CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computer Science

http://www.cit.ie

Natural Language Processing

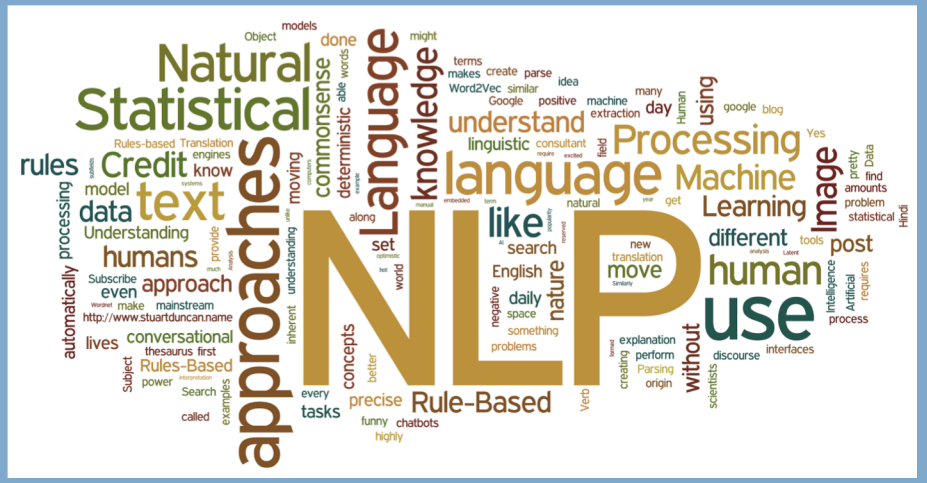
Week1: Introduction

Dr. Haithem Afli

Haithem. afli@cit.ie

@AfliHaithem

2020/2021



Roadmap

- Introduction
  - **Lecturer & module**
  - Topics
  - Syllabus discussion
  - Books and Resources
- Lecture 1

Dr Haithem Afli - Background

- **Computer Science Lecturer at CIT (J102)**
  - NLP, Data Analytics and ML
- **Science Foundation Ireland Funded Investigator**
  - Leader of ADAPT-CIT

- **Research Interest:**
  - Natural Language Processing
  - Social Media and UGC Analysis
  - Machine Translation
  - Data Analytics
- **Lecturing Experience**
  - 5 years in France
  - 3 years in DCU
  - 3 years in CIT

Engaging Content
Engaging People

Science Foundation Ireland sfi For what's next

Le Mans Unive

DCU

CORK INSTITUTE OF TECHNOLOGY

Selected research Projects

- **EU projects, PI** at

Horizon 2020
European Union Funding
for Research & Innovation

STOP
STop Obesity Platfor

ITFLOWS
IT tools and methods for managing migration FLOWS

SLICENET

https://slicenet.eu/

- **Examples of other research topics**

- Image/video caption generation
(Suzanne Considine – PhD, CIT)
- Machine translation of UGC
(Pintu Lohar – PhD, DCU)

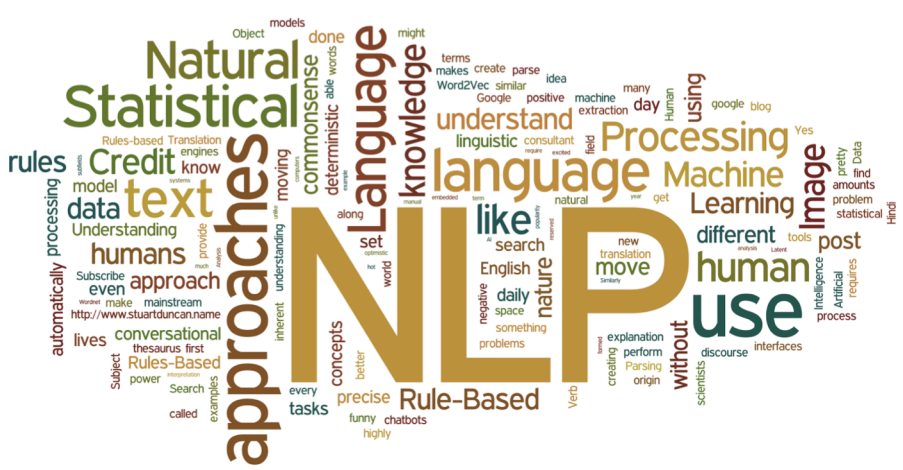
Labs and support

- Praveen Joshi
  - Casual Lecturer at the Department of Computer Science
  - Email: Praveen.joshi@mycit.ie
- Qualification:
  - Masters in Artificial Intelligence – CIT, 2018-2019
  - PhD candidate in AI
- Projects:
  - Slice Net
- Industrial Exp:
  - Infosys
  - Siemens
  - Accenture AI
  - Clear stream
  - AIP
  - Speire

COMP9066 - Natural Language Processing

We're going to study what goes into getting computers to perform useful and interesting tasks involving human languages.

