی انگلیسی حاصل تلاش در دانشگاه پنسیلوانیا است.
- این پیکره برای بیشتر لغات انگلیسی، ارتباط‌های معنایی بین واحدهای لغوی را تحت پوشش قرار می‌دهد.
- این سیستم بطور همزمان مزایای یک فرهنگ جامع را دارا است.

پیکره ها – Word Net

- اطلاعات موجود در این پیکره بر اساس یک تقسیم‌بندی منطقی به نام synset مرتب شده‌اند.

- هر synset شامل لیستی از لغات مترادف و یک اشاره‌گر معنایی است که این اشاره‌گر یک ارتباط معنایی بین synset جاری و synset دیگر برقرار می‌کند.

پیکره ها – Word Net

- برای مثال لغت car، auto، automobile و motocar تشکیل‌دهنده‌ی یک synset با یک تعریف مشترک هستند.

پیکره ها – Word Net

- برخی از اصطلاحاتی که در WordNet مطرح می‌شوند، به قرار زیر است:

- Sense (معنی): معنای یک لغت در WordNet است.

- Synset: مجموعه‌ای از لغات که در برخی متون به جای هم قابل استفاده هستند. (مترادف)

پیکره ها

- Hyponym: لفظی خاص که در مشخص کردن عضوی از یک مجموعه به کار می‌رود. X یک Hyponym از Y است یعنی X نوعی Y است (مگس Hyponym حشرات است).

- Hypernym: این اصطلاح در WordNet برعکس Hyponym عمل می‌کند (موجودات زنده Hypernym انسان است).

