Natural Language Processing

Lecture 27: NLP for Languages
Other Than English

The multilingual world

An Introduction

How many languages are there?

- Ethnologue: over 7400
- Lexico-statistical definition of languages:
  - percentage cognates in a vocabulary list
  - Sometimes decisions about language definition and classification are not explicit
- Linguists disagree about these matters
- We will return to this topic in the section on languages and dialects

http://langscape.umd.edu/map.php

A Common Situation Exemplified

Language

- Village language: **Kachai**
- Local language: **Tangkhul**
- Regional language: **Meithei**
- National language: **Hindi**
- Global language: **English**

Domain

- Family and village life
- Primary school, etc.
- Secondary school, etc.
- Military, etc.
- Higher education, etc.

Global languages: Mandarin, English, French, Arabic, Spanish, Russian, Portuguese, Japanese, German, Italian, Korean, and Turkish

Language Technologies for daily personal use

- Keyboard input
- Auto-complete
- Spell check
- Speech recognition
- Speech synthesis
- Information retrieval, search engines (and morphology)
- Grammar check
- Translation
- Question answering

Language technologies for commercial/government/research use

- Language detection
- Part of speech tagging
- Parsing
- Semantic role labeling
- Named Entity Recognition
- Summarization
- Translation
- Information extraction and question answering

Enabling infrastructure

- Character encoding (e.g. Unicode)
- Fonts and rendering technologies
- Input methods
- Standard orthography/spelling
- Enough text/speech to train models

Which Languages Have Significant Language Technologies?

- Mandarin
- Spanish
- English
- Hindi
- Arabic
- Portuguese
- Bengali
- Russian
- Japanese
- Punjabi
- German
- Javanese
- Wu
- Malay/Indonesian
- Telugu
- Vietnamese
- Korean
- French
- Marathi
- Tamil

Which Languages Have Significant Language Technologies?

- Mandarin
- Spanish
- English
- Hindi
- Arabic
- Portuguese
- Bengali
- Russian
- Japanese
- Punjabi
- German
- Javanese
- Wu
- Malay/Indonesian
- Telugu
- Vietnamese
- Korean
- French
- Marathi
- Tamil

Is it okay that language technologies only exist for the main global languages?

- Even for many of these languages, significant language technologies do not exist.
- But why do we need language technologies in other languages?
  - Why aren't the top twelve languages enough?
  - Why aren't Mandarin, English, French, Arabic, and Spanish enough?
  - Why isn't English enough?

On the proper scope of language technologies

Computers, Languages, and Dialects

A Chinese Example

- In China, there are a very large number of language variants—people speak "Chinese" in widely divergent ways.
- These ways include:
  - Standard Mandarin (Putonghua)
  - Other varieties of Mandarin
  - Shanghinese and other Wu varieties
  - Cantonese and other Yue varieties
  - Hokkienese and other Min varieties
  - Other groups of varieties like Jin, Gan, and Xiang

Fangyan

- Chinese speakers, both linguists and laypeople, refer to these varieties as *fangyan* (方言)
- *Fangyan* is conventionally translated into English as 'dialect,' but it means something drastically different than the English term *dialect* as used by linguists
- For linguists, a **dialect** is a language variety that belongs to a set of mutually-intelligible varieties
- Chinese fangyans are written with the same script and are descended from a common ancestor, but they are often not mutually intelligible
- For English-speaking linguists, the best translation of *fangyan* is probably…

Language

Why?