- **Text preprocessing** (stemmings, lemmatization, tokenization, stopword removal)
- **Language models**
- **Text classifiers**
- **Word Embeddings**
- **Semantics**
- **Machine translation**
- **Dialogue Systems**



COMP9066 - Natural Language Processing

Module Workload

Workload: Full Time

Workload Type	*Workload Description*	*Hours*	*Frequency*	*Average Weekly Learner Workload*
Lecture	Delivers the concepts and theories underpinning the learning outcomes.	2.0	Every Week	2.00
Lab	Application of learning to case studies and project work.	2.0	Every Week	2.00
Independent Learning	Student undertakes independent study. The student reads recommended papers and practices implementation.	3.0	Every Week	3.00
			Total Hours	7.00
			Total Weekly Learner Workload	7.00
			Total Weekly Contact Hours	4.00

Workload: Part Time

Workload Type	*Workload Description*	*Hours*	*Frequency*	*Average Weekly Learner Workload*
Lecture	Delivers the concepts and theories underpinning the learning outcomes.	2.0	Every Week	2.00
Lab	Application of learning to case studies and project work.	2.0	Every Week	2.00
Independent Learning	Student undertakes independent study. Student reads recommended papers and practices implementation.	3.0	Every Week	3.00
			Total Hours	7.00
			Total Weekly Learner Workload	7.00
			Total Weekly Contact Hours	4.00

COMP9066 - Natural Language Processing

Assessment Breakdown	%
Course Work	100.00%

Course Work

Assessment Type	*Assessment Description*	*Outcome addressed*	*% of total*	*Assessment Date*
Project	Build a language model and use it in a given natural language processing application such as text generation. Produce a report that critically analyses the performance of the model.	1,2,3	50.0	Week 8
Project	Implement a machine model such as a neural model with vector-based representations for tasks of Machine Translation or Question answering. Assess the performance of the model using standard techniques such as BLEU or WER.	3,4	50.0	Week 12

No End of Module Formal Examination

Reassessment Requirement

Coursework Only
This module is reassessed solely on the basis of re-submitted coursework. There is no repeat written examination.

The institute reserves the right to alter the nature and timings of assessment

Roadmap

- Introduction
  - Lecturer & module
  - **Topics**
  - Syllabus discussion
  - Books and Resources
- Lecture 1

Natural Language Processing (NLP)

- Also known as
  - Computational Linguistics
  - Human Language Technology
- NLP is an interdisciplinary field
  - Computer science
  - Linguistics
  - Cognitive science, psychology, pedagogy, mathematics, etc.
- Applied natural language processing
  - Develop practical applications modeling human languages
- Theoretical computational linguistics
  - Focus on theoretical linguistics and cognitive science

Natural Language Processing

- Applications
  - Machine Translation (MT)
  - Information Retrieval (IR)
  - Automatic Speech Recognition (ASR)
  - Optical Character Recognition (OCR)
  - Automatic Summarization, Speech Synthesis, etc.
- Enabling Technologies
  - Tokenization
  - Part-of-Speech Tagging
  - Syntactic Parsing
  - Lemmatization
  - Word Sense Disambiguation, etc.

Natural Language Processing

- Rule-based/Symbolic Approaches
  - Linguists write rules that are applied by the machines
- Corpus-based/Statistical Approaches
  - Machines learn the “rules” from training data
    - Annotated data – supervised methods
      - Parallel Corpora: translated text collections
      - Treebanks: manually syntactically analyzed texts
      - Speech Corpora with transcripts
    - Unannotated data – unsupervised methods
    - Semi-supervised methods
  - Machine learning approaches are dominant in the field
- Hybrid Approaches
  - The best of **Smart**/Slow Humans and Dumb/**Fast** Machines

Class Topics

- Basic text processing
- Finite state machines and word morphology modeling
- Language modeling
- Text Classification
- Sentiment analysis
- Conversational systems
- Machine translation
- A quick review of information retrieval.

Assignments

1. Basic text processing
2. **Language Modelling**
3. **Text Classification**
4. **Machine translation / conversational systems**

Roadmap

- Introduction
  - Lecturer & module
  - Topics
  - Syllabus discussion
  - **Books and Resources**
- Lecture 1

Books and Resources

- **JM:** Daniel Jurafsky and James H. Martin. "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition." (2nd or 3rd Edition).

- **MH:** Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

  https://nlp.stanford.edu/fsnlp/

Books and Resources

- Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems By Aurélien Géron.

Though not particularly dedicated to natural language processing, this practice-oriented book presents the most popular libraries that may be used for NLP and text analysis.