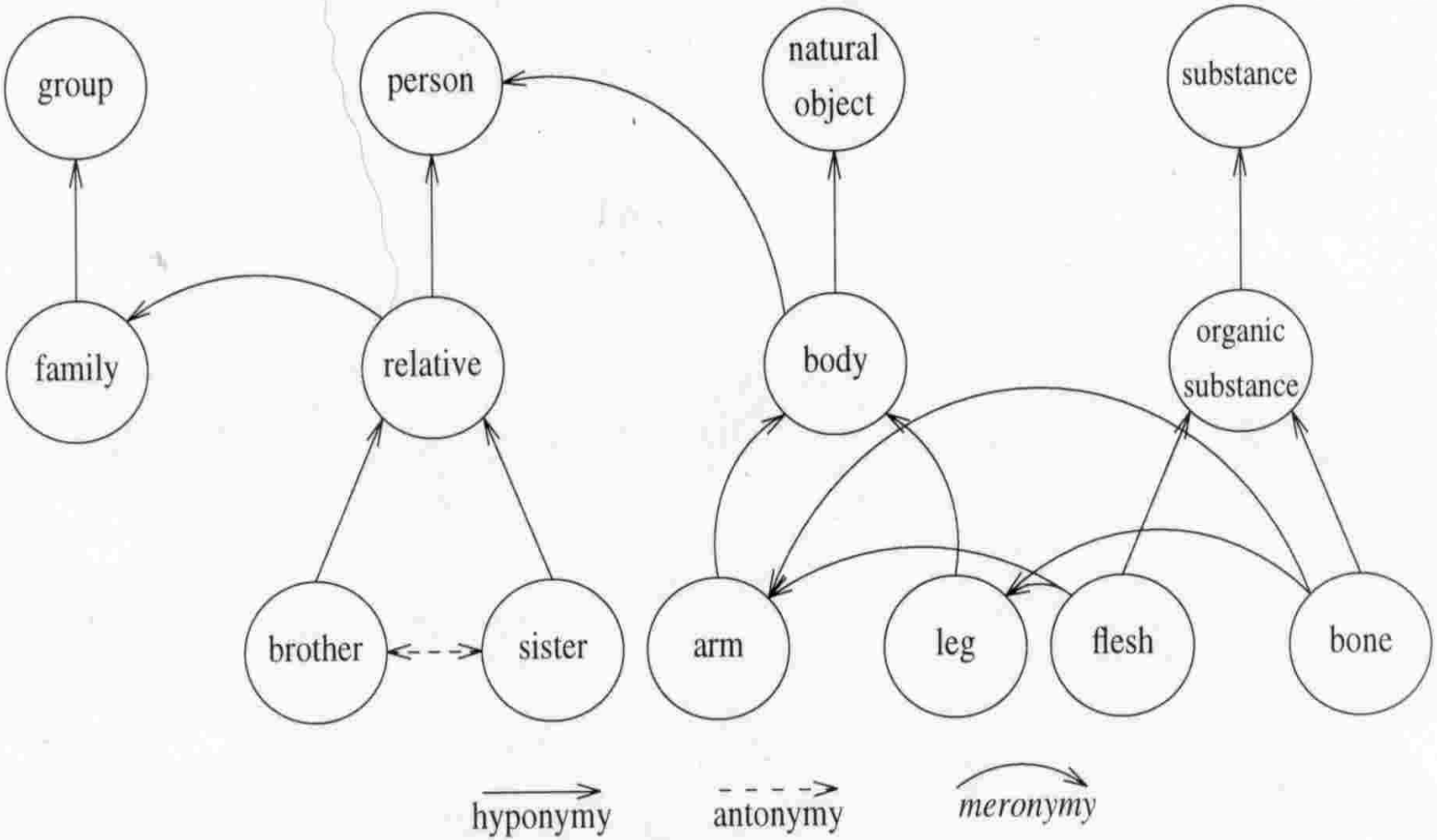
پیکره ها

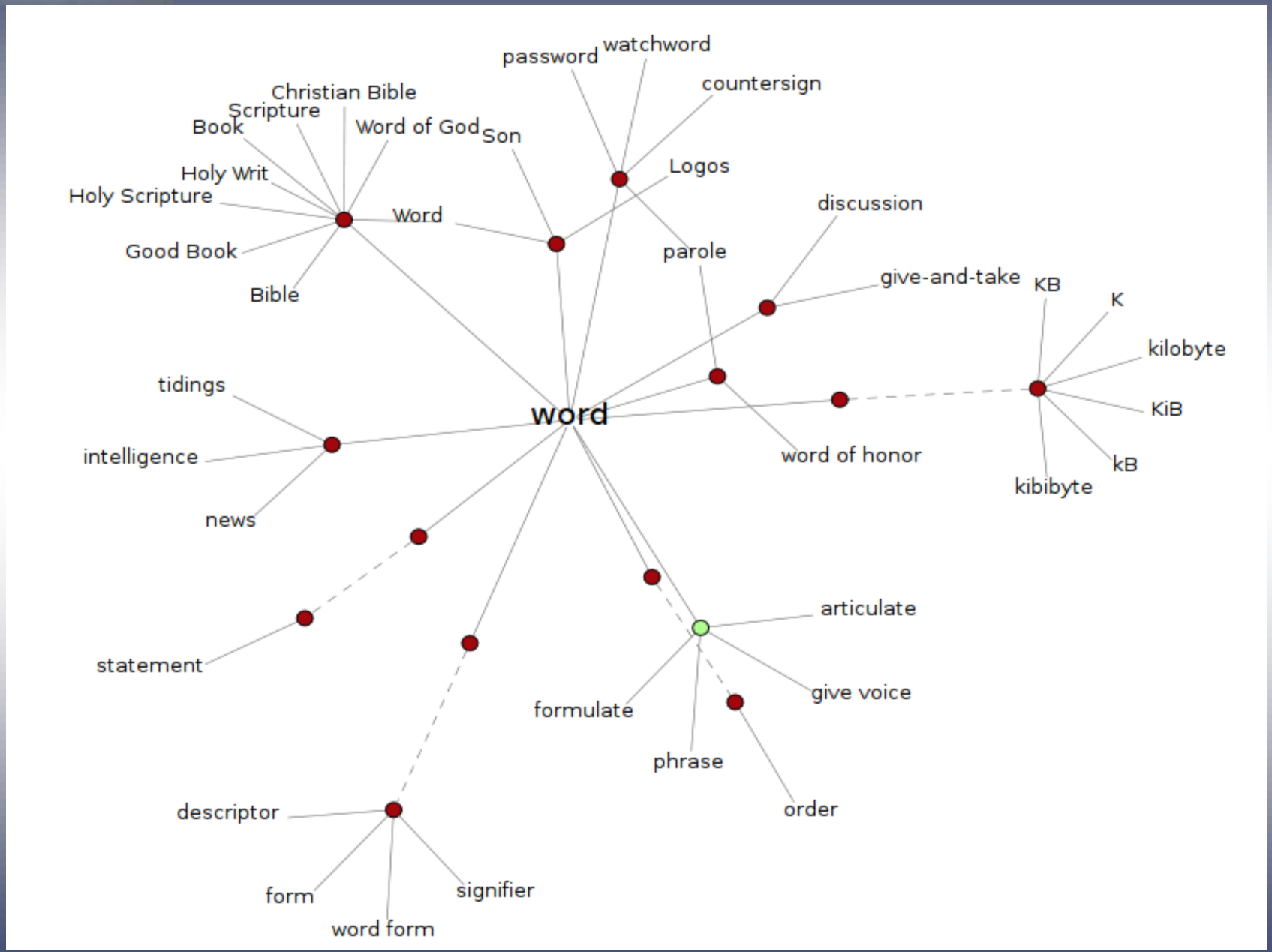
- Meronym: هرگاه گفته شود X یک Meronym از Y است به این معنا که X بخشی از Y است (پنجره Meronym ساختمان است).
- Holonym: این اصطلاح در WordNet برعکس Meronym عمل می‌کند .

پیکره ها

- Meronym: هرگاه گفته شود X یک Meronym از Y است به این معنا که X بخشی از Y است (پنجره Meronym ساختمان است).
- Holonym: این اصطلاح در WordNet برعکس Meronym عمل می‌کند .

Figure 2. Network representation of three semantic relations among an illustrative variety of lexical concepts




password
watchword
countersign
Christian Bible
Scripture
Book
Word of God
Son
Logos
Holy Writ
Holy Scripture
Word
discussion
parole
Good Book
Bible
give-and-take
KB
K
kilobyte
tidings
word
KiB
intelligence
word of honor
kB
kibibyte
news
articulate
statement
give voice
formulate
phrase
order
descriptor
form
signifier
word form

پیکره ها – Word Net

- کاربردهای Word Net

  - تشخیص نقش لغات در متون
  - دسته بندی متون بصورت خودکار
  - خلاصه سازی متون به صورت خودکار
  - استفاده در پردازش های وب معنایی

پیکره ها – Frame Net

- در دانشگاه برکلی به عنوان یک منبع آنلاین لغات انگلیسی (شامل اسم، صفت و فعل) ایجاد شده است.
- از داده‌ها و اطلاعات این پروژه، در بحث زبان‌شناسی و پردازش زبان طبیعی به‌منظور برچسب‌زنی نقش‌های معنایی استفاده می‌شود.

