CSCI 5582
Artificial Intelligence

Lecture 23

Jim Martin

Today 11/30

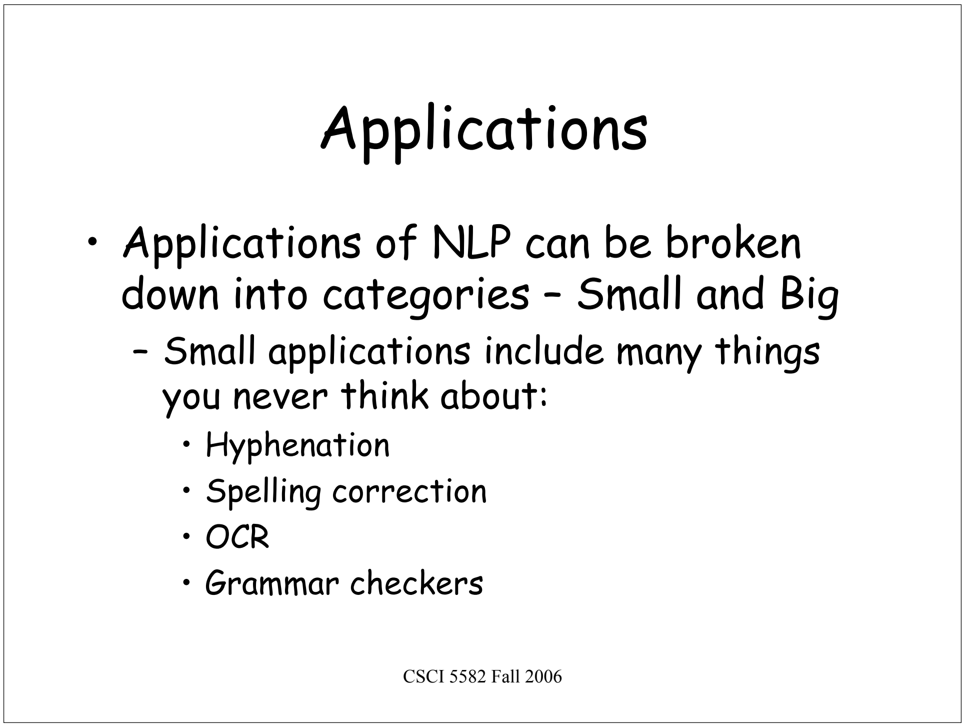
- Natural Language Processing
  - Overview
    - 2 sub-problems
  - Machine Translation
  - Question Answering

Readings

- Chapters 22 and 23 in Russell and Norvig
- Chapter 24 of Jurafsky and Martin

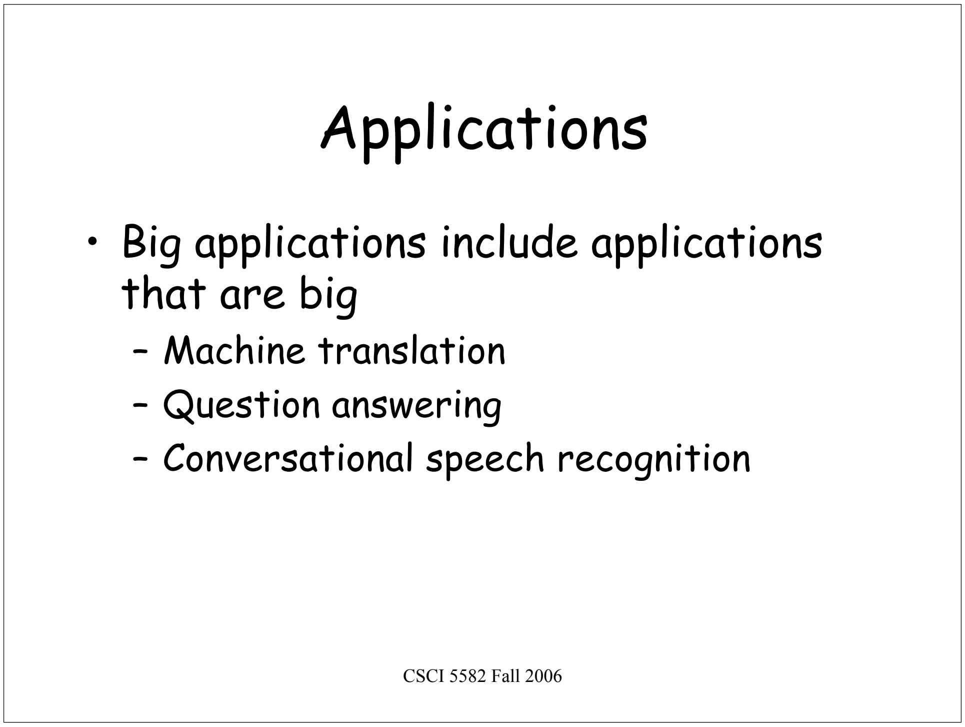
Speech and Language Processing

- Getting computers to do reasonably intelligent things with human language is the domain of Computational Linguistics (or Natural Language Processing or Human Language Technology)