- Unix Lab – contact the IT service desk to get an account if you do not have one.

Programming Resources

We will be using Python 3 as our programming language in this module.

Basic Python

Python 3 Tutorial – Clear and focused overview of Python 3 syntax, control structures, data structures etc.

Video Python 3 Tutorials – A set of very basic Python 3 video tutorials. More focused on beginners.

NumPy and Pandas

DataCamp NumPy Tutorial – Accessible and easy to understand tutorial to get started with NumPy.

NumPy Tutorial – Short overview of NumPy and basic Python data structures. It also covers SciPy (which you don't need) and basic Matplotlib (which you will be covering later in the programme as part of visualization).

DataCamp Pandas Tutorial – Short and easy to understand tutorial on using Pandas.

Libraries and Tools

Spacy
NLTK
Transformers (Huggingface)
Gensim
Stanza
AllenNLP
Fast.ai
pattern
TextBlob
CoreNLP
CaMel Tools (https://github.com/CAMeL-Lab/camel_tools)

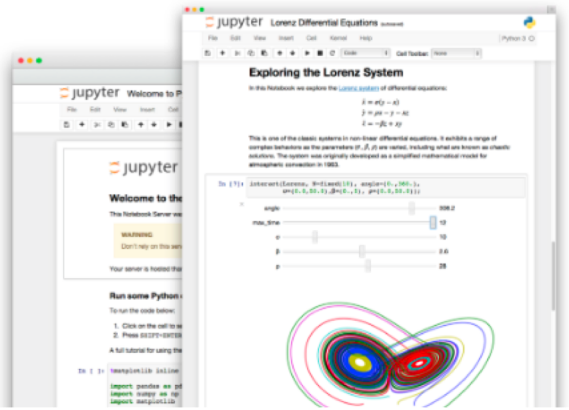
The Jupyter Notebook

jupyter INSTALL PROJECT DOCUMENTATION BLOG DONATE

https://jupyter.org/

Jupyter Notebook

The Jupyter Notebook is a web-based interactive computing platform that allows users to author data- and code-driven narratives that combine live code, equations, narrative text, visualizations, interactive dashboards and other media.



Discussion

Roadmap

- Introduction
  - Lecturer & module
  - Topics
  - Syllabus discussion
  - Books and Resources
- **Lecture 1**

Natural Language Processing

- We’re going to study what goes into getting computers to perform useful and interesting tasks involving human languages.
- We are also concerned with the insights that such computational work gives us into human processing of language.

Why Should You Care?

Important trends

1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication
3. Much of human-human communication is now mediated by computers

Very cool stuff! And with lots of commercial interest.

Machine Translation

LIFESTYLE JULY 5, 2018 / 4:32 PM / 2 MONTHS AGO

At Russia's World Cup, Google Translate breaks language barriers

3 MIN READ

KAZAN, Russia (Reuters) - Soccer might be the most universal language on the planet. But when it comes to deciphering the Cyrillic alphabet or communicating with locals at the World Cup in Russia, the love of the game is sometimes not enough.

Weblog Analytics

- Data-mining of Weblogs, discussion forums, message boards, user groups, and other forms of user generated media
  - Product marketing information
  - Political opinion tracking
  - Social network analysis
  - Buzz analysis (what's hot, what topics are people talking about right now).

Social Media Analysis

ADAPT Centre for Digital Content Technology analyse Irish General Election activity

16 Feb 2016

Researchers at the ADAPT Centre for Digital Content Technology are working with RTE to monitor and analyse activity on Twitter related to the 2016 Irish General Election. The system that has been developed within ADAPT can quickly track the volume of tweets and gauge reaction and sentiment from the tweets in relation to election topics, parties and candidates. The project presents a graphical analysis of tweets that can be filtered by party or candidate allowing the user quickly understand and gain insights into the conversations on election topics taking place on Twitter.

http://sma.adaptcentre.ie/ge16/#!/

Haithem Afli, Sorcha McGuire, and Andy Way. 2017. Sentiment translation for low resourced languages: Experiments on irish general election tweets. In 18th International Conference on Computational Linguistics and Intelligent Text Processing.

