An Introduction to Natural Language Processing

Mohammad Fathi

Agenda

- Natural language processing introduction
- Brief history
- NLP applications
- Machine and speech translation
- Future work
- Summary

What is Natural Language Processing?

- A branch of AI in computer science
  - Computational Linguistics if you're a linguist and use computers to study language
  - Natural Language Processing (NLP) if you're a computer scientist and work on applications involving language
- Focused on developing systems that enable interaction between computers and human users (natural language)
  - Includes text and speech

NLP Goal – Zero UI

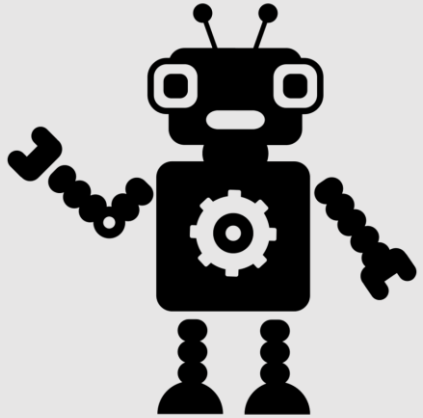
Where is Star Wars playing in San Jose tomorrow?

Star Wars will be playing at the Cinelux Theater on Almaden.

When is it playing there?

It's playing at 2pm, 5pm, and 8pm.

OK. I'd like 2 adults and 2 children for the first show. How much would that cost?



We need linguistic knowledge, domain knowledge, discourse knowledge, world knowledge

History

- 30s – first patents for automatic bilingual dictionary using paper tape
- 50s, 60s – Alan Turing and the “Turing test”, heavy military funding, little understanding of syntax, semantics. Realization that it’s hard!
- 70s, 80s – Ideas that would revolutionize NLP! Foundational work on speech recognition, Rules driven AI, syntactic parsing algorithms
- 90s – probabilistic modeling, supervised learning, more data, powerful machines, realistic expectations.
- 00s – sophisticated statistical modeling and machine learning algorithms, supervised to unsupervised learning

Drivers of NLP

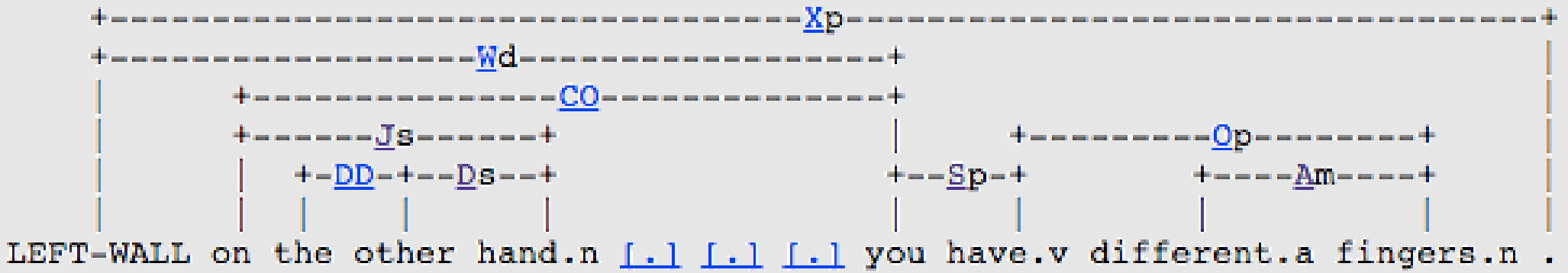
- Ever increasing corpus of machine readable natural language text
  - Generated by humans: web pages, emails, texts, instant messaging, tweets, docs, search history, etc
  - Generated by devices: wearable technology, GPS, Nest devices
- Proliferation of human-computer interaction in natural language
  - Apple’s Siri, Amazon’s Alexa
- Technology advancements
  - Hardware enhancements - beefier machines, server farms, and network, storage and memory advancements
  - Software suites - Fast, scalable data structures, databases for big data support

