Natural Language Processing

anoopsarkar.github.io/nlp-class

Programming Languages

C, C++, Java, Python, …

- unambiguous
- fixed
- designed
- learnable?
- known simple semantics

Natural Languages

French, English, Korean, Chinese, Tagalog, …

- ambiguous
- evolving
- transmitted
- learnable
- complex semantics

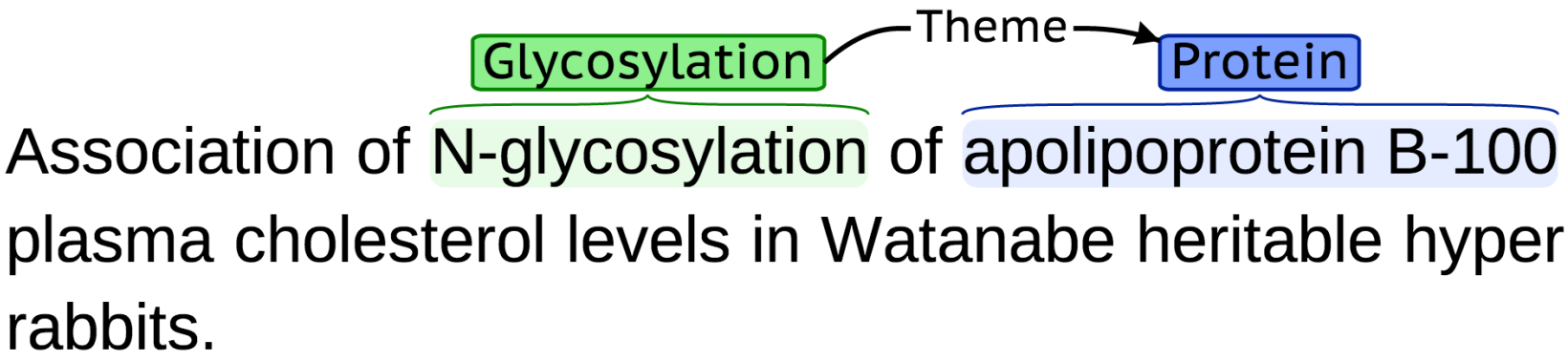
Language is ambiguous

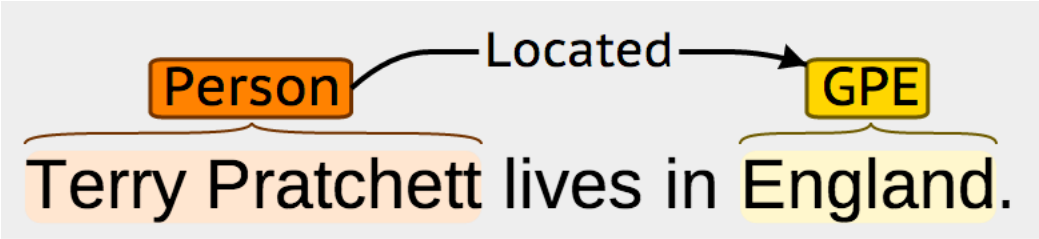
- Lung cancer in women mushrooms
  - Mushrooms is noun or a verb?
- Ban on nude dancing on governor's desk
  - Similar to "if-then-else" ambiguity
- Island Monks Fly in Satellite to Watch Pope Funeral
  - "fly in" vs. "fly [$_{OBJ}$ in Satellite]" hidden segmentation
- British Left Waffles on Falkland Islands
  - Is it British/Noun Left/Verb or British Left/NP Waffles/Verb?

Information Extraction

Así lo explicó hoy el presidente del Gobierno (ORG) español, José María Aznar (PER), en la conferencia de prensa con la que concluyó la XIII Cumbre Hispano-francesa (MISC), celebrada en Santander (LOC), con asistencia del presidente francés, Jacques Chirac (PER) ; del primer ministro, Lionel Jospin (PER), y trece miembros de ambos gabinetes.

Information Extraction





LocalizationID 1022 PSID 10126

1) Select "valid" if the passage contains strong evidence of an experimentally determined localization.

PubMed Entrez PMID 9811664

PubMed Centreal PMCID 107680

The cytoplasmic membrane proteins ExbB and ExbD support TonB-dependent active transport of iron siderophores and vitamin B12 across the essentially unenergized outer membrane of Escherichia coli.