Information Extraction & Sentiment Analysis

Attributes:

- zoom
- affordability
- size and weight
- flash
- ease of use

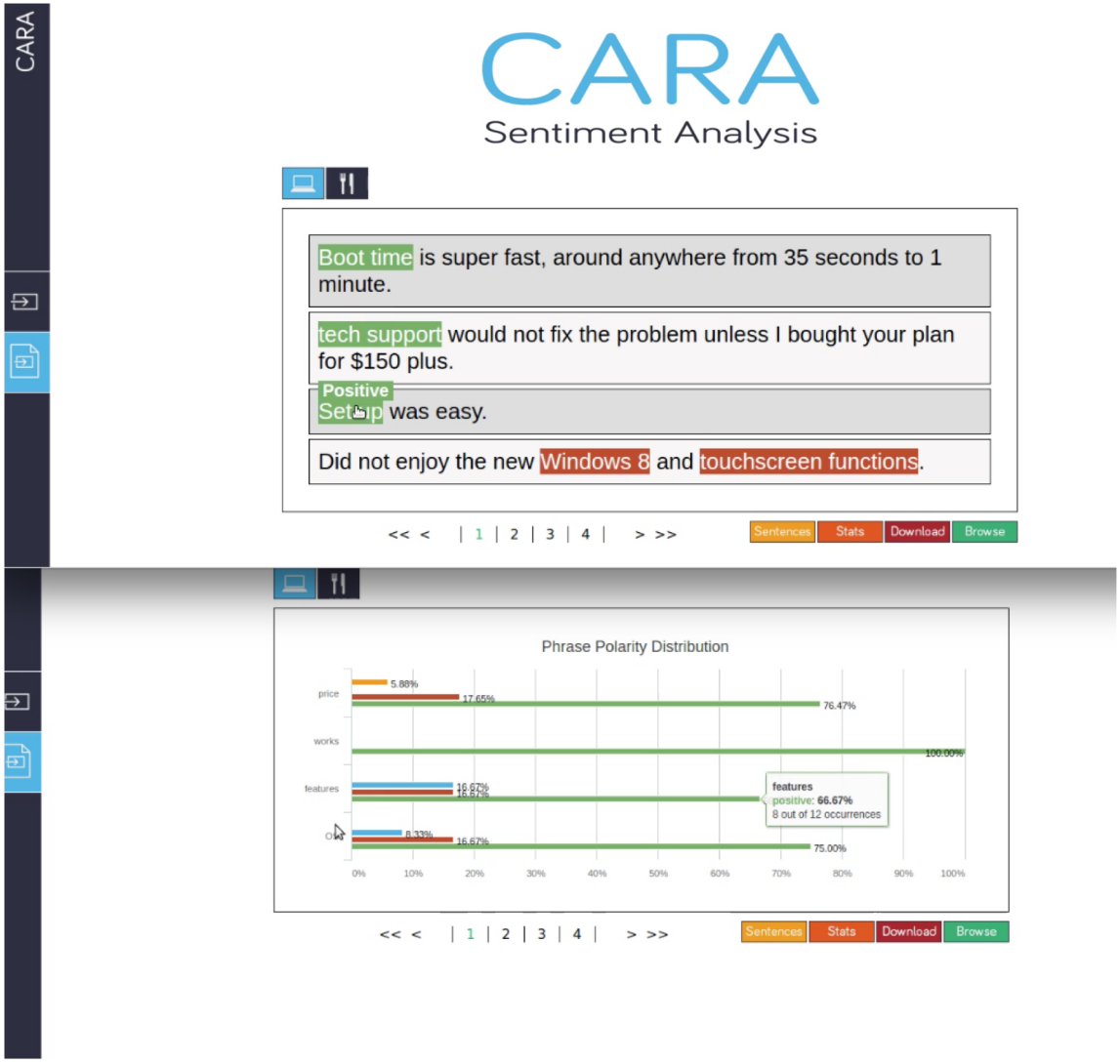
Size and weight

✓ nice and compact to carry!

✓ since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!

✗ the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Sentiment Analysis (Aspect-based)



UGC Machine Translation - Braziliator

The 2014 FIFA World Cup was the biggest event yet for Twitter with **672 million tweets**

Requested translation from Twitter (words)				Grand Total from all World Cup matches
	6,459,830	5,141,360	4,847,590	**85,047,110**

Top 3 languages

English | Portuguese | Spanish

- Source→Target traffic:
- EN→ES 13,614,450 (EN to all languages: 50,545,460)
- ES→EN 5,569,200 (ES to all languages: 10,609,420)
- PT→EN 1,831,750 (PT to all languages: 4,230,880)

UGC Machine Translation - Braziliator

The 2014 FIFA World Cup was the biggest event yet for Twitter with **672 million tweets**

Requested translation from Twitter (words)				Grand Total from all World Cup matches
	6,459			**85,047,110**