NLP Applications

mostly solved	making good progress	still really hard
Spam detection OK, let's meet by the big … D*** too small? Buy V1AGRA …	**Sentiment analysis** The pho was authentic and yummy. Waiter ignored us for 20 minutes.	**Semantic search** people protesting globalization Search …demonstrators stormed IMF offices…
Text categorization Phillies shut down Rangers 2-0 SPORTS Jobless rate hits two-year low BUSINESS	**Coreference resolution** Obama told Mubarak he shouldn't run again.	**Question answering (QA)** Q. What currency is used in China? A. The yuan
Part-of-speech (POS) tagging ADJ ADJ NOUN VERB ADV Colorless green ideas sleep furiously.	**Word sense disambiguation (WSD)** I need new batteries for my *mouse*.	**Textual inference & paraphrase** T. Thirteen soldiers lost their lives … H. Several troops were killed in the … YES
Named entity recognition (NER) PERSON ORG LOC Obama met with UAW leaders in Detroit …	**Syntactic parsing** I can see Russia from my house!	**Summarization** Sheen continues rant against … Sheen is nuts
Information extraction (IE) You're invited to our bunga bunga party, Friday May 27 at 8:30pm in Cordura Hall Party May 27 add	**Machine translation (MT)** Our specialty is panda fried rice. 我们的专长是熊猫炒饭	**Discourse & dialog** Where is Thor playing in SF? Metreon at 4:30 and 7:30

Source: Computational Linguistics, Bill MacCartney, Stanford University, 2011.

Hierarchical Structure Representation



Constituent tree:

```
(S (PP On
       (NP the other hand))
   . . .
   (S (NP you)
      (VP have
          (NP different fingers)))
   .)
```

"On the other hand … you have different fingers."