- They are not necessarily mutually intelligible
  - Different phonology
  - Different morphosyntax
  - Different lexicon, lexical semantics
- In many cases, it is not possible to use the same language technologies for all (or even a large subset) of these language varieties
- But people use these varieties every day, in their day-to-day lives
- Despite efforts by the Chinese state to promote Putonghua, evidence suggests that people will continue using these **languages** for the foreseeable future

Not Unique to China

- There are numerous varieties of Arabic which are conventionally called *dialects*
  - They are not mutually unintelligible
  - Totally different speech (and often language) technologies needed
- India has numerous minority languages which are conventionally called *dialects*
  - Usually not mutually intelligible with regional or dominant languages
  - Totally different language technologies required

Languages differ—

Smaller Languages are Different

A Naïve View

- A prominent NLP researcher once concluded a talk with the assertion that—he having developed English, Arabic, and Mandarin language technologies—the range of existing linguistic phenomena had largely been covered
- Is this likely to be true?
- Truism repeated by linguists
  - The most divergent languages are usually small languages
  - Small languages are also more likely to be structurally complex than global languages
  - Big languages tend to converge towards a structurally simple prototype

An Empirical Perspective

- Typological databases—characterize languages according to structural features (like SVO/SOV/VSO).
- Experiments conducted using typological databases to find which languages are most and least "typical" of languages generally
- Find English to be very **atypical**
- Nevertheless, researchers find significant qualitative differences between big languages and smaller languages

The Take Home Message

- It is unlikely that implementations of language technologies, no matter how clever the machine learning behind them, will be able to deal with languages that are not English in a perfectly language independent way (at least in the near term).

Based on the observations of Kevin Knight and Mark Steedman

A Linguistic Report Card for Google Translate

Humans vs Machines in NLP

Humans

- Pros
  - Understand the structure of language and the differences between languages
  - Write precise rules
- Cons
  - Slow: Takes person-decades to build a good system
  - Fragile: no graceful failure
  - Humans need to know the languages they are working on or have a lot of experience. Humans who know the language and know NLP might not be available.
  - Even when the humans know the languages and NLP, there are not enough humans who are talented enough to do it well. It takes a huge amount of expertise.

Machine Learning

- Pros
  - Nobody needs to know the language. The techniques are **"language independent".** Specially trained human linguists are not required.
  - Fast: Once you have data, just fire up your model.
  - Robust: Graceful treatment of unseen data.
- Cons
  - Make strange mistakes that a human wouldn't make.
  - Require a lot of data. Not all languages have enough data. But that's ok because we can get funding for low-resource NLP and cross-lingual transfer of models, and we have something to write papers about.

Promise vs Reality

Machine Learning: Promise

1. Nobody needs to know the language. The techniques are **"language independent".** Specially trained humans are not required.
2. Fast: Once you have data, just fire up your model.
3. Robust: Graceful treatment of unseen data.

Machine Learning: Reality

1. The techniques do not work equally well for all languages, and nobody knows why because they don't employ specially trained humans.
2. The first versions can come up quickly, but it has taken person-centuries and hundreds of millions of dollars to refine NLP systems to the current level of performance in English, Chinese, and Arabic.
3. Chronically unable to handle some basic linguistic constructions, resulting in wrong meanings.

Arabic vs Chinese MT

Using Google Translate, April 4, 2016

- Original
  - The Supreme Court ruled unanimously that states may count all residents in drawing election districts, whether or not they are eligible to vote.
- English-Arabic-English
  - The Supreme Court ruled unanimously that the role may count all residents in drawing electoral districts, whether or were not eligible to vote.
- English-Chinese-English
  - Supreme Court unanimously ruled that the State could not expect all residents in drawing the selection, regardless of whether they are eligible to vote.

Arabic vs Chinese MT

Using Google Translate, May 1, 2017

- Original
  - The Supreme Court ruled unanimously that states may count all residents in drawing election districts, whether or not they are eligible to vote.
- English-Arabic-English
  - The Supreme Court unanimously ruled that States may count on all residents of constituencies, whether they are eligible to vote or not.
- English-Chinese-English
  - The Supreme Court unanimously ruled that the states could count all residents' picks, whether or not they were eligible to vote.

Arabic vs Chinese MT (updated)

Using Google Translate, April 17, 2019

- Original
  - The Supreme Court ruled unanimously that states may count all residents in drawing election districts, whether or not they are eligible to vote.
- English-Arabic-English
  - The Supreme Court unanimously ruled that states may count all residents of constituencies, whether they are eligible to vote or not.
- English-Chinese-English
  - The Supreme Court unanimously ruled that states can count all residents in the electoral district, regardless of they are eligible to vote.