85,047,110 total words = 2,835,000 words per day (30 days)

equivalent to 1,134 human translators working full-time for 30 days

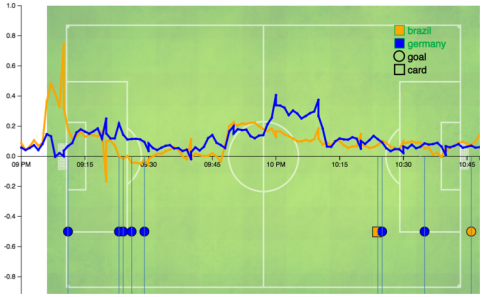
Top 3 languages

ish

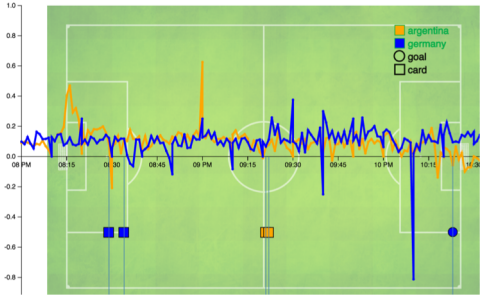
- Source→Targe
- EN→ES 13, ages: 50,545,460)
- ES→EN 5, all languages: 10,609,420)
- PT→EN 1,8 (PT to all languages: 4,230,880)