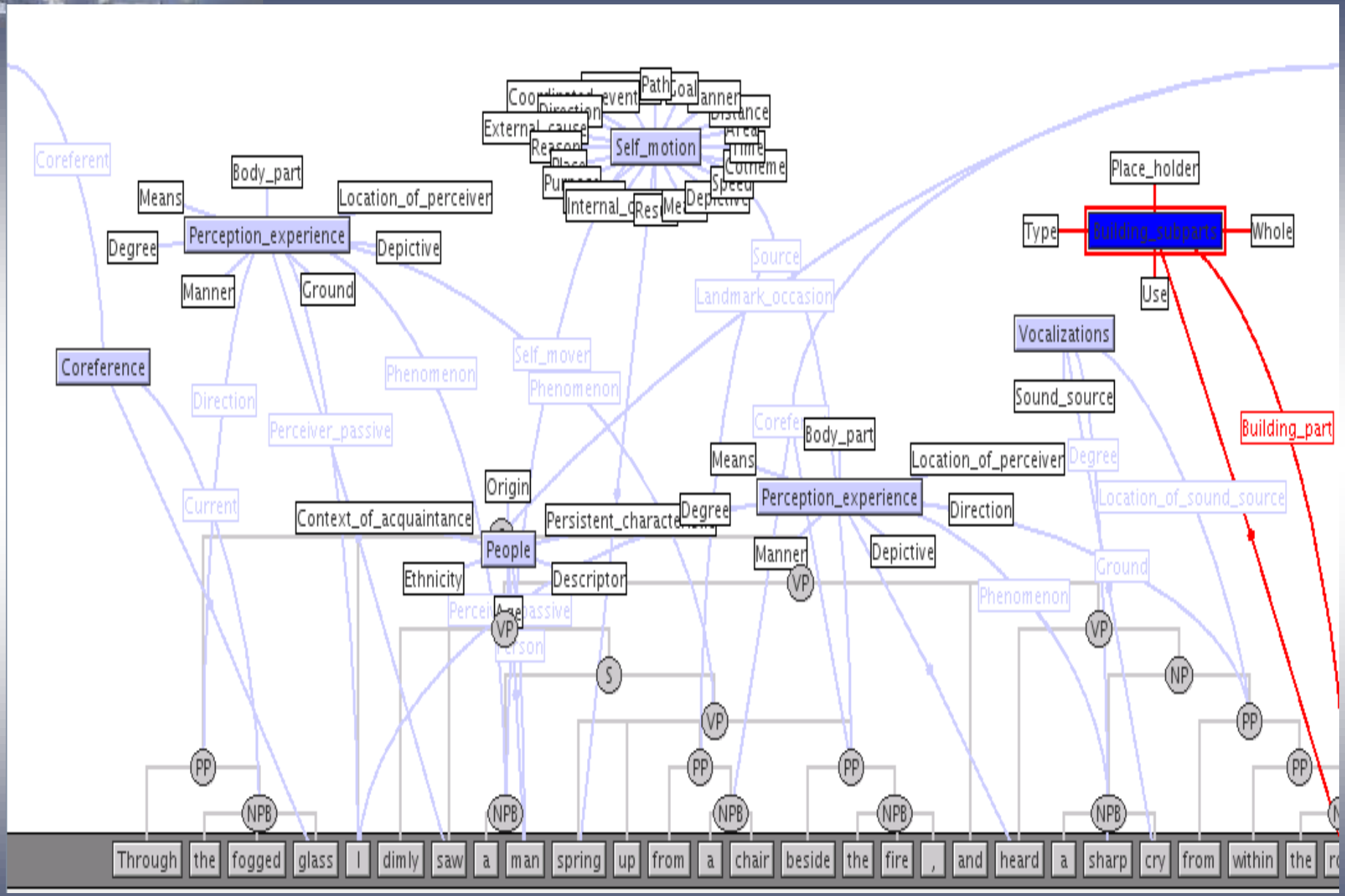
پیکره ها – Frame Net

- پروژه‌های تحقیقاتی فعال با محوریت فریم‌نت، به دنبال تولید قاب‌های معنایی واژگان برای زبان‌های دیگر و ساخت ابزارهای برچسب‌زنی خودکار برای قاب‌های معنایی هستند

- در حال حاضر پایگاه‌داده واژگان FrameNet، شامل ۱۲۰۰ قاب معنایی، ۱۳۰۰۰ واحد واژگانی و بیش از ۱۹۰۰۰۰ جمله تشریح شده می‌باشد.

پیکره ها – Frame Net

- در FrameNet قاب معنایی را می‌توان به عنوان یک ساختار مفهومی توصیف‌کننده‌ی یک رویداد، ارتباط و یا شی با شرکت‌کنندگان در آن تصور کرد.
- اجزای یک قاب معنایی عبارتند از:
  - واحد واژگانی: یک کلمه همراه با معنای آن می‌باشد. هر دریافتی که از یک کلمه با معانی متعدد وجود دارد، به یک قاب معنایی تعلق می‌گیرد.
  - جمله نمونه: قاب‌ها با جملات نمونه مرتبط هستند و عناصر قاب در جملات مشخص شده‌اند.


Through the fogged glass I dimly saw a man spring up from a chair beside the fire , and heard a sharp cry from within the
Self_motion
Perception_experience
Body_part
Means
Location_of_perceiver
Degree
Depictive
Manner
Ground
Coreference
Place_holder
Type
Building_subparts
Whole
Use
Vocalizations
Sound_source
Building_part
Location_of_sound_source
Context_of_acquaintance
People
Origin
Ethnicity
Descriptor
Persistent_characteristic
Direction
Phenomenon
Source
Landmark_occasion
Self_mover
Perceiver_passive
Current
S
VP
NP
PP
NPB

پیکره ها – Verb Net

- این منبع به شکل سلسله‌مراتبی و همراه با اطلاعات نحوی و معنایی برای افعال انگلیسی، ساخته شده است.
- هر گره در سلسله‌مراتب، توسط مجموعه ای از افعال و لیستی از ساختار آن افعال و اطلاعات معنایی و نحوی مرتبط با آنها تشریح می‌شود.

پیکره ها – Verb Net

- همه‌ی اعضای یک کلاس، معانی مشابهی داشته و یک مجموعه‌ی مشترک از نقش‌های موضوعی و قاب‌های نحوی را به اشتراک می‌گذارند

ابهام ها در پردازش زبان

ابهام معنایی (Semantic ambiguity): دارای پیچیدگی زیادی است و وقتی ناشی می شود که بین کلمات جمله، سطوح ارتباطی متفاوت برقرار گردد و باعث شود از جمله معناهای متفاوتی ادراک شود مانند :

- حسین دوست ۵ ساله من است.
- علی با پدر و مادر معلمش به سفر رفتند.

ابهام ها در پردازش زبان

ابهام کاربردی (Pragmatic ambiguity): وقتی ایجاد می شود که جمله در حالت های مختلف بیان شود و شرایط زمان و مکان معناهای متفاوتی ایجاد کند. مانند:

- تلفن زنگ می زند.
- وعده سرخرمن می دهد.

ابهام ها در پردازش زبان

ابهام کمیت پذیر (Quantifiable ambiguity) :
روی کلماتی رخ می دهد که خود کلمه از نظر مقدار اندازه پذیر است و اندازه آن ابهام دارد.

- مانند کلمات : تعدادی – خیلی - کم و بیش – داغ - باهوش

ابهام ها در پردازش زبان

ابهام کمیت ناپذیر (Nonquantifiable ambiguity)
: روی کلماتی ایجاد می شود که نمی توانند ارزش عددی بگیرند ولی ارزش های غیر عددی می توانند داشته باشند.

- مانند کلمات : زیبا – خوشرنگ - توانمند

تجزیه و تحلیل آوایی

(phonetics & phonology)

تجزیه و تحلیل آوایی

- هر زبانی کلمات خود را از تعداد محدودی از آواها (Phonemes) می سازد.
- صدا شناسی (Phonology) به ترتیب آواها و درک قوانین آنها در گفتار مرتبط می باشد
- آوا شناسی (Phonotic) به ویژگیهای صوتی گفتار و شیوه بیان آنها مرتبط است

تجزیه و تحلیل صرفی
(Morphology)

تجزیه و تحلیل صرفی

- در این سطح، تشخیص تک تک کلمات ورودی زبان طبیعی انجام می گیرد.

- برای انجام این کار نیاز به یک فرهنگ لغات است. اما می توان برخی از کلمات را از برخی دیگر بدست آورد.

تجزیه و تحلیل صرفی

- ریشه کلمات در فرهنگ لغت جای گرفته و بقیه کلمات با استفاده از قواعد لازم، ساخته شوند. این قواعد را قواعد صرفی (Morpholigical rules)می نامند.
- با قواعد صرفی می توان بوسیله یک کلمه، مجموعه وسیعی از کلمات را تولید یا آنکه ریشه یک کلمه را بدست آورد.

تجزیه و تحلیل صرفی

- در سطح تجزیه و تحلیل صرفی، **کلمه** مورد بررسی قرار می گیرد تا در صورت بکار رفتن قواعد صرفی، **ریشه** آن استخراج گردد.

Morphology & Finite state transdure

- عبارات با قاعده می تواند برای یافتن ریشه کلمات مورد استفاده قرار گیرد . اما این روش همیشه در تعداد کمی از کلمات ما را با مشکل روبرو می کند.
- مثال: رهایی ، اژدها – مردم (مُردم یا مَردم)

- سه رهیافت رایج برای ریشه یابی عبارتند از:
  - **رهیافت ساختاری**
  - **رهیافت جدول مراجعه**
  - **رهیافت آماری**

- الگوریتمهای مربوط به **رهیافت ساختاری** (مبتنی بر قاعده)
  - وابسته به تحلیل ساخت واژی زبان میباشند.
  - در این الگوریتمها، با توجه به یک سری قواعد از پیش تعریف شده، به حذف برخی وندها جهت استخراج ریشه پرداخته میشود.
  - الگوریتم پورتر مثالی از این دسته از الگوریتمهاست.
  - این ریشه یاب از ۵ مرحله تشکیل شده است. در طی این مراحل قواعدی بر روی کلمه اعمال میشود و بزرگترین پسوند آن حذف میشود.