Valid ☐ Invalid ☐ Maybe ☐ Reviewer

Comments

2) If the passage is valid then select whether the protein, organism, and location names are also valid. (If you want to defer your decision then select neither valid nor invalid)

Protein: ExbB

☐ Valid Invalid ☐

Organism: Escherichia coli

☐ Valid Invalid ☐

Location: cytoplasmic membrane

☐ Valid Invalid ☐

SQuASH: SFU QA Summarization System

Input: 25 news articles, Complex question **Output:** 250-word summary

Q. Describe developments in the movement for the independence of Quebec from Canada.

A. Canadian Prime Minister Jean Chretien has ruled out the possibility that the French-speaking province of Quebec could declare independence even if a majority of Quebeckers vote for secession in the next referendum. Premier Lucien Bouchard and the separatist Parti Quebecois have retained control of Canada's largest province, possibly setting the stage for a bitter confrontation with the rest of Canada, and possibly another referendum on independence. Canada's Supreme Court managed to rule on the question of Quebec secession without infuriating leaders on either side of a bitter ideological divide. ...

Headline Generation

Headline A: US launches air raids in Somalia

Headline B: Somalia says dozens killed in US attack

Headline C: Many dead after US strike in Somalia

Headline D: US Launches New Attacks in Somalia

Headline E: US strikes terrorist targets in Somalia

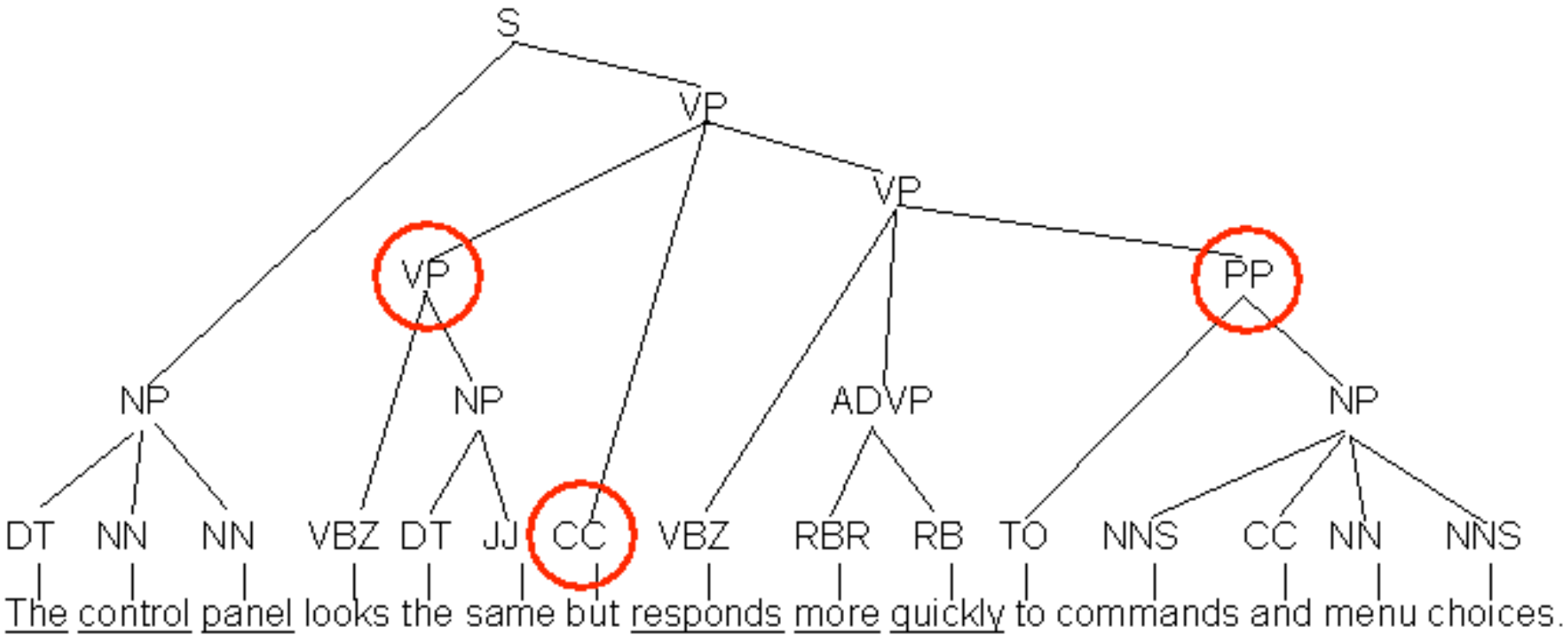
Cluster of headlines for an event on Google News