UGC Machine Translation - Braziliator

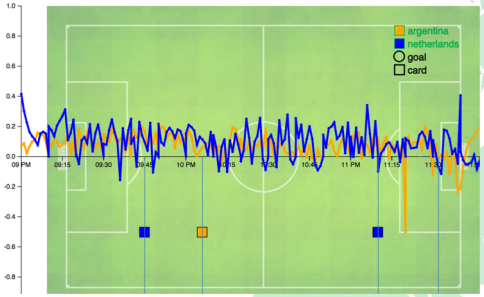
UI: Sentiment pitch



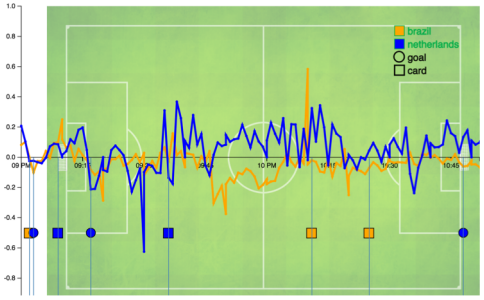
Semi-final: Germany 7-1 Brazil



Final: Germany 1-0 Argentina

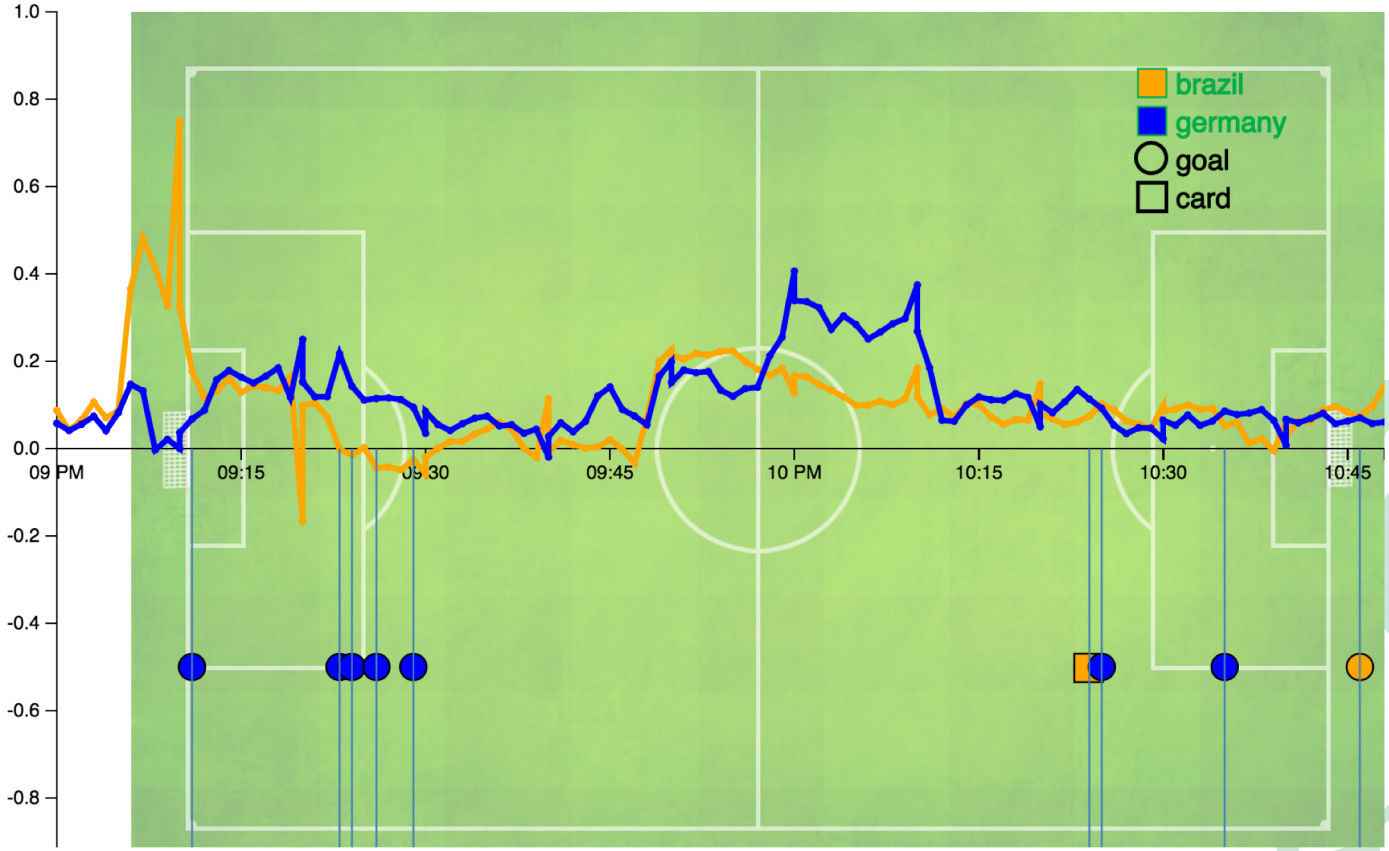


Semi-final: Argentina 1-0 Netherlands



3rd Place: Netherlands 3-0 Brazil

UGC Machine Translation - Braziliator



Semi-final: Germany 7-1 Brazil

Applications

- First, what makes an application a *language processing application* (as opposed to any other piece of software)?
  - An application that requires the use of knowledge about human languages
    - Example: Is Unix wc (word count) an example of a language processing application?

Applications

- Word count?
  - When it counts words: Yes
    - To count words you need to know what a word is. That's knowledge of language.
  - When it counts lines and bytes: No
    - Lines and bytes are computer artifacts, not linguistic entities