- در **رهیافت جدول مراجعه**
  - هر لغت و ریشه مربوط به آن در یک داده ساختار ذخیره شده اند.
  - ریشه هر لغت ذخیره شده را میتوان یافت.
  - این رهیافت نیاز به فضای حافظه زیادی دارد.
  - همچنین برای هر لغت جدید، جدول بایستی به طور دستی به روز رسانی شود.

- در **رهیافت آماری**،
  - از میان یک فرایند استنتاج و بر مبنای آماره های یک پیکره متنی، قواعدی با توجه به ساختمان لغت، استخراج میشود.
    - تعداد رخداد، N-gramها،
  - رهیافت آماری به هیچ وجه نیازی به دانش زبانشناسی ندارد و در کل، از ساختار ساخت واژه ی زبان، مستقل است.

- در مبحث ریشه یابی دو اصطلاح مرسوم وجود دارد:
  - Stemming : طراحی یک فرایند اکتشافی برای جدا کردن پیشوندها و پسوندهای یک کلمه به منظور رسیدن به ریشه
  - Lemmatization : استفاده از لغت نامه و تحلیل مورفولوژی کلمات به منظور حذف مقداری از انتهای کلمه تا رسیدن به ریشه

- الگوریتم پورتر
  - در ۵ فاز کاهشی سعی در حذف پسوندها دارد
  - در هر فاز قواعدی برای انتخاب یک قانون برای کاهش کلمه اعمال می شود .
  - در مراحل جداسازی، بعد از حذف پسوند کلمه باقی مانده مورد سنجش قرار می گیرد که آیا طول منطقی را دارد یا خیر ؟

Rule	Example
SSES ⇒ SS	Caresses ⇒ caress
IES ⇒ I	Ponies ⇒ poni
SS ⇒ SS	Caress ⇒ caress
S ⇒	Cats ⇒ cat

برخی قوانین ساخت صرفی کلمات فارسی:

[اسم] + [(تکواژ جمع)] + [(تکواژ نکره‌ساز) (تکواژ بند موصولی) (واژه‌بست‌های شخصی) (کسرهٔ اضافه)] + [(واژه بست‌های ربطی)]

صفت + [(تکواژ صفت تفصیلی‌ساز) (تکواژ صفت عالی‌ساز)] + [(تکواژ جمع)] + [(تکواژ نکره‌ساز) (تکواژ بند موصولی) (واژه‌بست‌های شخصی) (کسرهٔ اضافه)] + [(واژه‌بست‌های ربطی)]

- مثال: فرهنگ، فرهنگ‌ها، فرهنگ‌های، فرهنگ‌هایشان، فرهنگ‌مان
- برای ساختار بالا ذکر چند نکته لازم است:
- وجود هر یک از موقعیت‌های ۱، ۲ و ۳ در ساختار تصریف اسم اختیاری بوده و عدم حضور هر یک از آنها در این ساختار امکان‌پذیر است.
- با حضور یکی از وندهای موقعیت ۲ امکان حضور سایر وندهای این موقعیت در اسم وجود ندارد.

تجزیه و تحلیل نحوی
(Syntax)

تجزیه و تحلیل نحوی (Syntatic)

- نحو در زبان، عبارت است از مجموعه قواعدی که جملات صحیح دستوری را مشخص می نماید.
- در این سطح پردازنده زبان طبیعی بر روی اطلاعات ساختاری و ارتباط کلمات متمرکز می گردد.
- جمله ورودی از لحاظ مراعات قواعد دستوری مورد بررسی قرار می گیرد.

تجزیه و تحلیل نحوی (Syntatic)

- فرآیند تعیین و مشخص کردن کلمات در متن بر پایه معنا و محتوای متنی که در آن قرار دارد
- رابطه آن کلمه با کلمات مجاورش در عبارت
- نحو، مطالعه رابطه رسمی بین کلمات است.
- یک کلمه می تواند بر اساس موقعیتش در جمله بیشتر از یک برچسب نحوی داشته باشد

نحو – Syntax

- در سیستمهای پردازش زبان طبیعی به علت تنوع قواعد نحوی، معمولا زیرمجموعه ای که بتواند حداکثر جملات ممکن در یک محدوده یا کاربرد را پوشش دهد، انتخاب می گردد.
- برخی از برچسب های نحوی زبان : فعل، اسم، صفت، قید، ضمیر، حروف اضافه، عطف، ندا

نحو – Syntax

- اولین پیکره برچسب خورده زبان انگلیسی Brown و اولین پیکیره برچسب خورده زبان فارسی پیکره بی جن خان است .
- مدل مخفی مارکوف برای برچسب زنی کلمات از سال ۱۹۸۰ مورد استفاده قرار گرفت

نحو – Syntax

- این مدل با تشکیل جدول احتمالاتی برای یک دنباله احتمالی از برچسب های مختلف عمل برچسب زنی را انجام می دهد . مثلا اگر در متنی با حرف تعریف the برخورد کردیم به احتمال ۴۰ درصد کلمه بعدی اسم و به احتمال ۴۰ درصد کلمه بعدی صفت و به احتمال ۲۰ درصد کلمه بعدی یک عدد است.

نحو – Syntax

- از مزایای دیگر مدل مخفی مارکوف تعیین دنباله احتمالاتی بیش از دو کلمه است. برای مثال با مشاهده یک حرف تعریف و بعد ازآن یک فعل کلمه بعد به احتمال زیاد حرف اضافه ، حرف تعریف یا اسم است و به احتمال ضعیف یک فعل دیگر می باشد.
- سیستم CLAWS سیستمی است که از این روش برای تعیین برچسب کلمات استفاده کرده که دارای دقت بین ۹۳ تا ۹۵ درصد است

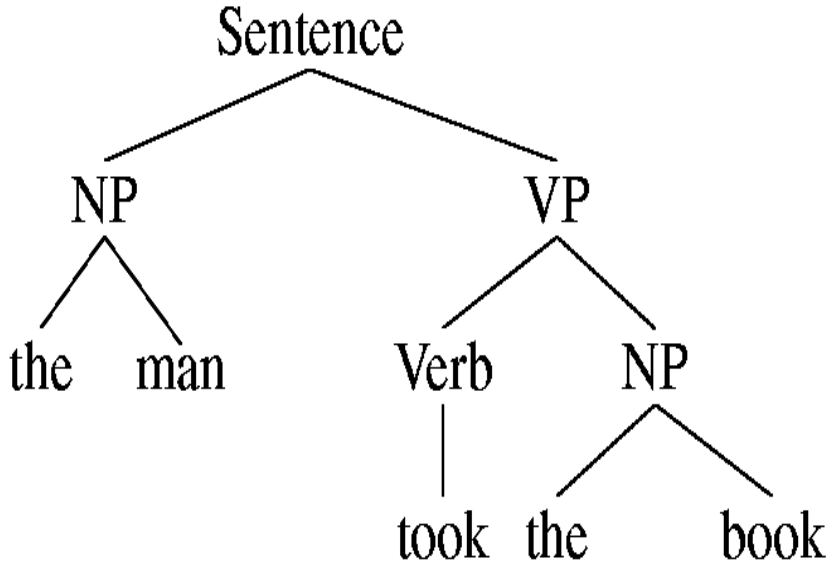
نحو- در این سطح، موارد ذیل مطالعه می گردد:

۱- چگونگی دسته بندی کلمات به کلاسهایی که بخشهای گفتار (Part-Of-Speech) نامیده می شود.