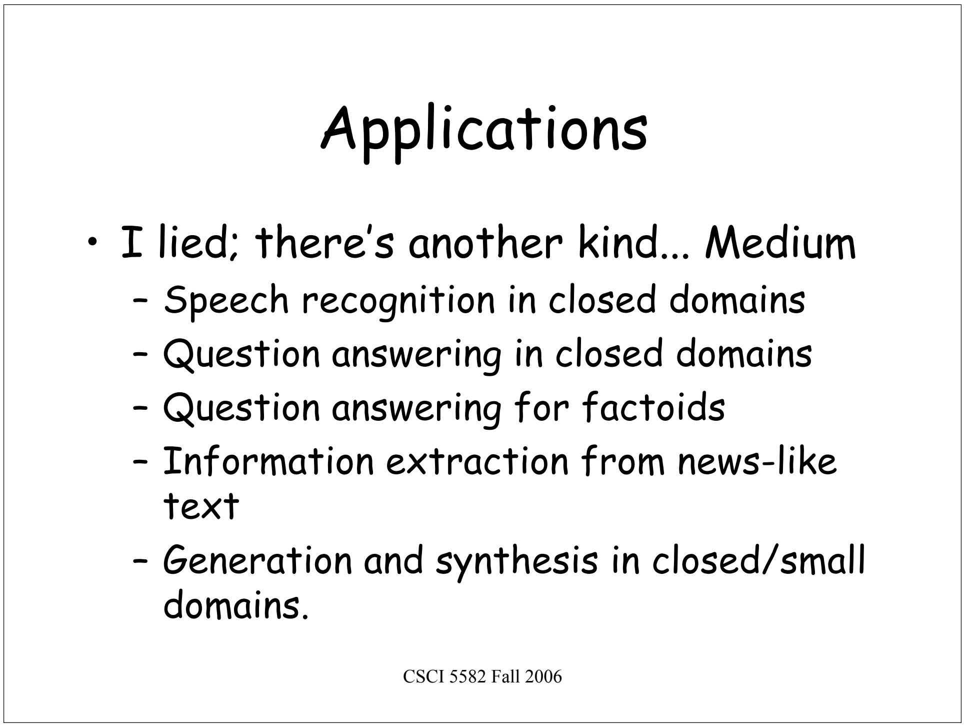
Applications

- Applications of NLP can be broken down into categories – Small and Big
  – Small applications include many things you never think about:
    - Hyphenation
    - Spelling correction
    - OCR
    - Grammar checkers



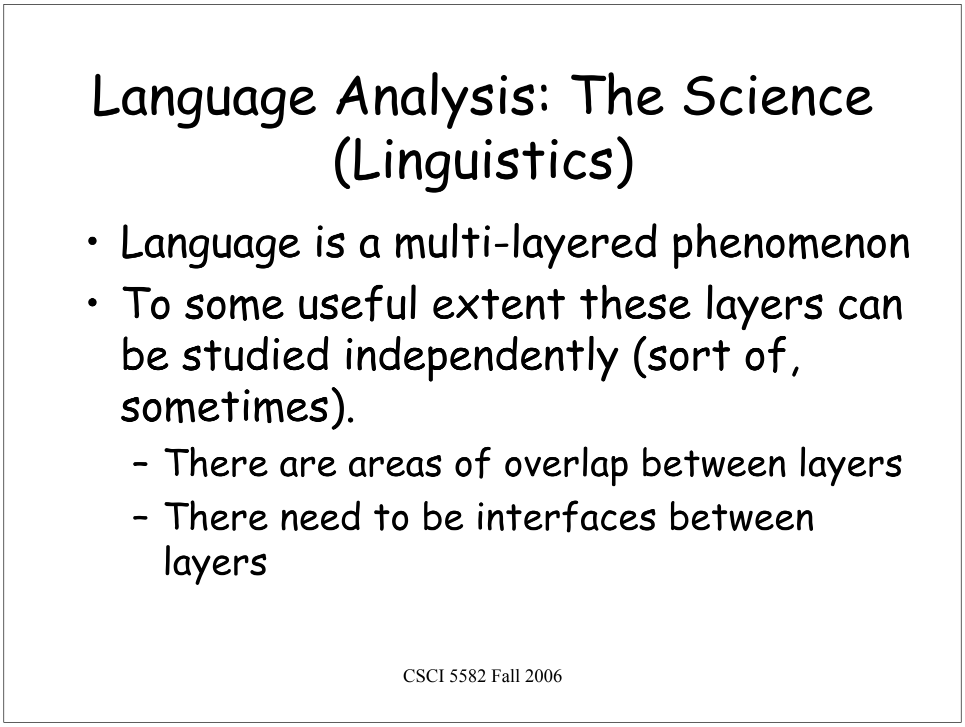
Applications

- Big applications include applications that are big
  – Machine translation
  – Question answering
  – Conversational speech recognition



Applications

- I lied; there's another kind... Medium
  - Speech recognition in closed domains
  - Question answering in closed domains
  - Question answering for factoids
  - Information extraction from news-like text
  - Generation and synthesis in closed/small domains.



Language Analysis: The Science (Linguistics)

- Language is a multi-layered phenomenon
- To some useful extent these layers can be studied independently (sort of, sometimes).
  - There are areas of overlap between layers
  - There need to be interfaces between layers

The Layers

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Phonology

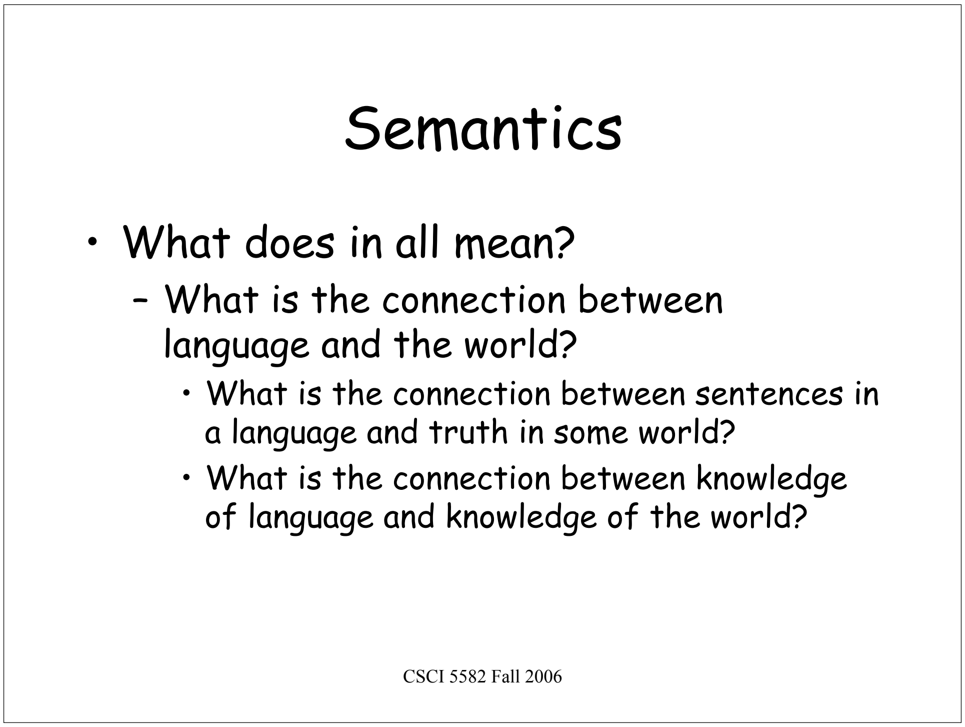
- The noises you make and understand

Morphology

- What you know about the structure of the words in your language, including their derivational and inflectional behavior.

