



Aalto University

Statistical Natural Language Processing: an introduction *+contents of the 2021 course*

Presented by: Mikko Kurimo

(Adapted content from Timo Honkela, Mathias Creutz etc. - thanks!)

Part I: Statistical natural language processing

1. Introduction

2. Applications

3. Why is it so hard?

- Challenges of natural language data

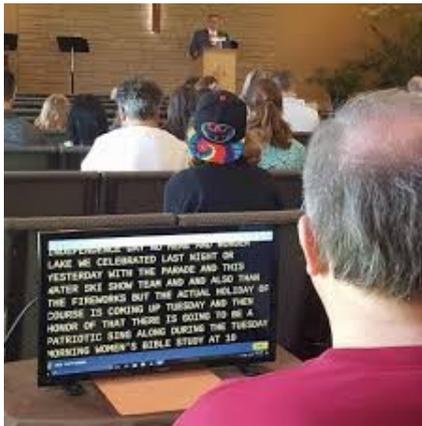
Language is processed in our phones and homes

Including televisions, phones, new assistance devices, toys



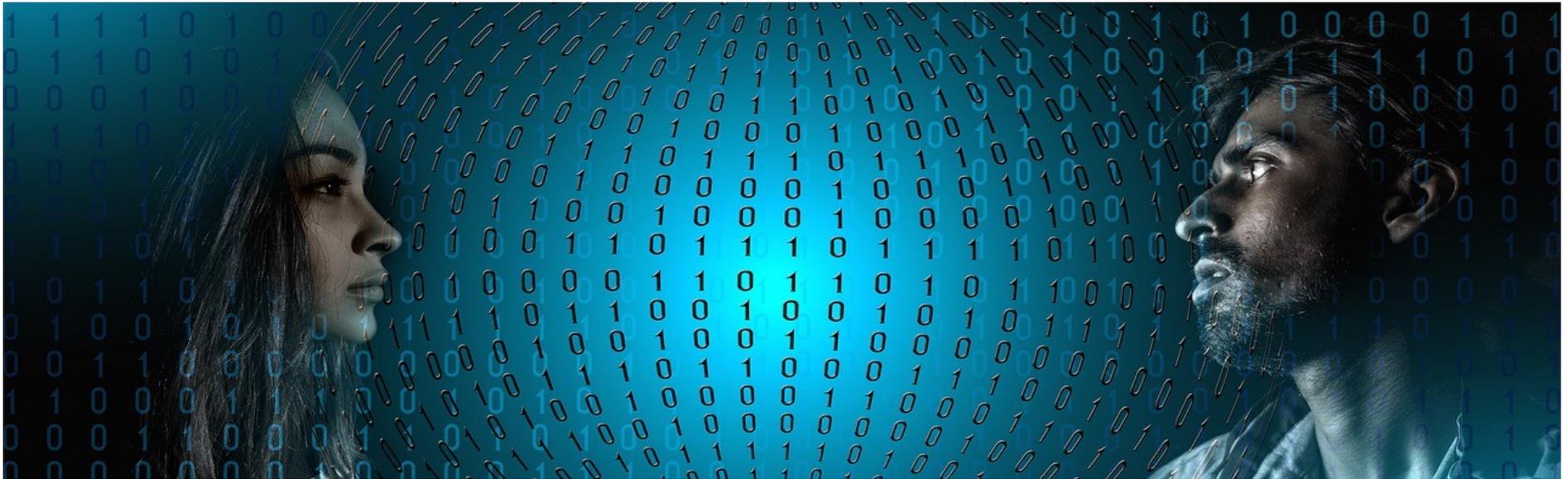
Language is used for several everyday tasks

Including dictation, captioning, translation, interpretation, information retrieval, conversational assistants, language learning



Language is human communication

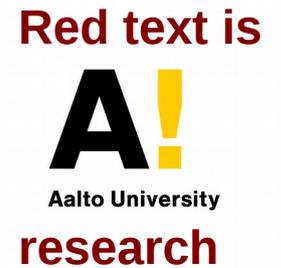
- Rich communication signal **between humans**
- Human speech is the most complex of all biosignals
- speech => text + emotion, loudness, speed, emphasis, ...
- text + *emotion, loudness, speed, emphasis, ...* => speech
- How much language “understanding” is needed?
- People perceive the use of language as a sign of “intelligence”



Modeling of language

- Language is complex, adaptive system
 - Storing and processing text and speech
 - Large datasets
- We want to make systems that 'understand'
 - Take into account language related phenomena
- Building models about natural language using large data sets

Statistical Natural Language Processing



Methodological basis:

- machine learning
- pattern recognition
- probability theory
- statistics
- signal processing

Related fields:

- computational linguistics
- corpus linguistics
- Phonetics
- speech processing
- discourse analysis
- cognitive science
- artificial intelligence

What is in a language?