Headline Generation

Headline Candidate	Score
Bush to sign of	-22.614
Bush to sign bill on	-26.652
Bush to sign of the	-26.835
the House of The Internet gambling	-29.946
The bill of the Internet gambling	-29.982
Bush to end of the Internet gambling	-32.576
Bush to sign bill on the Internet gambling	-35.746
Bush to sign bill on the Internet gambling law	-39.710
Bush to end of the Internet gambling on The Senate bill	-46.988
Bush to sign bill on the Internet gambling site of The law	-50.912

Table 5.9: Top Headlines for "Law on Internet Gambling" news story

Sentence Compression



Paraphrasing

- open borders imply increasing racial fragmentation in ***european countries*** .
- open borders imply increasing racial fragmentation in *the countries of europe* .
- open borders imply increasing racial fragmentation in *european states* .
- open borders imply increasing racial fragmentation in *europe* .
- open borders imply increasing racial fragmentation in *european nations* .
- open borders imply increasing racial fragmentation in *the european countries* .

Why is paraphrasing useful?

http://www.machinedlearnings.com/2011/01/happiness-is-warm-tweet.html

Sentiment detection

Annotate tweets using labels from http://en.wikipedia.org/wiki/List_of_emoticons

10 Happiest Tweets

- @WRiTExMiND no doubt! <--guess who I got tht from? Bwahaha anyway doe I like surprising people it's kinda my thing so ur welcome! And hi :)
- @skvillain yeh wiz is dope, got his own lil wave poppin! I'm fuccin wid big sean too he signed to kanye label g.o.o.d music
- And @pumahbeatz opened for @MarshaAmbrosius & blazed! So proud of him! Go bro! & Marsha was absolutely amazing! Awesome night all around. =)
- Awesome! RT @robscoms: Great 24 hours with nephews. Watched Tron, homemade mac & cheese for dinner, Wii, pancakes & Despicable Me this am!
- Good Morning 2 U Too RT @mzmonique718: Morninggggg twitt birds!...up and getting ready for church...have a good day and LETS GO GIANTS!
- Goodmorning #cleveland, have a blessed day stay focused and be productive and thank god for life
- AMEN!!!>>>RT @DrSanlare: Daddy looks soooo good!!! God is amazing! To GOD be the glory and victory #TeamJesus Glad I serve an awesome God
- AGREED!! RT @ILoveElizCruz: Amen to dat... We're some awesome people! RT @itsVonnell_Mars: @ILoveElizCruz gotta love my sign lol
- #word thanks! :) RT @Steph0e: @IBtunes HAppy Birthday love!!! =) still a fan of ya movement... yay you get another year to be dope!!! YES!!
- Happy bday isaannRT @isan_coy: Selamatt ulang tahun yaaa RT @Phitz_bow: Selamat siangg RT @isan_coy: Slamat pagiiii

http://www.machinedlearnings.com/2011/01/happiness-is-warm-tweet.html

Sentiment detection

Annotate tweets using labels from http://en.wikipedia.org/wiki/List_of_emoticons

10 Saddest Tweets

- Migraine, sore throat, cough & stomach pains. Why me God?
- Ik moet werken omg !! Ik lig nog in bed en ben zo moe .. Moet alleen opstaan en tis koud buitn :(
- I Feel Horrible ' My Voice Is Gone Nd I'm Coughing Every 5 Minutes ' I Hate Feeling Like This :-/
- SMFH !!! Stomach Hurting ; Aggy ; Upset ; Tired ;; Madd Mixxy Shyt Yo !
- Worrying about my dad got me feeling sick I hate this!! I wish I could solve all these problems but I am only 1 person & can do so much..
- Malam2 menggigil+ga bs napas+sakit kepala....badan remuk redam *I miss my husband's hug....#nangismanja#
- Waking up with a sore throat = no bueno. Hoping someone didn't get me ill and it's just from sleeping. D:
- Aaaa ini tenggorokan gak enak, idung gatel bgt bawaannya pengen bersin terus. Calon2 mau sakit nih -___-
- I'm scared of being alone, I can't see to breathe when I am lost in this dream, I need you to hold me?
- Why the hell is suzie so afraid of evelyn! Smfh no bitch is gonna hav me scared I dnt see it being possible its not!