Xp – connect punctuation symbols to words

Wd - used in ordinary declarative sentences

Co - to connect "openers" to subjects of clauses

J– connects prepositions to their objects

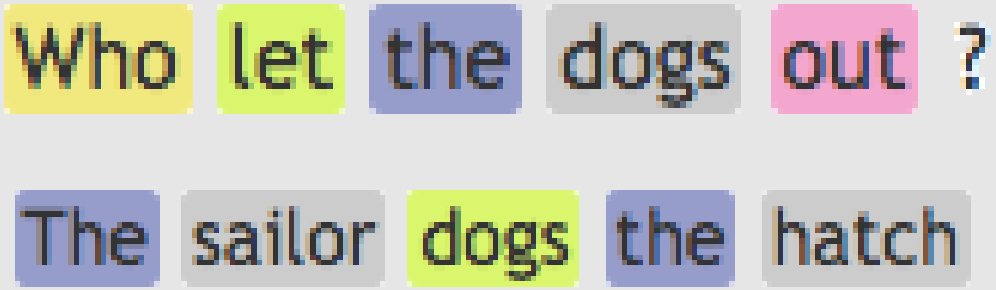
DD - connects definite determiners

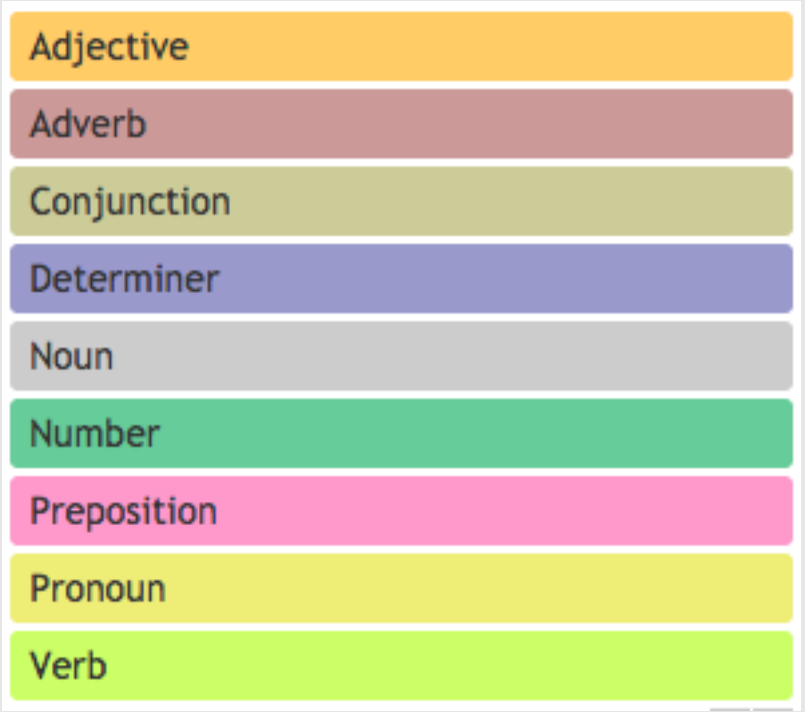
Ds- connects determiners to nouns

Source: http://www.link.cs.cmu.edu/link/dict/section-X.html

Application: Parts of Speech Tagging

Marking up words in text based on the parts of speech





Information Extraction

Your trip confirmation and receipt

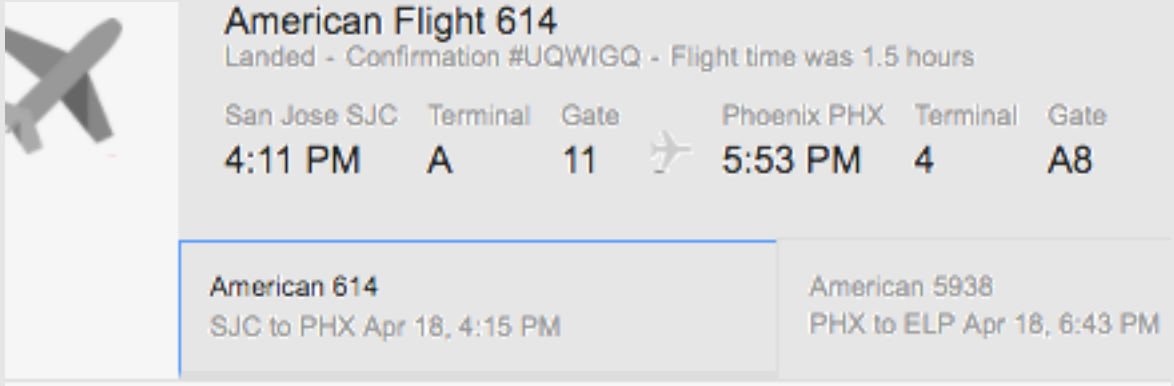
Record locator: **UQWIGQ**

View your trip

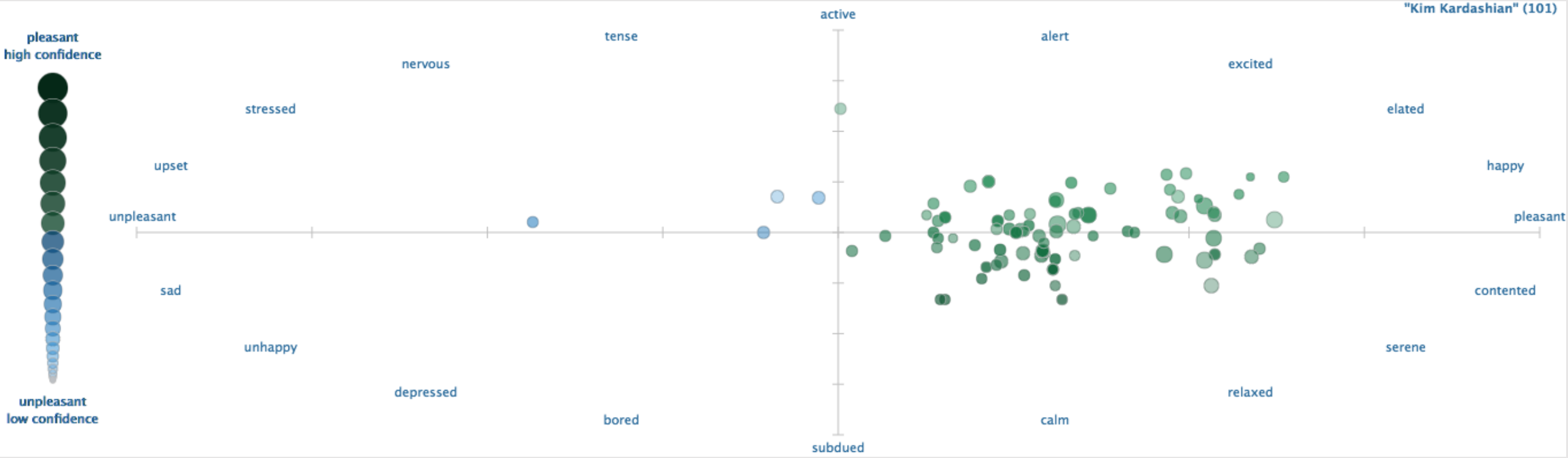
Tuesday, April 18, 2017

SJC 4:15 PM San Jose	→	PHX 6:06 PM Phoenix	Seats: -- Class: Economy (Q) Meals:	American Airlines 614
PHX 6:43 PM Phoenix	→	ELP 8:54 PM El Paso	Seats: -- Class: Economy (Q) Meals:	American Airlines 5938 OPERATED BY MESA AIRLINES AS AMERICAN EAGLE.