۲- چگونگی گروه بندی آنها به همسایگانشان درون عبارت

۳- چگونگی وابستگی کلمات به سایر کلمات در جمله

Context-free-Grammar for English



The first context-free grammar parse tree (Chomsky, 1956)

برخی کاربردهای NLP

خلاصه سازی

خلاصه سازی

منظور از «خلاصه‌سازی»، دریافت یک متن و تولید یا استخراج یک متن دیگر از آن متن است؛ به گونه‌ای که:

- متن به دست‌آمده از متن اصلی کوتاه‌تر باشد
- نکات اصلی و مهم آن را دربرداشته باشد
- خوانا باشد
- بین جملات آن پیوستگی وجود داشته باشد.

خلاصه سازی

- اگر متن خلاصه با انتخاب جملاتی از متن اصلی به دست آید، نوع خلاصه‌سازی، «استخراجی» یا «گزینشی» است
- اگر خلاصه متن پس از فهم مطالب موجود در متن اصلی تولید شود، خلاصه‌سازی از نوع «چکیده» است.

خلاصه سازی

- در روش چکیده نیز باید ابتدا متن اصلی فهمیده شود و بر اساس معنای موجود در متن و به صورت معنایی، چکیده‌ای از متن اصلی تولید شود.
- در این روش، با چالش‌های موجود در زمینه پردازش زبان طبیعی و تجزیه و تحلیل معنایی متن، برای درک و تفسیر متن روبه‌رو هستیم.

خلاصه سازی

- فازهای فرآیند خلاصه‌سازی متون:
  - پیش‌پردازش
  - تحلیل
  - انتخاب

خلاصه سازی

در فاز پیش‌پردازش، پردازش‌های اولیه مورد نیاز بر روی متن صورت می‌گیرد. کارهای صورت گرفته در این مرحله :

۱ - یکسان‌سازی نگارشی:

- یکسان سازی فاصله بین پیشوندها و پسوندها با کلمه اصلی
- یکسان سازی حروف مانند حرف ی و ک در فارسی و عربی

یکسان سازی اعداد فارسی ، عربی و انگلیسی

- یکسان سازی فاصله بین کلمات مرکب

خلاصه سازی

۲- یکسان سازی کلمات مترادف

- کلماتی مانند: «کامپیوتر» و «رایانه» مترادف هستند. در مرحله پیش‌پردازش لازم است کلمات مترادف یافته و به یک شکل تبدیل شوند.

- برای این مرحله دو پیکره قابل استفاده است:

- فرهنگ طیفی (تزاروس زبان فارسی)
- فرهنگ مترادف های زبان فارسی

خلاصه سازی

۳ - حذف لغات و بخش‌های اضافی

- حروف اضافه
- کلمات توقف
- بخش هایی از متن که حکم توضیحات اضافی را دارند :
  - داخل پرانتز
  - خط تیره
  - داخل پاورقی
  - داخل براکت

خلاصه سازی

۴ - تجزیه جمله و استخراج کلمات کلیدی

یک جمله، از تعدادی کلمه تشکیل شده است.

شناسایی کلمات موجود در یک جمله و میزان اهمیت آنها، می‌تواند تا حد بسیاری میزان اهمیت کل جمله را تعیین نماید.

خلاصه سازی

۵ – ریشه یابی کلمات

کلمات استخراج شده از جملات باید ریشه‌یابی شده و با ریشه خود جایگزین شوند.

۶ – یافتن ارتباط بین جملات

در مرحله پیش‌پردازش می‌توان ارتباط بین جملات یا معانی آنها را یافت و تشابه یا تفاوت بین برخی از بخش‌های متن را شناسایی کرد.

خلاصه سازی

اولین تلاش های برای ساخت سیستمی که قادر به خلاصه سازی خودکار یک متن باشد در ۱۹۵۸ توسط Luhn صورت گرفت.

این سیستم قادر بود جملات مهم متن را بر اساس اینکه کدام جمله دارای بیشترین تعداد کلمه مهم است استخراج کند.

خلاصه سازی

در سال ۱۹۶۹ روشی بر پایه روش سیستم Luha معرفی شد که در آن خصوصیات نوع متن را نیز در نظر می گرفت و از خصوصیاتی نظیر موقعیت جمله، کلمات موجود در عنوان و عبارات خاص برای بهبود کارایی سیستم خلاصه سازی استفاده می نمود

از سال ۱۹۷۰ روش های خلاصه سازی بر پایه مفهوم در هوش مصنوعی مرسوم و مورد توجه قرار گرفت.

انواع مدل های خلاصه سازی

- روش بالا به پایین
  - خلاصه سازی بر پایه ی درخواست کاربر می باشد .
  - معیارهایی به عنوان علایق کاربر دریافت می شود و سپس سیستم با استفاده از این خصوصیات قسمت های مختلف متن را آنالیز می کند.
- روش پایین به بالا
  - بر پایه ی متن موجود می باشد و اولویت های متن مدنظر قرار خواهد گرفت.

انواع مدل های خلاصه سازی

- در این روش معیارهای عمومی به صورت یکسری استراتژی کد می شود سپس سیستم این استراتژی ها را بر روی قسمت های مختلف متن اعمال می کند

روشهای اولویت دهی و انتخاب یک جمله

- روش بر پایه موقعیت
  - ایده این روش بدین صورت است که جملات مهم متن در ابتدا یا انتهای متن قرار دارند.
  - تجربه نشان داده است که در ۸۵ درصد موارد، جملات مهم در مکان های ابتدایی و در ۷ درصد موارد در مکان های پایانی قرار دارند
  - نقاط ضعف : محل جملات مهم می تواند متناسب موضوع متون متفاوت باشد.

روشهای اولویت دهی و انتخاب یک جمله

- روش سیاست مکانی بهینه
  - با توجه به اینکه موقعیت مناسب جملات و پاراگراف را به صورت داینامیک و متناسب با نوع متن به دست می آورد از کارایی بهتری برخوردار است.