The techniques are not language-independent

- The next slides illustrate this for Arabic and Chinese, the two languages that have been most funded by US government projects.
  - Let's say they have equal amount of funding and research effort.
- Arabic consistently gets higher BLEU scores than Chinese.
- The following slides show higher fluency in English-Arabic-English MT than English-Chinese-English MT.
- The following slides show different research issues in Arabic and Chinese.

Level 1: Simplified English

http://www.newsinlevels.com/products/explosion-in-mexico-level-1/

- **Original:** This news is from Mexico. There is an explosion. It happens at a petrochemical plant. Three people die. The explosion injures more than 100 people.
- **Arabic:** This news from Mexico. There explosion. It happens at a petrochemical plant. Three people die. Explosion infects more than 100 people.
- **Chinese:** This message is from Mexico. I had a blast. It occurs in a petrochemical plant. Three deaths. Explosion damage of more than 100 people.

Level 2: Simplified English

- **Original:** On the southern coast of the Gulf of Mexico, there was an explosion at a petrochemical plant. The explosion killed at least 3 people and injured more than 100, including 58 workers
- **Arabic:** On the southern coast of the Gulf of Mexico, there was an explosion at a petrochemical plant. The blast killed at least three people and injuring more than 100, including 58 workers.
- **Chinese:** In the southern coast of the Gulf of Mexico, an explosion occurred in a petrochemical plant. Killing at least three people, injured more than 100, including 58 workers.

Level 3: English-Arabic-English

- **Original:** An explosion has ripped through a petrochemical plant on the southern coast of the Gulf of Mexico, sending a toxin-filled cloud into the air. At least 3 people are known to have died with more than 100 injured, including 58 workers.
- **Arabic:** The explosion occurred at a petrochemical plant on the southern coast of the Gulf of Mexico and a cloud filled with poison in the air. It is known that at least three people were killed with more than 100 injured, including 58 workers.
- **Chinese:** Bombings by the southern coast of the Gulf of Mexico ripped a petrochemical plant sent clouds filled with toxins into the air. At least three people are known to have died more than 100 people were injured, including 58 workers.

Other languages

- MT is not developed as well for other areas of the world as it is for Chinese, Arabic, and English.
  - India, Indonesia, Philippines, Africa, Southeast Asia
- Languages from these areas all bring specific linguistic issues that are not addressed by the current “language independent” methods.

Current MT consistently misses patterns that are obvious to humans

- Grammar conveys who did what to who(m).
- Making errors in subject and object leads to semantic errors in agent and patient.
  - The company bought the bank.
  - The bank bought the company.
  - The bank was bought by the company
  - The company was bought by the bank.
  - This is the company that bought the bank.
  - This is the company that the bank bought.

- The following slides are based on observations by Kevin Knight and Mark Steedman.

Constructions that are obvious to linguists

- Passive:
  - Agent verb-ed patient.
  - Patient was verb-ed by agent.
- Relative clause with subject gap:
  - The company that ___ bought Google.
- Relative clause with object gap:
  - The company that Google bought___.

- We've been checking these in Japanese, Chinese, and Arabic since 2010.

Passive: English-Chinese-English

- My wallet was stolen.
- 我的钱包被偷了。
- Wǒ de qiánbāo bèi tōule.
- My wallet was stolen.

- My friend's wallet was stolen.
- 我朋友的钱包被偷了。
- Wǒ péngyǒu de qiánbāo bèi tōule.
- My friend's wallet was stolen.

Passive: Japanese

- My wallet was stolen.
- Saifu o nusuma remashita.
- 財布を盗まれました。
- My wallet was stolen.

- My friend’s wallet was stolen.
- 私の友人の財布を盗まれました。
- Watashi no yūjin no saifu o nusuma remashita.
- It was stolen my friend's purse.