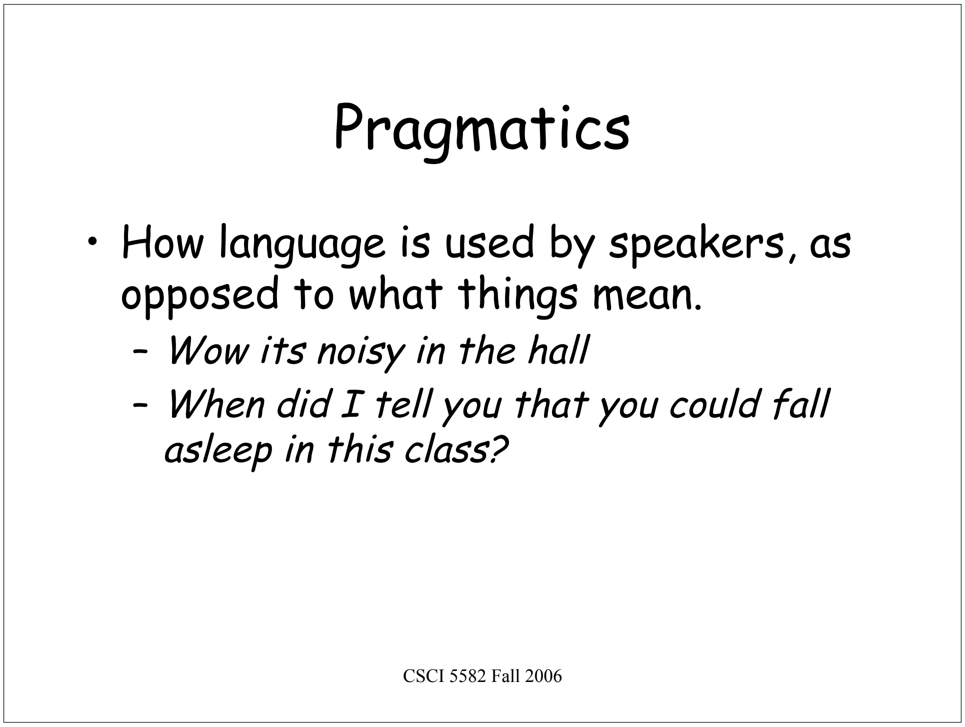
Syntax

- What you know about the order and constituency of the utterances you spout.



Semantics

- What does in all mean?
  - What is the connection between language and the world?
    - What is the connection between sentences in a language and truth in some world?
    - What is the connection between knowledge of language and knowledge of the world?



Pragmatics

- How language is used by speakers, as opposed to what things mean.
  - *Wow its noisy in the hall*
  - *When did I tell you that you could fall asleep in this class?*

Discourse

- Dealing with larger chunks of language
- Dealing with language in context

Break

- Reminders
  - The class is over real soon now
    - Last lecture is 12/14 (review lecture)
  - NLP for the next three classes
  - The final is Monday 12/18, 1:30 to 4

HW Questions

- Testing will be on “normal to largish” chunks of text.
  - I won’t test on single utterances, or words.
  - Each test case will be separated by a blank line.
  - You should design your system with this in mind.

HW Questions

- Code: You can use whatever learning code you can find or write.
- You can’t use a canned solution to this problem. In other words…
  - Yes you can use Naïve Bayes
  - No you can’t just find and use a Naïve Bayes solution to this problem
  - The HW is an exercise in feature development as well as ML.

NLP Research

- In between the linguistics and the big applications are a host of hard problems.
  - Robust Parsing
  - Word Sense Disambiguation
  - Semantic Analysis
  - etc

NLP Research

- Not too surprisingly, solving these problems involves
  - Choosing the right logical representations
  - Managing hard search problems
  - Dealing with uncertainty
  - Using machine learning to train systems to do what we need

Example

- Suppose you worked for a Text-to-Speech company and you encountered the following…
  - I read about a man who played the bass fiddle.

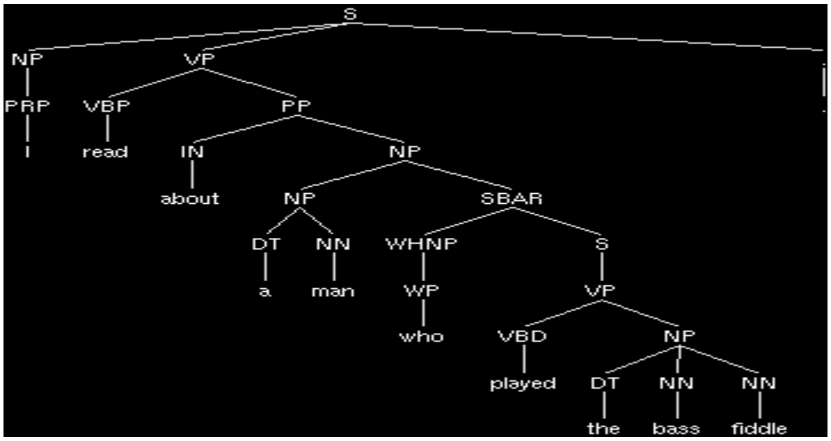
Example

- I read about a man who played the bass fiddle
- There are two separate problems here.
  - For read, we need to know that it’s the past tense of the verb (probably).
  - For bass, we need to know that it’s the musical rather than fish sense.