روشهای اولویت دهی و انتخاب یک جمله

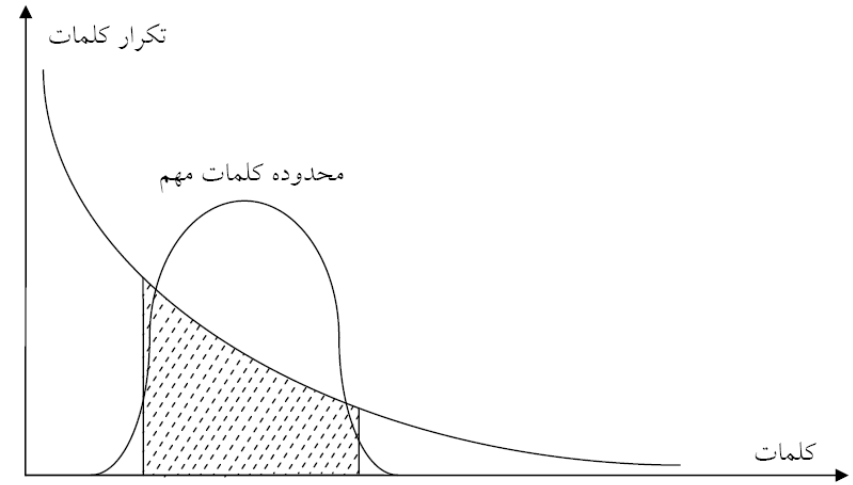
- روش بر پایه عنوان
  - ایده این روش به این صورت است که کلمات واقع شده در عنوان های کلی و عنوان های بخش های جزئی تر متن ارتباط مستقیمی با خلاصه تولید شده دارند.
  - تجربه نشان داده ۹۹ درصد این جملات به صورت آماری از اهمیت برخوردار هستند و در خلاصه کردن مفید واقع می شوند

روشهای اولویت دهی و انتخاب یک جمله

- روش بر پایه عبارات خاص
  - جملات مهم متن اغلب شامل عبارات خاصی هستند.
  - این عبارات می توانند عبارت مثبت یا منفی باشند.
  - همیشه مجموعه ای از عبارات خاص برای همه مستندات موجود نیست.

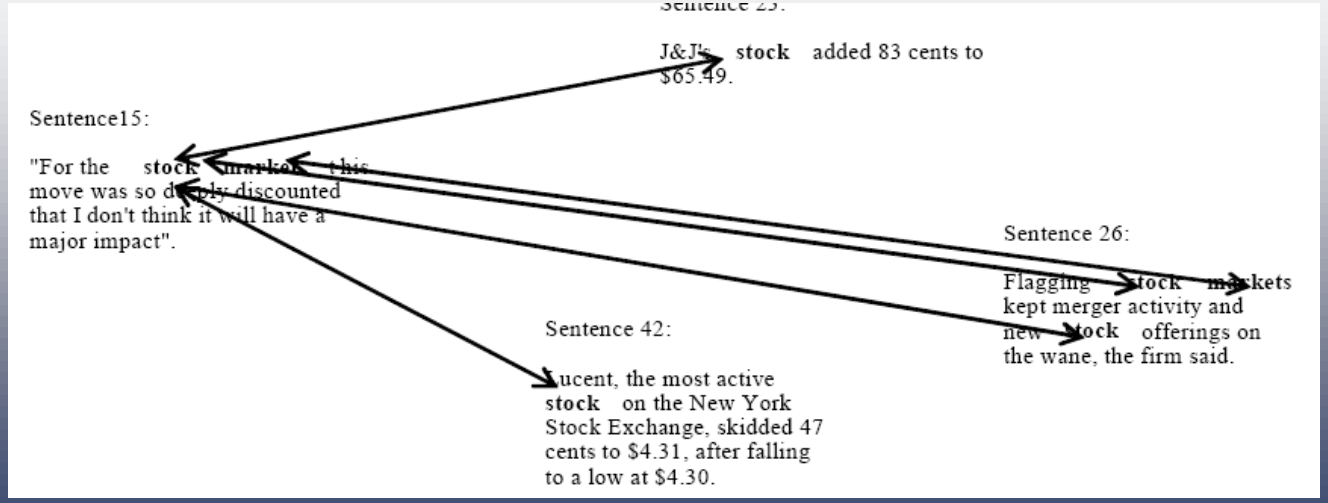
روشهای اولویت دهی و انتخاب یک جمله

- روش برپایه بسامد لغوی
  - ایده در این روش بر این است که جملات مهم حاوی کلماتی هستند که دارای تکرار بالایی هستند.
  - در این روش از روش وزن دهی TFIDF استفاده می شود.



روشهای اولویت دهی و انتخاب یک جمله

- روش بر پایه پیوستگی:
  - در این روش جملات و عبارات مهم دارای بیشترین اتصال در گراف شباهت خود هستند. این گراف خود می تواند بر اساس تکرار یک کلمه یا شباهت معنایی یا هم مرجعی و غیره باشد.



روشهای اولویت دهی و انتخاب یک جمله

- روش هم وقوعی کلمات
  - در این روش ابتدا باید یک واحد برای متن خود جهت تشابه انتخاب کنیم. (مثلا پاراگراف) در قدم بعدی هر پاراگراف را به صورت یک بردار درنظر می گیریم . عناصر این بردار میزان اهمیت کلمه در پراگراف است.
  - $D_i = (d_{i1} , \dots , d_{ij})$

روشهای اولویت دهی و انتخاب یک جمله

در قدم بعدی باید بتوانیم شباهت بین دو پاراگراف را مشخص کنیم.

$$Sim(D_i, D_j) = \Sigma\, d_{ik} \cdot d_{jk}$$

- بعد از این مرحله کلیه ی پاراگراف ها را به هم متصل نموده و وزن اتصالات را برابر مقدار شباهت قرار می دهیم . در مرحله بعد یال هایی که وزن کمتر از یک مقدار آستانه را دارند حذف می کنیم .

روشهای اولویت دهی و انتخاب یک جمله

- با حذف ارتباطات کمتر از مقدار آستانه گراف به چند بخش تقسیم می شود که اتصالات درون هر بخش بیشینه و اتصال بین بخش ها کمینه است. این زیر گراف ها در واقع پراگراف هایی هستند که از لحاظ معنایی بیشترین ارتباط را دارند و همگی تقریبا یک مفهوم را دربر می گیرند.

- در مرحله بعد باید از بین پاراگراف های یک بخش مهمترین آنها را انتخاب کرد. برای این انتخاب پاراگرافی انتخاب می شود که بیشترین اتصال را با بقیه داشته باشد.