Passive: Arabic

- My wallet was stolen
- محفظتي سرقت.
- muhaffazati saraqt.
- My wallet was stolen.

- My book was read by many people.
- وقد قرأ كتابي من قبل كثير من الناس.
- waqad qara kitabi min qibal kthyr min alnnas.
- I read my book by many people.

- My friend's wallet was stolen
- سرقت محفظة صديقي.
- saraqat muhfizat sadiqi.
- Stolen purse my friend.

Subject vs Object Gap: English-Arabic-English

- This is the company that bought Google.
- .هذه هي الشركة التي اشترت جوجل
- hadhih hi alshsharikat alty aishtarat jujal .
- This is a company that Google bought.

- This is the company that Google bought.
- .هذه هي الشركة التي اشترت جوجل
- hadhih hi alshsharikat alty aishtarat jawjl.
- This is a company that Google bought.

Subject vs Object Gap: English-Chinese-English

- This is the company that bought Google.
- 这是谷歌收购该公司。
- Zhè shì gǔgē shōugòu gāi gōngsī.
- This is Google acquired the company.

- This is the company that Google bought.
- 这是谷歌收购了该公司。
- Zhè shì gǔgē shōugòule gāi gōngsī.
- This is Google acquired the company.

Subject vs Object Gap: English-Japanese-English

- This is the company that bought Google.
- これは、Googleを買った会社です。
- Kore wa, gūguru o katta kaishadesu.
- This is a company who bought Google.

- This is the company that Google bought.
- これは、Googleが買った会社です。
- Kore wa, gūguru ga katta kaishadesu.
- This is Google bought the company.

Extra Slides

Level 1: English-Arabic-English

- This news is from Mexico. There is an explosion. It happens at a **petrochemical plant**. Three people die. The explosion injures more than 100 people.

  هذا الخبر من المكسيك. هناك انفجار. يحدث ذلك في مصنع للبتروكيماويات. ثلاثة أشخاص يموتون. انفجار يصيب أكثر من 100 شخص.
- This news from Mexico. There explosion. It happens at a petrochemical plant. Three people die. Explosion infects more than 100 people.

Level 1: English-Chinese-English

- This news is from Mexico. There is an explosion. It happens at a **petrochemical plant**. Three people die. The explosion injures more than 100 people.

  這個消息是從墨西哥。有一個爆炸。它發生在一個石化廠。三人死亡。爆炸傷害100餘人。
- This message is from Mexico. I had a blast. It occurs in a petrochemical plant. Three deaths. Explosion damage of more than 100 people.

Level 1: English-Japanese-English

- This news is from Mexico. There is an explosion. It happens at a **petrochemical plant**. Three people die. The explosion injures more than 100 people.
- このニュースは、メキシコからのものです。爆発があります。これは、石油化学プラントで発生します。3人が死亡しています。爆発は、100人以上を傷つけます。
- This news is from Mexico. There is an explosion. This occurs in a petrochemical plant. 3 people have died. Explosion Hurts more than 100 people.

Level 2: English-Arabic-English

- On the southern coast of the Gulf of Mexico, there was an explosion at a **petrochemical plant**. The explosion killed at least 3 people and injured more than 100, including 58 workers.
- على الساحل الجنوبي للخليج المكسيك، كان هناك انفجار في مصنع للبتروكيماويات. أسفر الانفجار عن مقتل 3 أشخاص على الأقل وجرح أكثر من 100، بينهم 58 عامل.
- On the southern coast of the Gulf of Mexico, there was an explosion at a petrochemical plant. The blast killed at least three people and injuring more than 100, including 58 workers.

Level 2: English-Chinese-English

- On the southern coast of the Gulf of Mexico, there was an explosion at a **petrochemical plant**. The explosion killed at least 3 people and injured more than 100, including 58 workers.
- 在墨西哥灣南部海岸，有在石化廠發生爆炸。爆炸造成至少3人，受傷100餘名，其中包括58工人。
- In the southern coast of the Gulf of Mexico, an explosion occurred in a petrochemical plant. Killing at least three people, injured more than 100, including 58 workers.