Solution One

- Syntactically parse the sentence
  - This reveals the past tense
- Semantically analyze the sentence (based on the parse)
  - This reveals the musical use of bass

Syntactic Parse



Solution Two

- Assign part of speech tags to the words in the sentence as a stand-alone task
  - Part of speech tagging
- Disambiguate the senses of the words in the sentence independent of the overall semantics of the sentence.
  - Word sense disambiguation

Solution 2

- I read about a man who played the bass fiddle.

I/**PRP** read/**VBD** about/**IN** a/**DT** man/**NN** who/**WP** played/**VBD** the/**DT** bass/**NN** fiddle/**NN** ./**.**

Part of Speech Tagging

- Given an input sequence of words, find the correct sequence of tags to go along with those words.

  Argmax P(Tags|Words)

  = Argmax P(Words|Tags)P(Tags)/P(Words)
- Example
  - Time flies
  - Minimally time can be a noun or a verb, flies can be a noun or a verb. So the tag sequence could be N V, N N, V V, or V N.
  - So…
    - P(N V | Time flies) = P(Time flies| N V)P(N V)

Part of Speech Tagging

- P(N V|Time flies) = P(Time flies|N V)P(N V)
- First

  P(Time flies|N V) = P(Time|N)*P(Flies|V)
- Then

  P(N V) = P(N)*P(V|N)
- So
  - P(N V| Time flies) =

    P(N)P(V|N)P(Time|Noun)(Flies|Verb)

Part of Speech Tagging

- So given all that how do we do it?

Word Sense Disambiguation

- Ambiguous words in context are objects to be classified based on their context; the classes are the word senses (possibly based on a dictionary.
  - … *played the bass fiddle.*
  - Label *bass* with bass_1 or bass_2

Word Sense Disambiguation

- So given that characterization how do we do it?

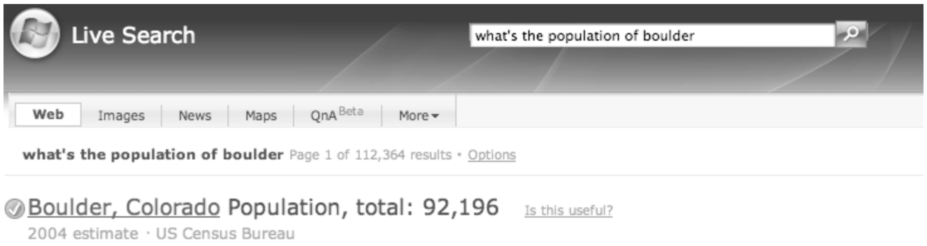
Big Applications

- POS tagging, parsing and WSD are all medium-sized enabling applications.
  - They don’t actually do anything that anyone actually cares about.
  - MT and QA are things people seem to care about.

Q/A

- Q/A systems come in lots of different flavors…
  - We'll discuss open-domain factoidish question answering

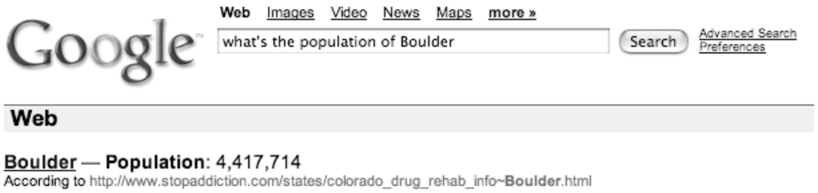
Q/A



what's the population of boulder

what's the population of boulder Page 1 of 112,364 results · Options

Boulder, Colorado Population, total: 92,196 Is this useful?
2004 estimate · US Census Bureau



what's the population of Boulder

Web

Boulder — Population: 4,417,714
According to http://www.stopaddiction.com/states/colorado_drug_rehab_info~Boulder.html

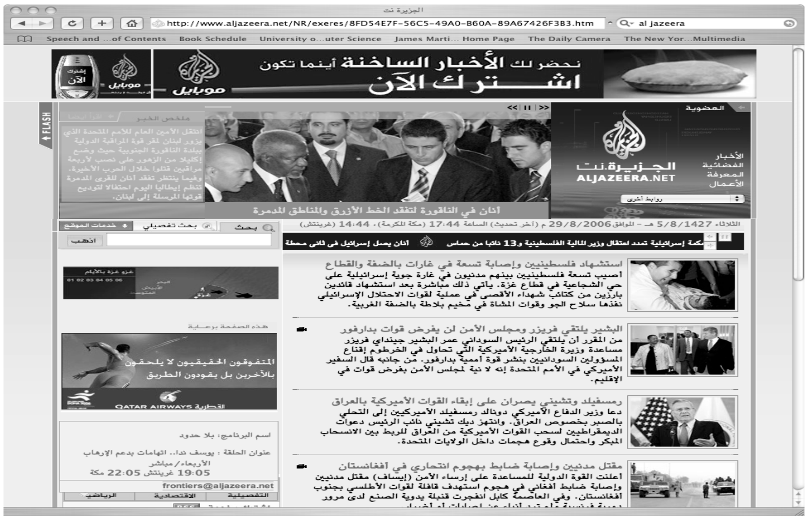
What is MT?

- Translating a text from one language to another automatically.

Warren Weaver (1947)

When I look at an article in Russian, I say to myself: This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.