Extraction of structured information from unstructured machine readable documents



Sentiment Analysis



Source: https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Sentiment Analysis

Save your $$ and buy the better one.
I have been a photographer for over 50 years. Disliked everything about this camera. Poor quality photos, dark, fuzzy and out of focus. Not intuitive at all. Read more ›
Published 35 minutes ago by James Snyder

good
good
Published 15 hours ago by owner

Delighted
Absolutely what i was looking for, an easy camera with great quality and a reasonable price.
Published 1 day ago by Jay F. A

Five Stars
Great price for great camera!
Published 3 days ago by Nicky

Works OK not great a few minor problems with double clicking when taking ...
Works OK not great a few minor problems with double clicking when taking pictures with a delay on actual picture taken. Would not recommend buying this camera to others.
Published 3 days ago by kittykat23

Great pics
Easy to use
Published 3 days ago by VALLERY HILL

Best pictures and a great buy
Best buy and pictures are great
Published 3 days ago by BigDog

Love this camera
Love this camera! It is so simple to use for a non professional like myself! Right out of the box, very easy clear directions. I need to get a tripod for better long shots. Read more ›
Published 7 days ago by Mick

Five Stars
i like it alot very easy to use and clear pictures
Published 7 days ago by joe moffatt

who really isn't very good at taking photos
I am an amateur photographer, who really isn't very good at taking photos. I wasn't sure how this camera would work out. Read more ›

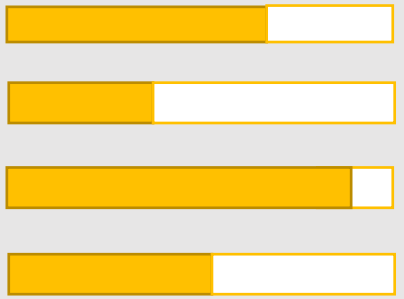
Attributes

Zoom

Weight, size

Flash

Easy of use



Weight, size:

- Very light weight compared to my other cameras
- Nice and compact. Love it!
- Its made of plastic and is flimsy and fragile.

Summarization

“Last week, Arthur Modica, the 76-year-old sculptor of Wall Street's ferocious "Charging Bull," demanded the city of New York remove the "Fearless Girl" sculpture currently obstructing the flow of testosterone toward Wall Street. Modica's lawyer claims both his reputation and his work of art have been severely damaged by the city's decision to permit placement of the bronze little girl near the 18-foot-long bronze bull. Though "Fearless Girl" may face a tough courtroom fight under current United States copyright law, the 4-foot sculpture could find a way to profit from the snarling bull if a lawsuit is filed. Because even if a court of law does not decide in her favor, a court of public opinion will -- as evidenced from the hordes of tourists flocking to her side each day. Instead of filing suit, then, Modica would do better to capitalize on the new tourism opportunities that "Fearless Girl's" addition creates. ”

"Charging Bull's" lawyers make some compelling arguments about how the placement of "Fearless Girl" has damaged their client's artistic rights. They note that "Charging Bull" was created in the aftermath of the 1987 stock market crash as a positive symbol for a demoralized Wall Street, and yet now that symbolism is lost. As the artist himself said in an interview with the New York Post and Marketwatch in March of this year: "I put it there for art. ... My bull is a symbol for America. My bull is a symbol of prosperity and for strength.” He should perhaps add to the list that the bull is also a symbol of a stealth public art placement done in the middle of the night without proper permitting and permission.

Arthur Modica, demanded the city of New York remove the "Fearless Girl" sculpture currently obstructing the flow of testosterone toward Wall Street. Instead of filing suit, then, Modica would do better to capitalize on the new tourism opportunities that "Fearless Girl's" addition creates.”

Source: http://www.cnn.com/2017/04/18/opinions/fearless-girl-lawsuit-outcome-opinion-callan/index.html

Challenges

Challenge	Example
Ambiguity	“I saw the Prudential building flying in to Boston.”
Non standard language	?4u - I have a question for you 10Q – Thank you 2m2h! – Too much to handle!
Idioms	Bob’s your uncle!
Context awareness	" I am a talented individual working for Microsoft Windows core OS team" " I am a janitor, I wash windows at Microsoft. "

Machine Translation

Google Translate

Global surface temperatures in March 2017 were the second-warmest for any March in records dating to the late 19th century, according to three independent analyses. NASA's Goddard Institute for Space Studies calculated the Earth's mean temperature over land and water in March was 1.12 degrees Celsius above average, second only to March 2016's 1.27 degree Celsius departure from average in 137 years of records.