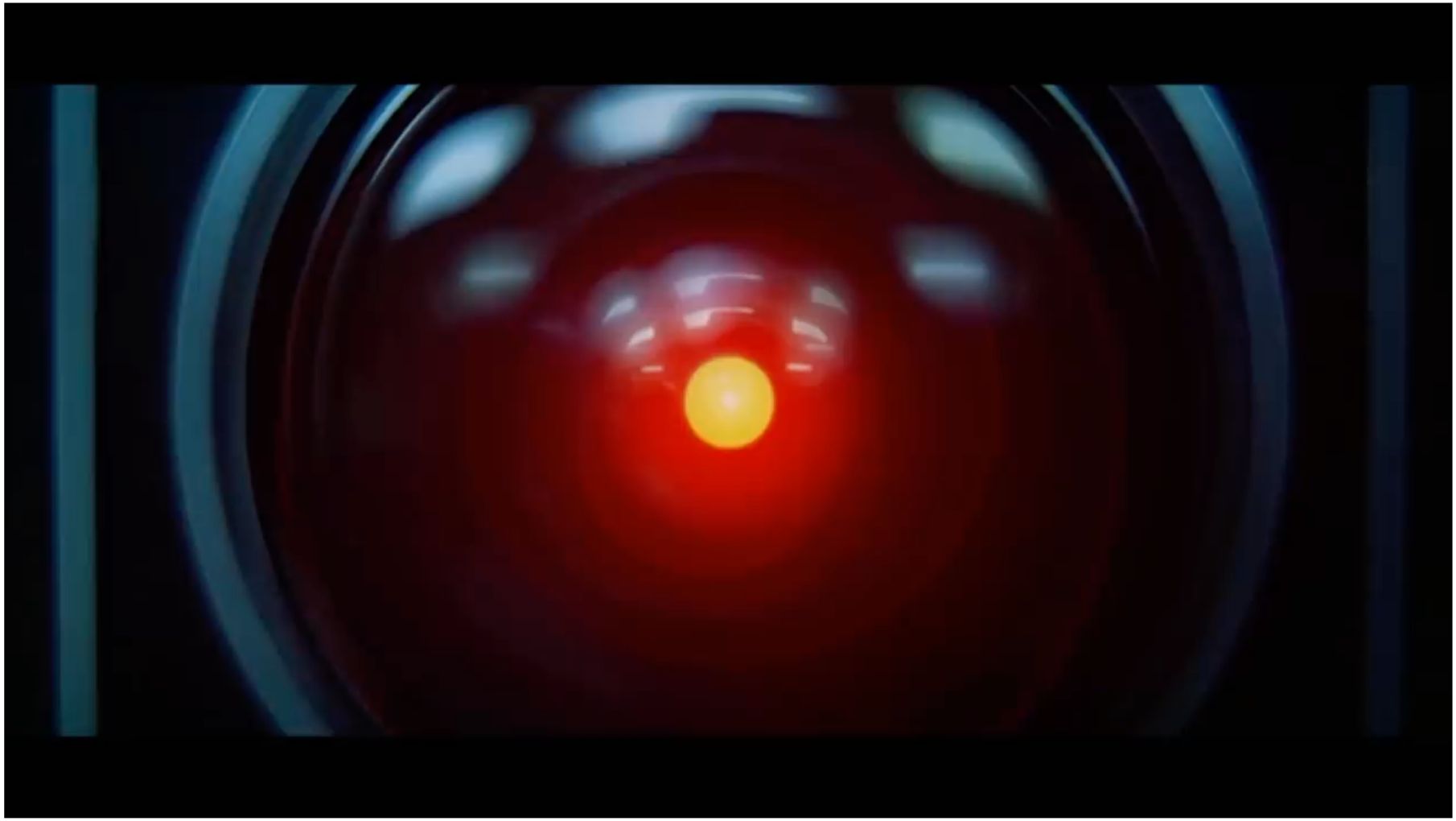
Big Applications

- Question answering
- Conversational agents
- Summarization
- Machine translation

Big Applications

- These kinds of applications require a tremendous amount of knowledge of language.
- Consider the interaction with HAL the computer from **2001: A Space Odyssey**
  - Dave: *Open the pod bay doors, Hal.*
  - HAL: *I'm sorry Dave, I'm afraid I can't do that.*

HAL 9000



https://www.youtube.com/watch?v=ARJ8cAGm6JE

What's needed?

- Speech recognition and synthesis
- Knowledge of the English words involved
  - What they mean
- How groups of words clump
  - What the clumps mean

What's needed?

- Dialog
  - It is polite to respond, even if you're planning to kill someone.
  - It is polite to pretend to want to be cooperative (I'm afraid, I can't…)

Caveat

NLP has an AI aspect to it.

- We're often dealing with ill-defined problems
- We don't often come up with exact solutions/algorithms
- We can't let either of those facts get in the way of making progress

Course Material

- We'll be intermingling discussions of:
  - Linguistic topics
    - E.g. Morphology, syntax, discourse structure
  - Formal systems
    - E.g. Regular languages, context-free grammars
  - Applications
    - E.g. Machine translation, conversational systems

Topics: Linguistics

- Word-level processing
- Syntactic processing
- Lexical and compositional semantics
- Discourse processing
- Dialogue structure

Topics: Applications

- Small
  - Spelling correction
  - Hyphenation
- Medium
  - Word-sense disambiguation
  - Named entity recognition
  - Information retrieval
- Large
  - Question answering
  - Conversational agents
  - Machine translation

- Stand-alone
- Enabling applications
- Funding/Business plans

Categories of Knowledge

- Phonology
- Morphology
- Syntax
- Semantics
- Discourse

- Each kind of knowledge has associated with it an encapsulated set of processes that make use of it.
- Interfaces are defined that allow the various levels to communicate.
- This usually leads to a pipeline architecture.

Ambiguity

- Computational linguists are obsessed with ambiguity
- Ambiguity is a fundamental problem of computational linguistics
- Resolving ambiguity is a crucial goal

Ambiguity

- Find at least 5 meanings of this sentence:
  - I made her duck

Ambiguity

- Find at least 5 meanings of this sentence:
  - I made her duck
- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity is Pervasive

- I caused her to quickly lower her head or body
  - **Lexical category**: “duck” can be a N or V
- I cooked waterfowl belonging to her.
  - **Lexical category:** “her” can be a possessive (“of her”) or dative (“for her”) pronoun
- I made the (plaster) duck statue she owns
  - **Lexical Semantics:** “make” can mean “create” or “cook”

Ambiguity is Pervasive

- **Grammar**: Make can be:
  - **Transitive: (verb has a noun direct object)**
    - I cooked [waterfowl belonging to her]
  - **Ditransitive: (verb has 2 noun objects)**
    - I made [her] (into) [undifferentiated waterfowl]
  - **Action-transitive (verb has a direct object and another verb)**
  - I caused [her] [to move her body]

Ambiguity is Pervasive

- **Phonetics!**
  - I mate or duck
  - I'm eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I'm aid her duck
  - I mate her duck
  - I'm ate her duck
  - I'm ate or duck
  - I mate or duck

Dealing with Ambiguity

Four possible approaches

1. Tightly coupled interaction among processing levels; knowledge from other levels can help decide at ambiguous levels.
2. Pipeline processing that ignores ambiguity as it occurs and hopes that other levels can eliminate incorrect structures.
3. Probabilistic approaches based on making the most likely choices.
4. Don't do anything, maybe it won't matter.
    1. *We'll leave when the duck is ready to eat.*
    2. *The duck is ready to eat now.*
        - Does the "duck" ambiguity matter with respect to whether we can leave?

Models and Algorithms

- By **models** we mean the formalisms that are used to capture the various kinds of **linguistic knowledge** we need.
- **Algorithms** are then used to manipulate the **knowledge representations** needed to tackle the task at hand.