Google/Arabic



http://www.aljazeera.net/NR/exeres/8FD54E7F-56C5-49A0-B60A-89A67426F3B3.htm

أنان في الناقورة لتفقد الخط الأزرق والمناطق المدمرة

الثلاثاء 5/8/1427 هـ - الموافق 29/8/2006 م (آخر تحديث) الساعة 17:44 (مكة المكرمة)، 14:44 (غرينتش)

محكمة إسرائيلية تمدد اعتقال وزير المالية الفلسطينية و13 نائبا من حماس
أنان يصل إسرائيل في ثاني محطة

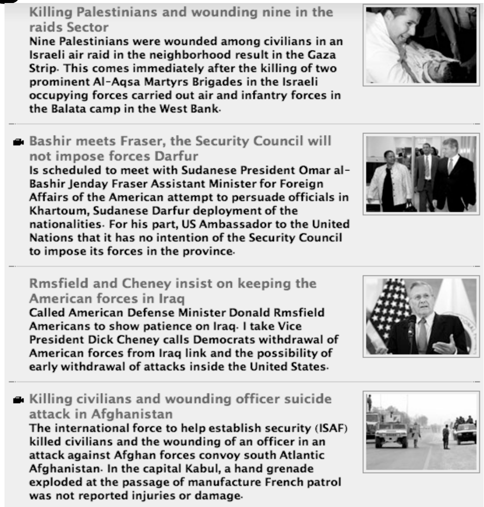
استشهاد فلسطينيين وإصابة تسعة في غارات بالضفة والقطاع
أصيب تسعة فلسطينيين بينهم مدنيون في غارة جوية إسرائيلية على حي الشجاعية في قطاع غزة. يأتي ذلك مباشرة بعد استشهاد قائدين بارزين من كتائب شهداء الأقصى في عملية لقوات الاحتلال الإسرائيلي نفذها سلاح الجو وقوات المشاة في مخيم بلاطة بالضفة الغربية.

البشير يلتقي فريزر ومجلس الأمن لن يفرض قوات بدارفور
من المقرر أن يلتقي الرئيس السوداني عمر البشير جينداي فريزر مساعدة وزيرة الخارجية الأميركية التي تحاول في الخرطوم إقناع المسؤولين السودانيين بنشر قوة أممية بدارفور. من جانبه قال السفير الأميركي في الأمم المتحدة إنه لا نية لمجلس الأمن بفرض قوات في الإقليم.

رمسفيلد وتشيني يصران على إبقاء القوات الأميركية بالعراق
دعا وزير الدفاع الأميركي دونالد رمسفيلد الأميركيين إلى التحلي بالصبر بخصوص العراق. وانتهز ديك تشيني نائب الرئيس دعوات الديمقراطيين لسحب القوات الأميركية من العراق للربط بين الانسحاب المبكر واحتمال وقوع هجمات داخل الولايات المتحدة.

مقتل مدنيين وإصابة ضابط بهجوم انتحاري في أفغانستان
أعلنت القوة الدولية للمساعدة على إرساء الأمن (إيساف) مقتل مدنيين وإصابة ضابط أفغاني في هجوم استهدف قافلة لقوات الأطلسي بجنوب أفغانستان. وفي العاصمة كابل انفجرت قنبلة يدوية الصنع لدى مرور

Google/Arabic Translation



Killing Palestinians and wounding nine in the raids Sector
Nine Palestinians were wounded among civilians in an Israeli air raid in the neighborhood result in the Gaza Strip. This comes immediately after the killing of two prominent Al-Aqsa Martyrs Brigades in the Israeli occupying forces carried out air and infantry forces in the Balata camp in the West Bank.

Bashir meets Fraser, the Security Council will not impose forces Darfur
Is scheduled to meet with Sudanese President Omar al-Bashir Jenday Fraser Assistant Minister for Foreign Affairs of the American attempt to persuade officials in Khartoum, Sudanese Darfur deployment of the nationalities. For his part, US Ambassador to the United Nations that it has no intention of the Security Council to impose its forces in the province.

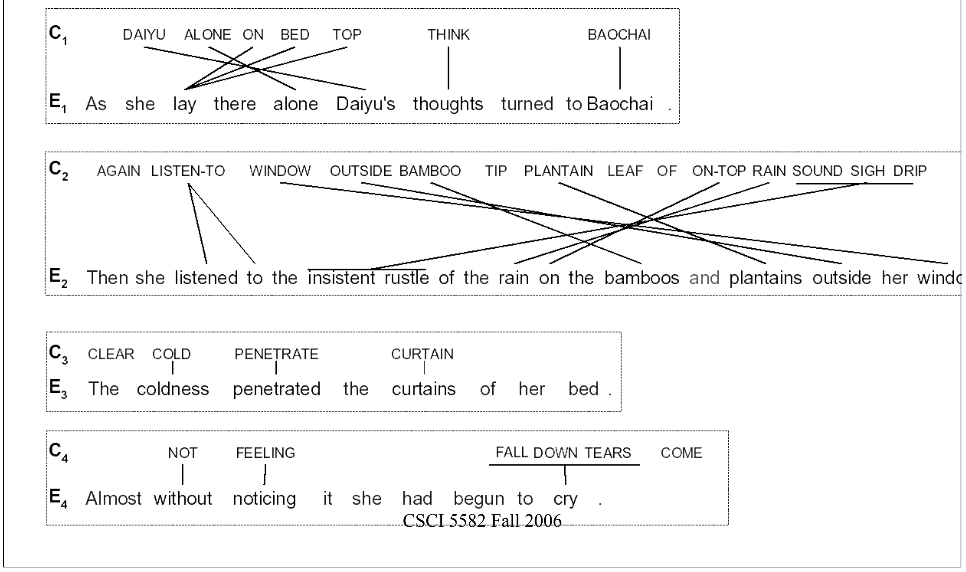
Rmsfield and Cheney insist on keeping the American forces in Iraq
Called American Defense Minister Donald Rmsfield Americans to show patience on Iraq. I take Vice President Dick Cheney calls Democrats withdrawal of American forces from Iraq link and the possibility of early withdrawal of attacks inside the United States.

Killing civilians and wounding officer suicide attack in Afghanistan
The international force to help establish security (ISAF) killed civilians and the wounding of an officer in an attack against Afghan forces convoy south Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of manufacture French patrol was not reported injuries or damage.

Machine Translation

- dai yu zi zai chuang shang gan nian bao chai you ting jian chuang wai zhu shao xiang ye zhe shang, yu sheng xi li, qing han tou mu, bu jue you di xia lei lai.
- Dai-yu alone on bed top think-of-with-gratitude Bao-chai again listen to window outside bamboo tip plantain leaf of on-top rain sound sigh drop clear cold penetrate curtain not feeling again fall down tears come
- As she lay there alone, Dai-yu’s thoughts turned to Bao-chai… Then she listened to the insistent rustle of the rain on the bamboos and plantains outside her window. The coldness penetrated the curtains of her bed. Almost without noticing it she had begun to cry.