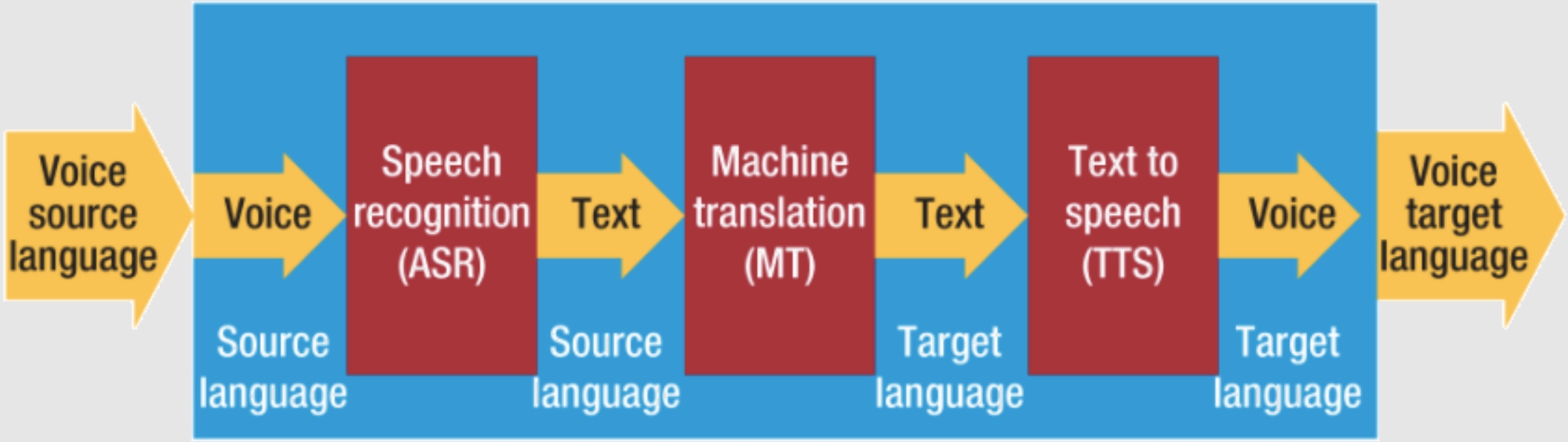
मार्च 2017 में ग्लोबल सतह के तापमान में 1 9वीं सदी के अंत तक होने वाले रिकॉर्ड में किसी भी मार्च के लिए दूसरा सबसे ऊंचा था, तीन स्वतंत्र विश्लेषण के अनुसार
नासा के गोडार्ड इंस्टीट्यूट फॉर स्पेस स्टडीज ने गणना की है कि पृथ्वी का औसत तापमान मार्च के औसत से 1.12 डिग्री सेल्सियस था, मार्च 2016 में 1.27 डिग्री सेल्सियस के औसत से 137 साल के रिकॉर्ड के औसत से दूसरा स्थान है।

Machine Translation

- Two/three steps involved:
  - “Understand” source text
  - Convert that into target language
  - Generate correct target text
- Depends on approach
- Understanding source text involves same problems as for any NLP application
- In addition, “contrastive” problems

Speech to Speech Translation

Conversational spoken phrases are instantly **translated** and spoken aloud in a second language



http://doi.ieeecomputersociety.org/10.1109/MS.2014.14

Challenges

- Words for word never works
- Word orders matter
- Choosing the right words is not trivial
- Rewriting text into another language. Eg. imperative mood in English infinitive in French
- Homographs. Eg. ”fan” a ventilator or an enthusiast
- Morphological: Eg. Chinese and Japanese do not use punctuations
- Misspellings!

Future work

Advanced NLP systems that can

- ✓Providing insights on annual reports, legal and compliance documents, transcripts
  - ✓Understanding of natural language
  - ✓Translation between one natural language to another
- ✓Chat Bots
- ✓Zero UI
- ✓Understand, learn, predict, adapt,…and operate autonomously!

Summary

- NLP is a very challenging area of research
  - Difficulr to cover all the rules and adjust them to all possible languages and variations
- There is a long way to go before MT systems replace human translators
- Machine Translation can be used in applications where the language is very specific