Models

- State machines
- Rule-based approaches
- Logical formalisms
- **Probabilistic models**

Algorithms

- Many of the algorithms that we'll study will turn out to be **transducers**; algorithms that take one kind of structure as input and output another.
- Unfortunately, ambiguity makes this process difficult. This leads us to employ algorithms that are designed to handle ambiguity of various kinds

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤️

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤️

segmentation issues

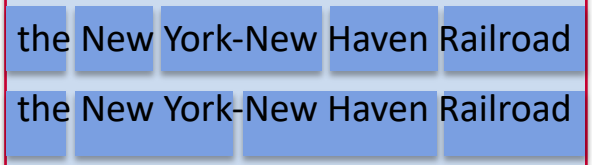
the	New	York-New	Haven	Railroad
the	New York-New	Haven	Railroad	

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤️

segmentation issues



idioms

dark horse
get cold feet
lose face
throw in the towel

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤️

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤️

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤️

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …
Let It Be sold millions…
… a mutation on the *for* gene …

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤️

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …
Let It Be sold millions…
… a mutation on the *for* gene …

But that's what makes it fun!

Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- How we generally do this:
  - Probabilistic models built from language data
    - P(“maison” → “house”) high
    - P(“L’avocat général” → “the general avocado”) low
  - Luckily, rough text features can often do half the job.

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

Dialog

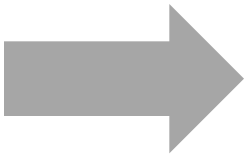
Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Real Success: IBM's Watson

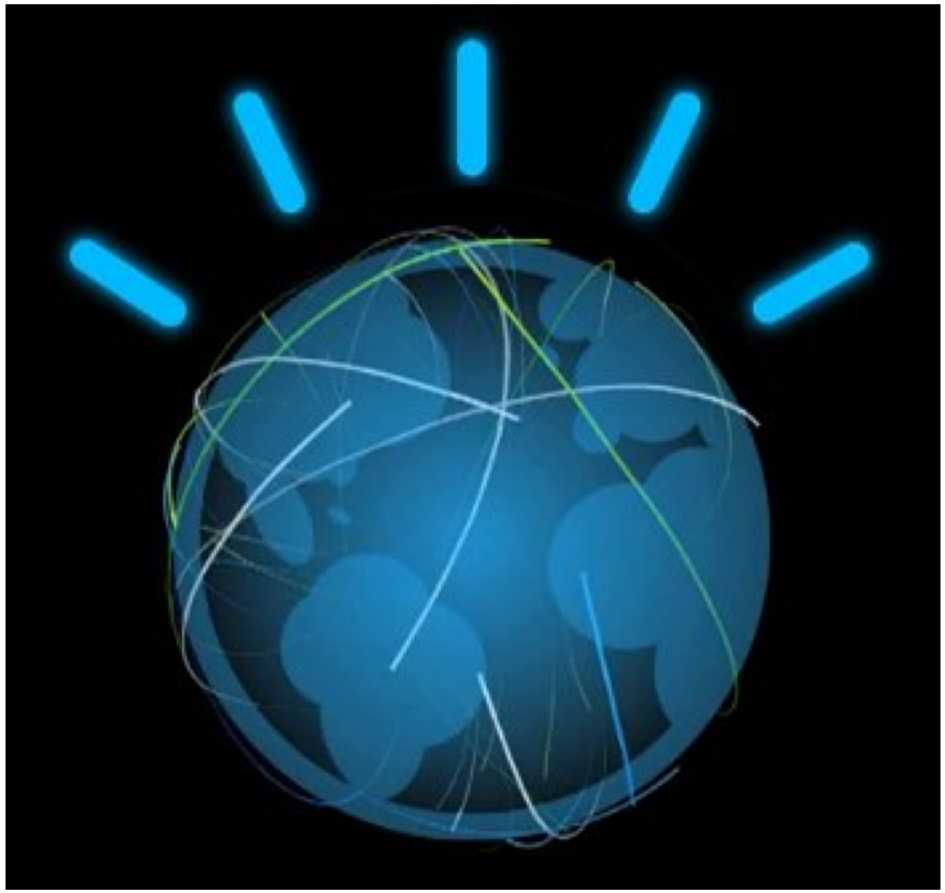
- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL



Bram Stoker

Real Success: Watson on Jeopardy



- https://www.youtube.com/watch?v=WFR3lOm_xhE

Next Week

- Basic text processing
  - Unix tools
  - Regular expressions

Discussion

Some content was adapted from Speech and Language Processing - Jurafsky and Martin

CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computer Science

http://www.cit.ie

Thank you

Haithem. afli@cit.ie

@AfliHaithem