Machine Translation



C_1 DAIYU ALONE ON BED TOP THINK BAOCHAI

E_1 As she lay there alone Daiyu's thoughts turned to Baochai .

C_2 AGAIN LISTEN-TO WINDOW OUTSIDE BAMBOO TIP PLANTAIN LEAF OF ON-TOP RAIN SOUND SIGH DRIP

E_2 Then she listened to the insistent rustle of the rain on the bamboos and plantains outside her windo

C_3 CLEAR COLD PENETRATE CURTAIN

E_3 The coldness penetrated the curtains of her bed .

C_4 NOT FEELING FALL DOWN TEARS COME

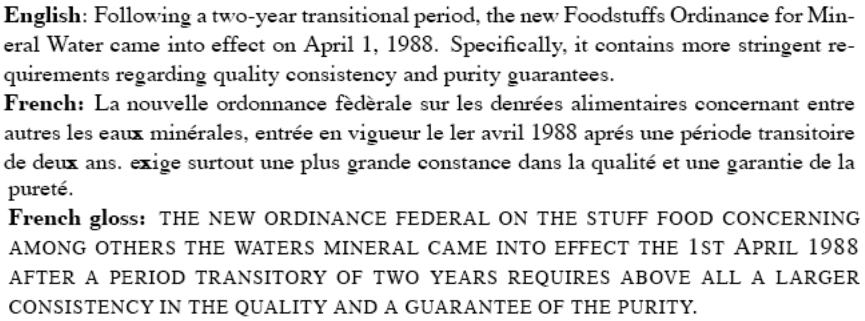
E_4 Almost without noticing it she had begun to cry .

Machine Translation

- Issues:
  - Word segmentation
  - Sentence segmentation: 4 English sentences to 1 Chinese
  - Grammatical differences
    - Chinese rarely marks tense:
      - As, turned to, had begun,
      - *tou* -> penetrated
    - Zero anaphora
    - No articles
  - Stylistic and cultural differences
    - Bamboo tip plaintain leaf -> bamboos and plantains
    - Ma 'curtain' -> curtains of her bed
    - Rain sound sigh drop -> insistent rustle of the rain

Not just literature

- Hansards: Canadian parliamentary



English: Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.

French: La nouvelle ordonnance fèdèrale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le ler avril 1988 aprés une période transitoire de deux ans. exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

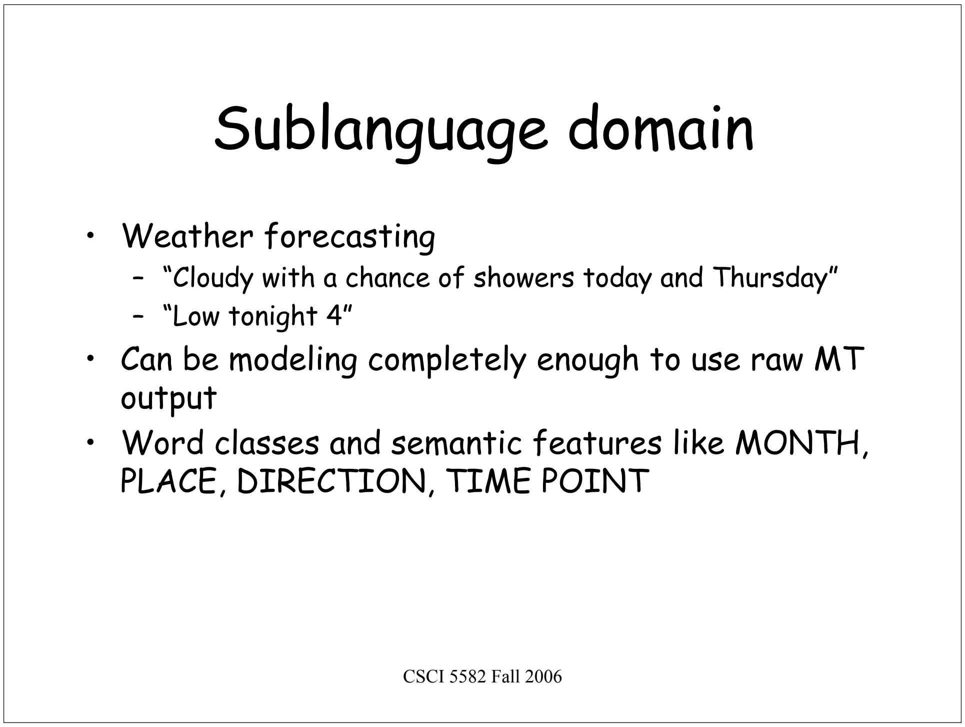
French gloss: THE NEW ORDINANCE FEDERAL ON THE STUFF FOOD CONCERNING AMONG OTHERS THE WATERS MINERAL CAME INTO EFFECT THE 1ST APRIL 1988 AFTER A PERIOD TRANSITORY OF TWO YEARS REQUIRES ABOVE ALL A LARGER CONSISTENCY IN THE QUALITY AND A GUARANTEE OF THE PURITY.

What is MT not good for?