روشهای اولویت دهی و انتخاب یک جمله

- روش زنجیره لغوی
  - در این روش جملات را بر اساس ارتباط کلمات بکار رفته در آنها ارزیابی می کنیم. برای یافتن ارتباط بین کلمات می توان از پیکره ای شبیه wordnet استفاده کرد.
  - ارتباط بین کلمات :
    - ارتباط فوق قوی (تکرار بالا در متن)
    - ارتباط قوی (wordnet)
    - ارتباط متوسط (ارتباط با بیش از یک واسطه)

روشهای اولویت دهی و انتخاب یک جمله

- روش **LEXRANK**

  - این روش نیز برپایه ساخت گراف شباهت جملات است. با این تفاوت که از یک ایده شبکه های اجتماعی بهره می گیرد.
  - مانند روش قبلی نودی در گراف که بیشترین درجه خروجی را دارد از بیشترین اهمیت برخوردار است . منتها موضوع اضافه شده در این روش بدین ترتیب است که علاوه بر این که یک نود چه تعداد ارتباط دارد و اینکه با چه نودهایی در ارتباط است نیز باید لحاظ گردد.

روشهای اولویت دهی و انتخاب یک جمله

- روش مبتنی بر دانش مباحثه ای
  - در این روش ساختار و ارتباط بین جملات مشخص شده و واحدهایی از متن که دارای محوریت اصلی و مرکزیت هستند بازگو کننده بخش های با اهمیت هستند
  - بنابراین روش دو جمله به چندین صورت می توانند به یکدگیر مرتبط باشند. بر این اساس یک جمله به عنوان پایه و یک جمله به عنوان وابسته.
  - جملات وابسته معمولا به وسیله عبارات خاصی به جملات اصلی مرتبط می شوند.
  - مثلا در زبان فارسی عباراتی مثلا بنابراین، اما ، و ، براین اساس و... متصل کننده دو جمله هستند.

روشهای اولویت دهی و انتخاب یک جمله

به عنوان مثال پاراگراف زیر را در نظربگیرد:

(A) Smart cards are becoming more attractive

(B) as the price of micro-computing power and storage continues to drop.

(C) They have two main advantages over magnetic strip cards.

(D) First, they can carry 10 or even 100 times as much information

(E) and hold it much more robustly.

(F) Second, they can execute complex tasks in conjunction with a terminal.

روشهای اولویت دهی و انتخاب یک جمله

به عنوان مثال پاراگراف زیر را در نظربگیرد:

(A) Smart cards are becoming more attractive

(B) as the price of micro-computing power and storage continues to drop.

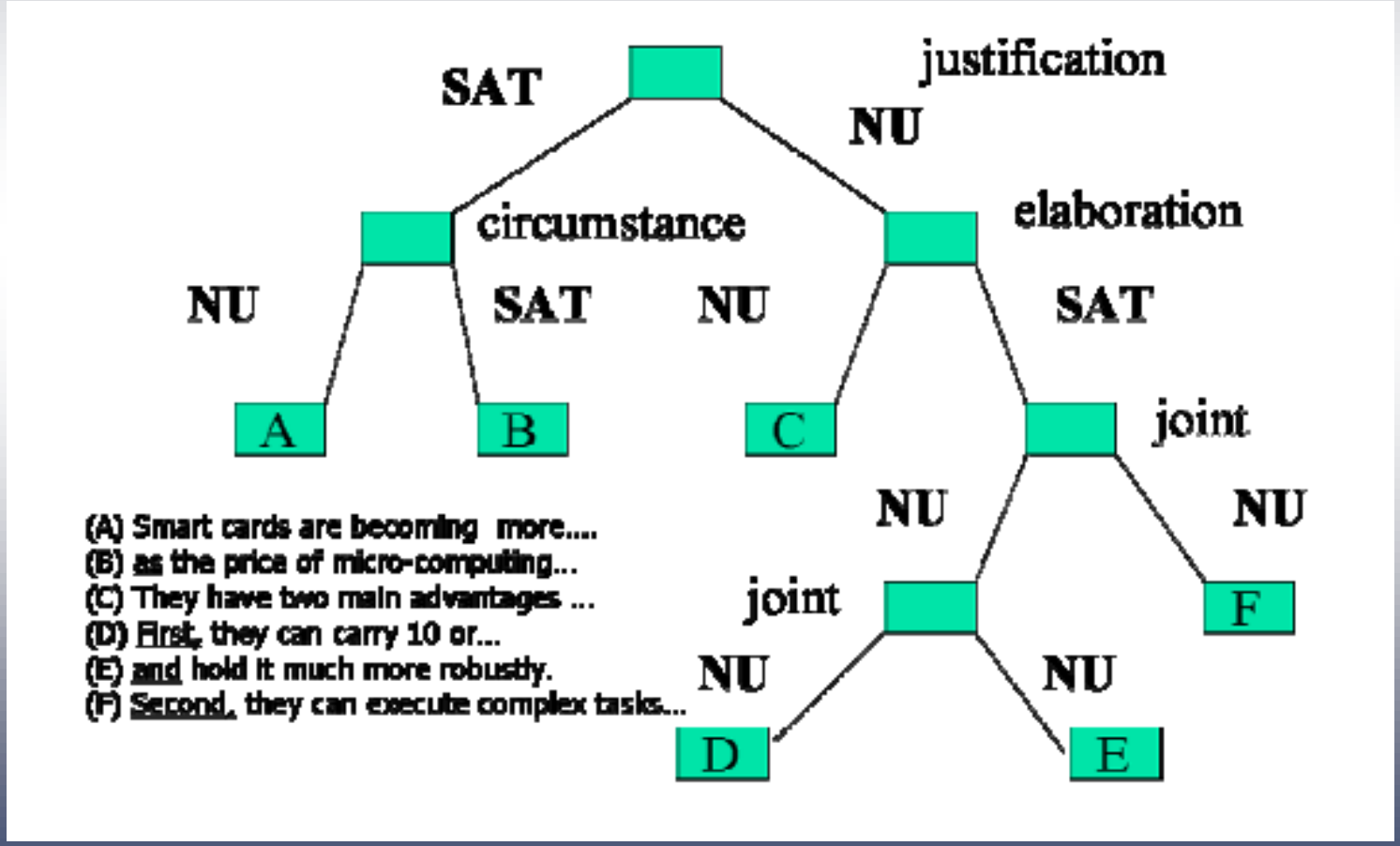
(C) They have two main advantages over magnetic strip cards.

(D) First, they can carry 10 or even 100 times as much information

(E) and hold it much more robustly.

(F) Second, they can execute complex tasks in conjunction with a terminal.