Word Segmentation (in Chinese)

北京大学生体育馆

- 北京 (Beijing) 大学生 (university students) 体育馆 (gym)
The gym for university students in Beijing.

-北京大学 (Peking University) 生 (give birth to) 体育馆 (gym)
Peking University gave birth to the gym?

Statistical Machine Translation

SMT uses parallel corpora to automatically learn a translation

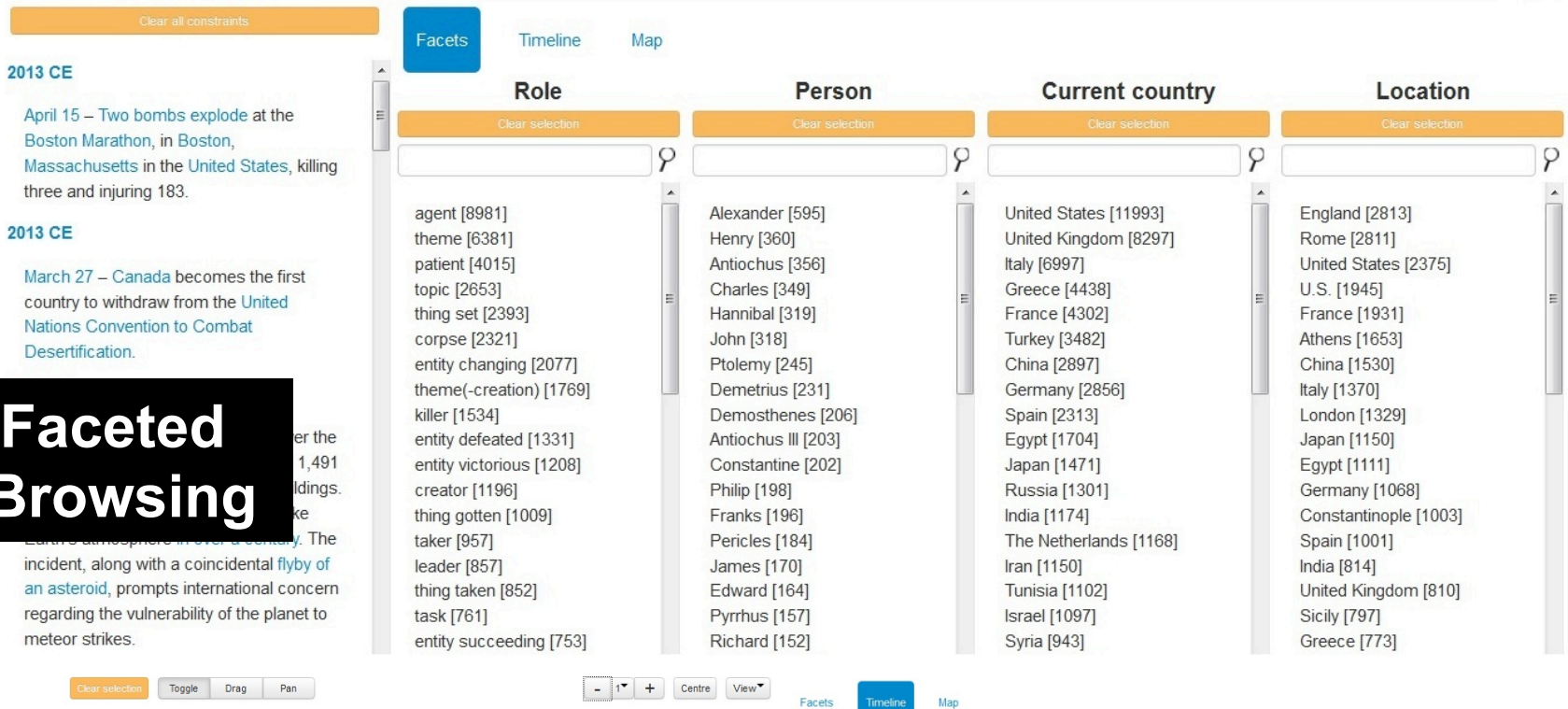
SOURCE: 目前 , 某些 西方 国家 已经 宣布 终止 对 津巴布韦 的 经济援助 .