- Really hard stuff
  - Literature
  - Natural spoken speech (meetings, court reporting)
- Really important stuff
  - Medical translation in hospitals, 911 calls

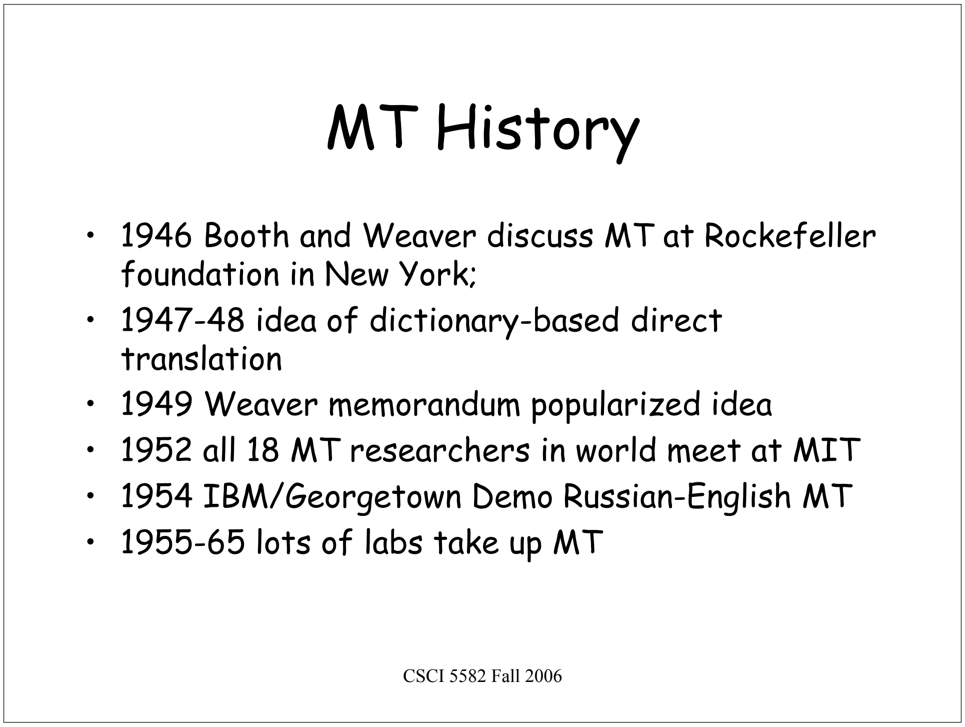
What is MT good for?

- Tasks for which a rough translation is fine
  - Web pages, email
- Tasks for which MT can be post-edited
  - MT as first pass
  - "Computer-aided human translation
- Tasks in sublanguage domains where high-quality MT is possible
  - FAHQT



Sublanguage domain

- Weather forecasting
  - “Cloudy with a chance of showers today and Thursday”
  - “Low tonight 4”
- Can be modeling completely enough to use raw MT output
- Word classes and semantic features like MONTH, PLACE, DIRECTION, TIME POINT



MT History

- 1946 Booth and Weaver discuss MT at Rockefeller foundation in New York;
- 1947-48 idea of dictionary-based direct translation
- 1949 Weaver memorandum popularized idea
- 1952 all 18 MT researchers in world meet at MIT
- 1954 IBM/Georgetown Demo Russian-English MT
- 1955-65 lots of labs take up MT

History of MT: Pessimism

- 1959/1960: Bar-Hillel “Report on the state of MT in US and GB”
  - Argued FAHQT too hard (semantic ambiguity, etc)
  - Should work on semi-automatic instead of automatic
  - His argument
    Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy.
  - Only human knowledge let’s us know that ‘playpens’ are bigger than boxes, but ‘writing pens’ are smaller
  - His claim: we would have to encode all of human knowledge

History of MT: Pessimism

- The ALPAC report
  - Headed by John R. Pierce of Bell Labs
  - Conclusions:
    - Supply of human translators exceeds demand
    - All the Soviet literature is already being translated
    - MT has been a failure: all current MT work had to be post-edited
    - Sponsored evaluations which showed that intelligibility and informativeness was worse than human translations
  - Results:
    - MT research suffered
      - Funding loss
      - Number of research labs declined
      - Association for Machine Translation and Computational Linguistics dropped MT from its name

History of MT

- 1976 Meteo, weather forecasts from English to French
- Systran (Babelfish) been used for 40 years
- 1970's:
  - European focus in MT; mainly ignored in US
- 1980's
  - ideas of using AI techniques in MT (KBMT, CMU)
- 1990's
  - Commercial MT systems
  - Statistical MT
  - Speech-to-speech translation

Language Similarities and Divergences

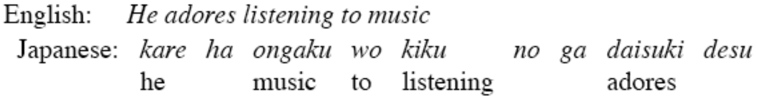
- Some aspects of human language are universal or near-universal, others diverge greatly.
- Typology: the study of systematic cross-linguistic similarities and differences
- What are the dimensions along with human languages vary?

Morphological Variation

- Isolating languages
  - Cantonese, Vietnamese: each word generally has one morpheme
- Vs. Polysynthetic languages
  - Siberian Yupik (`Eskimo'): single word may have very many morphemes
- Agglutinative languages
  - Turkish: morphemes have clean boundaries
- Vs. Fusion languages
  - Russian: single affix may have many morphemes

Syntactic Variation

- SVO (Subject-Verb-Object) languages
  - English, German, French, Mandarin
- SOV Languages



English: *He adores listening to music*

Japanese:	*kare*	*ha*	*ongaku*	*wo*	*kiku*	*no*	*ga*	*daisuki*	*desu*
	he		music	to	listening			adores	

- VSO languages
  - Irish, Classical Arabic
- SVO lgs generally prepositions: to Yuriko
- VSO lgs generally postpositions: Yuriko ni

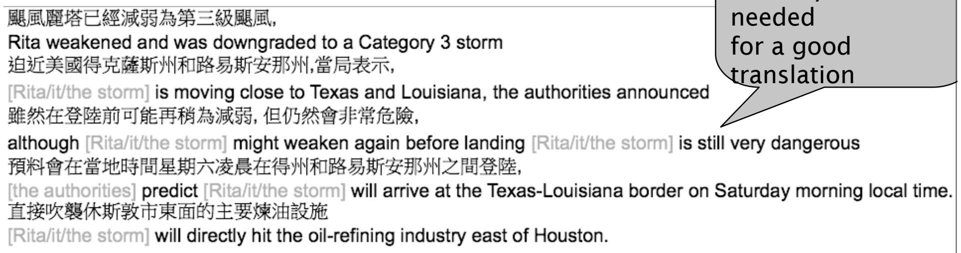
Segmentation Variation

- Not every writing system has word boundaries marked
  - Chinese, Japanese, Thai, Vietnamese
- Some languages tend to have sentences that are quite long, closer to English paragraphs than sentences:
  - Modern Standard Arabic, Chinese