Level 2: English-Japanese-English

- On the southern coast of the Gulf of Mexico, there was an explosion at a **petrochemical plant**. The explosion killed at least 3 people and injured more than 100, including 58 workers.
- メキシコ湾の南海岸には、石油化学工場で爆発がありました。爆発は、少なくとも3人が死亡、58労働者を含む、100以上の負傷します。
- To the south coast of the Gulf of Mexico, there was an explosion at a petrochemical plant. The explosion, at least three people were killed, including 58 workers and injured more than 100.

Level 3: English-Arabic-English

- An explosion has **ripped** through a petrochemical plant on the southern coast of the Gulf of Mexico, sending a toxin-filled cloud into the air. At least 3 people are known to have died with more than 100 injured, including 58 workers.
- وقد وقع انفجار في مصنع للبتروكيماويات على الساحل الجنوبي للخليج المكسيك وتصاعدت سحابة مليئة السم في الهواء. ومن المعروف أن 3 أشخاص على الأقل لقوا حتفهم مع أكثر من 100 جريح، بينهم 58 عامل.
- The explosion occurred at a petrochemical plant on the southern coast of the Gulf of Mexico and a cloud filled with poison in the air. It is known that at least three people were killed with more than 100 injured, including 58 workers.

Level 3: English-Chinese-English

- An explosion has **ripped** through a petrochemical plant on the southern coast of the Gulf of Mexico, sending a toxin-filled cloud into the air. At least 3 people are known to have died with more than 100 injured, including 58 workers.
- 發生爆炸事件，通過墨西哥灣南部海岸的一個石化廠撕開，發送毒素填充雲到空氣中。至少有3人已知有超過100人受傷死亡，其中包括58工人。
- Bombings by the southern coast of the Gulf of Mexico ripped a petrochemical plant sent clouds filled with toxins into the air. At least three people are known to have died more than 100 people were injured, including 58 workers.

Level 3: English-Japanese-English

- An explosion has **ripped** through a petrochemical plant on the southern coast of the Gulf of Mexico, sending a toxin-filled cloud into the air. At least 3 people are known to have died with more than 100 injured, including 58 workers.
- 爆発は、空気中に毒素が充填されたクラウドを送信し、メキシコ湾の南海岸に石油化学プラントを介してリッピングしました。少なくとも3人は、58労働者を含む、負傷者100以上で死亡していることが知られています。
- Explosion sends a cloud of toxins is filled in the air, it was ripping through a petrochemical plant on the south coast of the Gulf of Mexico. At least three people, including the 58 workers, have been known to have died in the injured more than 100.

The Supreme Court ruled unanimously that states may count all residents in drawing election districts, whether or not they are eligible to vote.

- قضت المحكمة العليا بالإجماع على أن الدول قد عد جميع المقيمين في رسم الدوائر

الانتخابية، سواء كانت أو لم تكن مؤهلة للتصويت.

- qadat almahkamat aleulya bial'iijmae ealaa 'an alddawr qad eud jmye almuqimin fi rusim alddawayir alaintikhabiati, swa' kanat 'aw lm takun muahhalat lilttaswit.
- The Supreme Court ruled unanimously that the role may count all residents in drawing electoral districts, whether or were not eligible to vote.

- The Supreme Court ruled unanimously that states may count all residents in drawing election districts, whether or not they are eligible to vote.
- 最高法院一致裁决，国家可能指望所有居民在绘制选区，不管他们是否有资格投票。
- Zuìgāo fǎyuàn yīzhì cáijué, guójiā kěnéng zhǐwàng suǒyǒu jūmín zài huìzhì xuǎnqū, bùguǎn tāmen shìfǒu yǒu zīgé tóupiào.
- Supreme Court unanimously ruled that the State could not expect all residents in drawing the selection, regardless of whether they are eligible to vote.