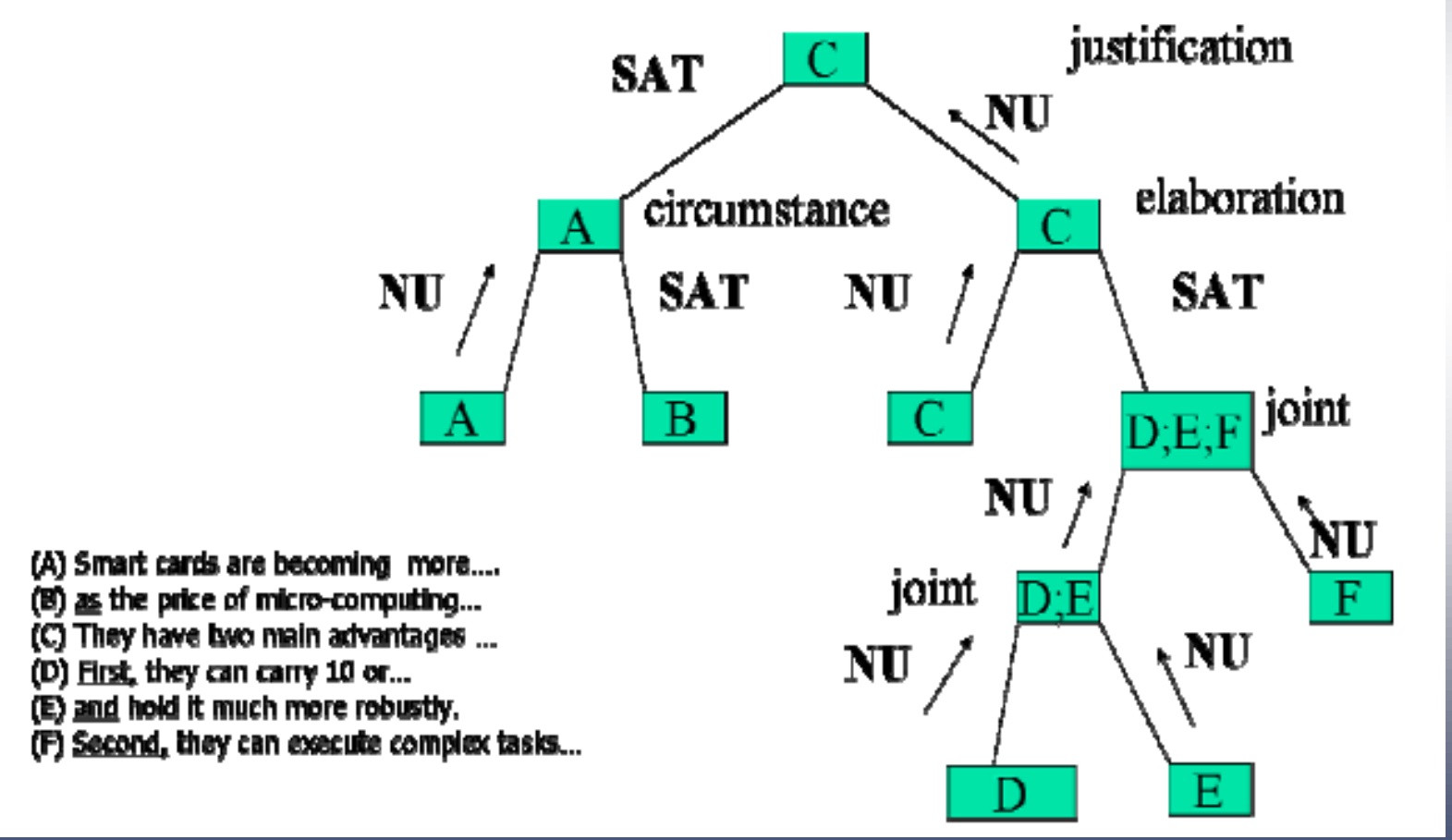
روشهای اولویت دهی و انتخاب یک جمله



روشهای اولویت دهی و انتخاب یک جمله

- عبارات پایه با برچسب NU مشخص شده اند.
- در این روش در ساختار درختی عبارات پایه که با NU مشخص شده را به نود بالایی انتقال می دهیم .

روشهای اولویت دهی و انتخاب یک جمله



برخی کاربردهای NLP

رده بندی

رده بندی

دسته بندی متون (کلاس بند ی متون) به عمل برچسب گذاری موضوعی متون زبان طبیعی بر مبنای یک مجموعه از پیش تعیین شده (مثلا ورزشی، سیاسی، خانواده، تغذیه و ...)، گفته می‌شود.

شناسایی رده، دسته یا طبقه یک متن می‌تواند اطلاعات مفیدی برای فرایندهایی همچون ترجمه ماشینی، تبدیل نوشتار به گفتار و نویسه‌خوان نوری (OCR) ارائه کند.

رده بندی

دسته بندی به صورت دستی علاوه بر داشتن هزینه بالا معایب زیر را دارد:

- برای زمینه های تخصصی خاص نیاز به دانش افراد خبره دارد (مانند بانکهای پزشکی، بانکهای حقوقی)
- از آن جا که برچسب گذاری دستی مبتنی بر دانش و تجربه فرد می باشد، بسیار خطا پذیر است.
- تصمیم دو فرد خبره در برچسب گذاری می تواند متفاوت و حتی ناسازگار باشد .

رده بندی

تعریف رسمی دسته بندی:

مجموعه ای از متون : (yi متعلق به مجموعه c)

$D = \{(d1, y1), \ldots, (di, y2), \ldots, (dn, yn)\}$

مجموعه ای از کلمات در یک مستند:

$di = [wi,1, \ldots, wi,k, \ldots, wi,|di|]$

مجموعه ای از دسته ها :

$C=\{c1, c2, \ldots, c|c|\}$

رده بندی

هدف در دسته بندی متون، استنتاج یک تابع رابطه ای f
است به نحو ی که yi = f(di) باشد.

انواع ابزارهای دسته بند (طبقه بندی کننده) :

- تک برچسبی (دو دویی)
- چندبرچسبی

رده بندی

دو تعریف :

- دسته بندی مبتنی بر متن (DPC): برای هر مستند ما باید تمام دسته هایی را که باید انتخاب شوند را بیابیم.
- دسته بندی مبتنی بر دسته (CPC): برای دسته های موجود باید مستند مناسب را بیابیم که باید جزء آن باشد

از آن جا که ممکن است مجموعه های C یا D از ابتدا موجود نباشد انتخاب روش اهمیت پیدا می کند.

رده بندی

DPC زمانی مناسب است که متن ها در طول زمان عرضه شوند . (پست الکترونیک)

CPC زمانی مناسب است که یک دسته جدید به مجموعه دسته های موجود اضافه شود.

گزینه DPC نسبت به CPC متداولتر و کاربردی تر است .

رده بندی

شاخصهای رده بندی متون:

- مبتنی بر لغت نامه کنترل شده
  - برای هر متن یک یا چند کلیدواژه که محتوای آن را توصیف می کند، تعیین می گردد.
  - این کلیدواژه ها و عبارات کلیدی متعلق به مجموعه محدودی است که اغلب شامل یک گنج واژه سلسله مراتبی (تزاروس) موضوعی می باشند .

رده بندی

- در این کاربرد نیاز به $k1 \leq x \leq k2$ کلید واژه که برای هر متن تعیین شده باشند، می باشد.

- مبتنی بر فراداده

- معمولاً در کتابخانه های دیجیتال برچسب گذاری متون به وسیله ی فرادادههایی که آنها را توصیف می کنند، صورت می گیرد.
- بعضی از این فرادادهها موضوعی هستند که نقششان شرح معنایی متون با استفاده از کلید واژهها یا عبارات کلیدی است.