Inferential Load: cold vs. hot lgs

- Some 'cold' languages require the hearer to do more "figuring out" of who the various actors in the various events are:
  - Japanese, Chinese,
- Other 'hot' languages are pretty explicit about saying who did what to whom.
  - English

Inferential Load (2)



颶風麗塔已經減弱為第三級颶風,
Rita weakened and was downgraded to a Category 3 storm
迫近美國得克薩斯州和路易斯安那州,當局表示,
[Rita/it/the storm] is moving close to Texas and Louisiana, the authorities announced
雖然在登陸前可能再稍為減弱, 但仍然會非常危險,
although [Rita/it/the storm] might weaken again before landing [Rita/it/the storm] is still very dangerous
預料會在當地時間星期六凌晨在得州和路易斯安那州之間登陸,
[the authorities] predict [Rita/it/the storm] will arrive at the Texas-Louisiana border on Saturday morning local time.
直接吹襲休斯敦市東面的主要煉油設施
[Rita/it/the storm] will directly hit the oil-refining industry east of Houston.

Noun phrases in blue do not appear in Chinese text …
But they are needed
for a good translation

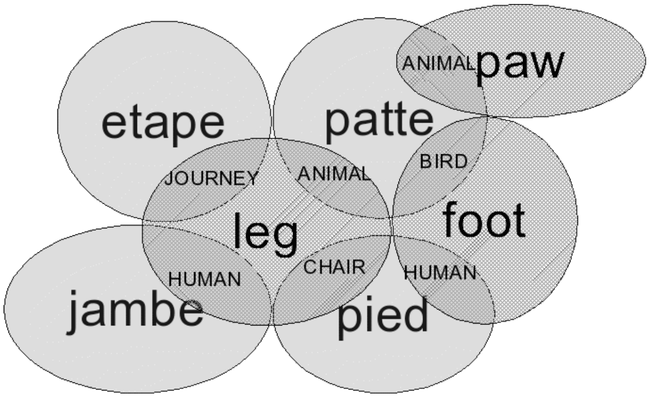
Lexical Divergences

- Word to phrases:
  - English “computer science” = French “informatique”
- POS divergences
  - Eng. ‘she likes/VERB to sing’
  - Ger. Sie singt gerne/ADV
  - Eng ‘I’m hungry/ADJ
  - Sp. ‘tengo hambre/NOUN

Lexical Divergences: Specificity

- Grammatical constraints
  - English has gender on pronouns, Mandarin not.
    - So translating “3rd person” from Chinese to English, need to figure out gender of the person!
    - Similarly from English “they” to French “ils/elles”
- Semantic constraints
  - English ` brother’
  - Mandarin ‘gege’ (older) versus ‘didi’ (younger)
  - English ‘wall’
  - German ‘Wand’ (inside) ‘Mauer’ (outside)
  - German ‘Berg’
  - English ‘hill’ or ‘mountain’

Lexical Divergence: many-to-many



Lexical Divergence: lexical gaps

- Japanese: no word for privacy
- English: no word for Cantonese 'haauseun' or Japanese 'oyakoko' (something like `filial piety')
- English 'cow' versus 'beef', Cantonese 'ngau'

Event-to-argument divergences

- English
  - The bottle floated out.
- Spanish
  - La botella salió flotando.
  - The bottle exited floating
- Verb-framed lg: mark direction of motion on verb
  - Spanish, French, Arabic, Hebrew, Japanese, Tamil, Polynesian, Mayan, Bantu familiies
- Satellite-framed lg: mark direction of motion on satellite
  - Crawl out, float off, jump down, walk over to, run after
  - Rest of Indo-European, Hungarian, Finnish, Chinese

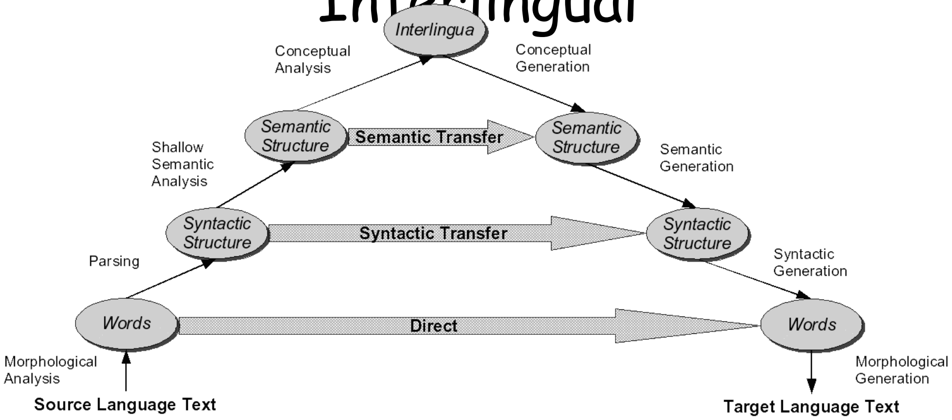
MT on the web

- Babelfish
  - http://babelfish.altavista.com/
  - Run by systran
- Google
  - Arabic research system. Otherwise farmed out (not sure to who).

3 methods for MT

- Direct
- Transfer
- Interlingua

Three MT Approaches: Direct, Transfer, Interlingual



Interlingua

Conceptual Analysis

Conceptual Generation

Semantic Structure

Semantic Transfer

Semantic Structure

Shallow Semantic Analysis

Semantic Generation

Syntactic Structure

Syntactic Transfer

Syntactic Structure

Parsing

Syntactic Generation

Words

Direct

Words

Morphological Analysis

Morphological Generation

Source Language Text

Target Language Text

Next Time

- Read Chapters 22 and 23 in Russell and Norvig, and 24 in Jurafsky and Martin