H1: at present , some western nations have already announced their termination of economic aid to zimbabwe .
H2: at present , certain western countries have already suspended their economic aids to zimbabwe .
H3: so far , some western countries have declared ending economic aid to zimbabwe .
H4: some western countries have already halted economic aid to zinbarbwe at present .

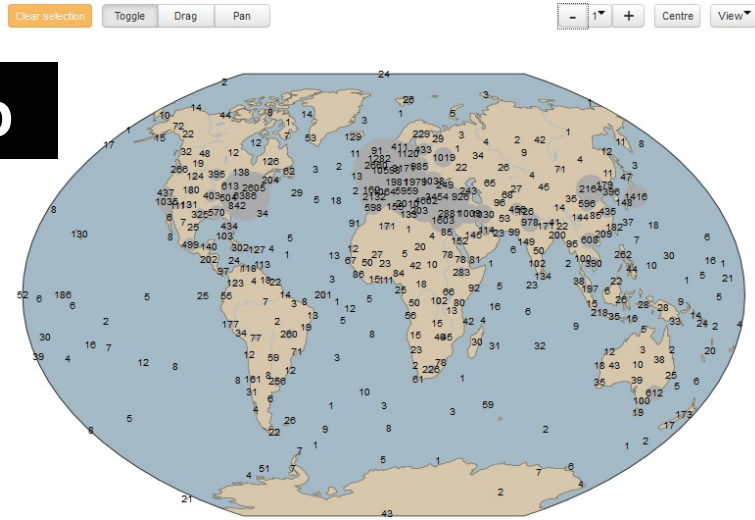
SYSTEM: at present , some western countries have announced the* end* of the* financial* assistance* to zimbabwe .

Open Source Machine Translation! www.statmt.org

Visualization of Information



Faceted Browsing



Map

Timeline

Identifying confusable drug names

G. Kondrak and B. Dorr

Table 4 Top 8 names that are most similar to *Toradol* according to the BI-SIM similarity measure, and the corresponding recall values

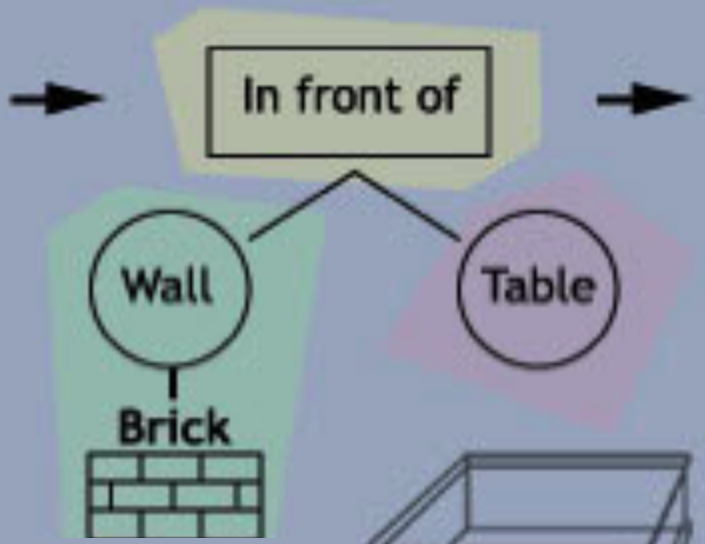
	Name	Score	+/−	Recall
1.	*Tramadol*	0.6875	+	0.25
2.	*Tobradex*	0.6250	−	0.25
3.	*Torecan*	0.5714	+	0.50
4.	*Stadol*	0.5714	−	0.50
5.	*Torsemide*	0.5000	−	0.50
6.	*Theraflu*	0.5000	−	0.50
7.	*Tegretol*	0.5000	+	0.75
8.	*Taxol*	0.5000	−	0.75

Holy Grail: Understanding Language

- Can we *generate* language from our knowledge of language?
- Can we convert a natural language utterance into a *model* (or some other fancy logic thing)
- Can we map it into a *database*?
- Can we map it into a *mental picture* (or a *real* one?)
- Demo: WordsEye (from Richard Sproat's group at AT&T)

Text to semantic model to image

The vase is on the Richard Sproat coffee table. The table is in front of the brick wall. The Van Gogh picture is on the wall. The Matisse sofa is next to the table. Mary is sitting on the sofa. She is playing the violin. She is wearing a straw hat.




details
details
The Devil is in the details
© 2000, AT&T Corp. Produced by Wordseye (TM)