Phonetics and phonology:

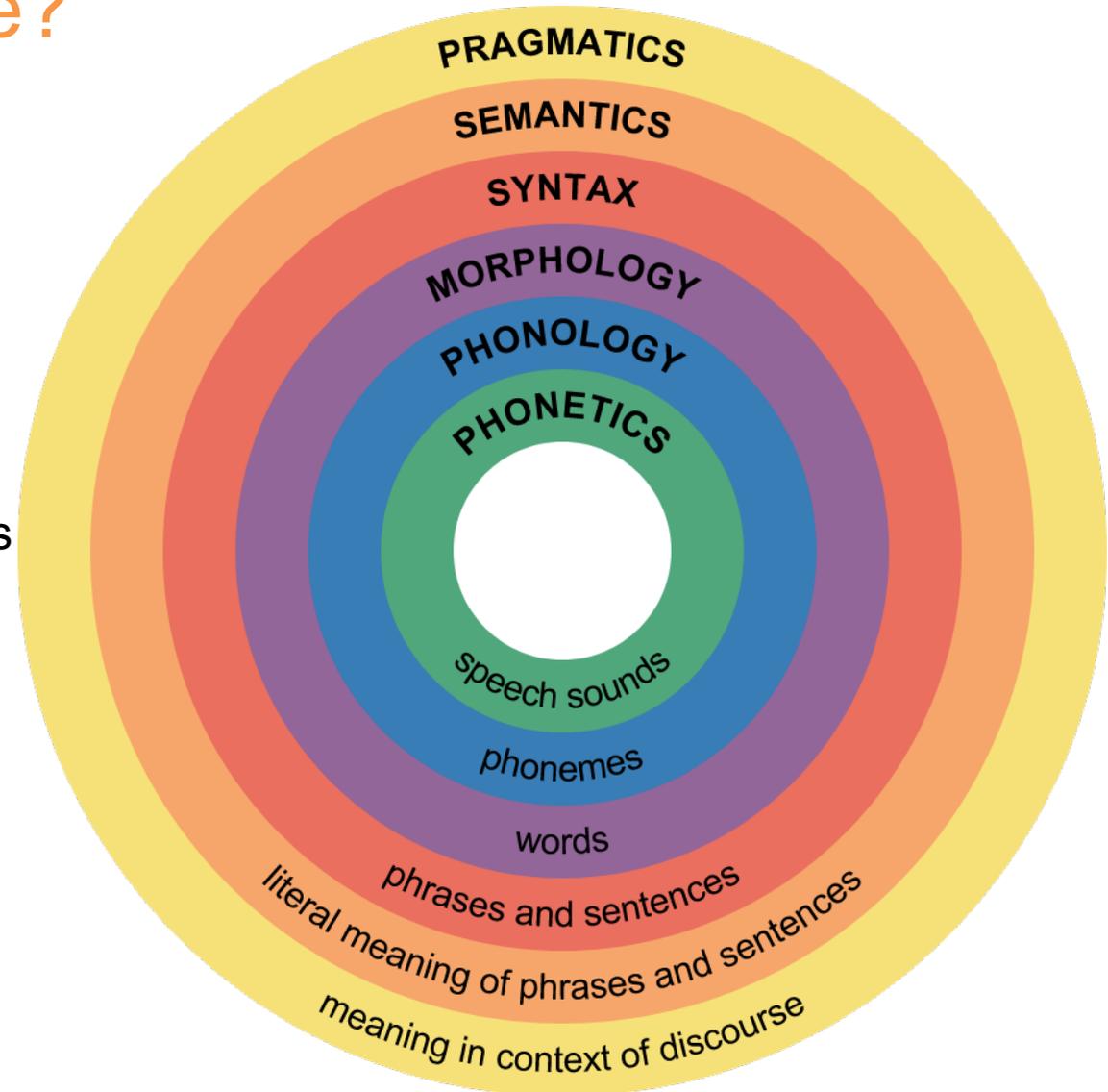
- physical sounds
- patterns of sounds

Morphology: building blocks of words

Syntax: grammatical structure

Semantics: meaning of words

Pragmatics, discourse, spoken interaction...



Application areas



- Information retrieval
- Text clustering and classification
- Automatic speech recognition
- Natural language interfaces
- Statistical machine translation
- ...

Information retrieval



teemu selänten

About 437,000 results (0.37 seconds)

 **Everything**

 Images

 Videos

 News

 Shopping

 Realtime

 More

[Teemu Selänne - Wikipedia, the free encyclopedia](#) ☆

Teemu Ilmari **Selänne** nicknamed "The Finnish Flash" (born July 3, 1970) is a Finnish professional ice hockey player and an alternate captain of the Anaheim Ducks. [Playing career](#) - [International](#) - [Personal](#) - [Career statistics](#)
en.wikipedia.org/wiki/Teemu_Selänne - [Cached](#) - [Similar](#)

[Teemu Selänne – Wikipedia](#) ☆ - [[Translate this page](#)]

Teemu Ilmari **Selänne** (s. 3. heinäkuuta 1970 Helsinki) on suomalainen jääkiekkoilija. [Peliura](#) - [Perhe ja vapaa-aika](#) - [Muuta](#) - [Tilastot](#)
fi.wikipedia.org/wiki/Teemu_Selänne - [Cached](#) - [Similar](#)

[+](#) [Show more results from wikipedia.org](#)

[Teemu Selanne hockey statistics & profile at hockeydb.com](#) ☆

3 Jul 1970 ... A profile of **Teemu Selanne**, a hockey player from Helsinki, Finland, born July 03, 1970.
www.hockeydb.com/ihdb/stats/pdisplay.php?pid=4863 - [Cached](#) - [Similar](#)

PageRank algorithm

The PageRank Citation Ranking: Bringing Order to the Web (1998)

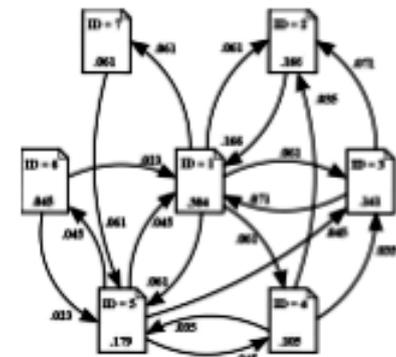
Larry Page, Sergey Brin, R. Motwani, T. Winograd

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them. We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

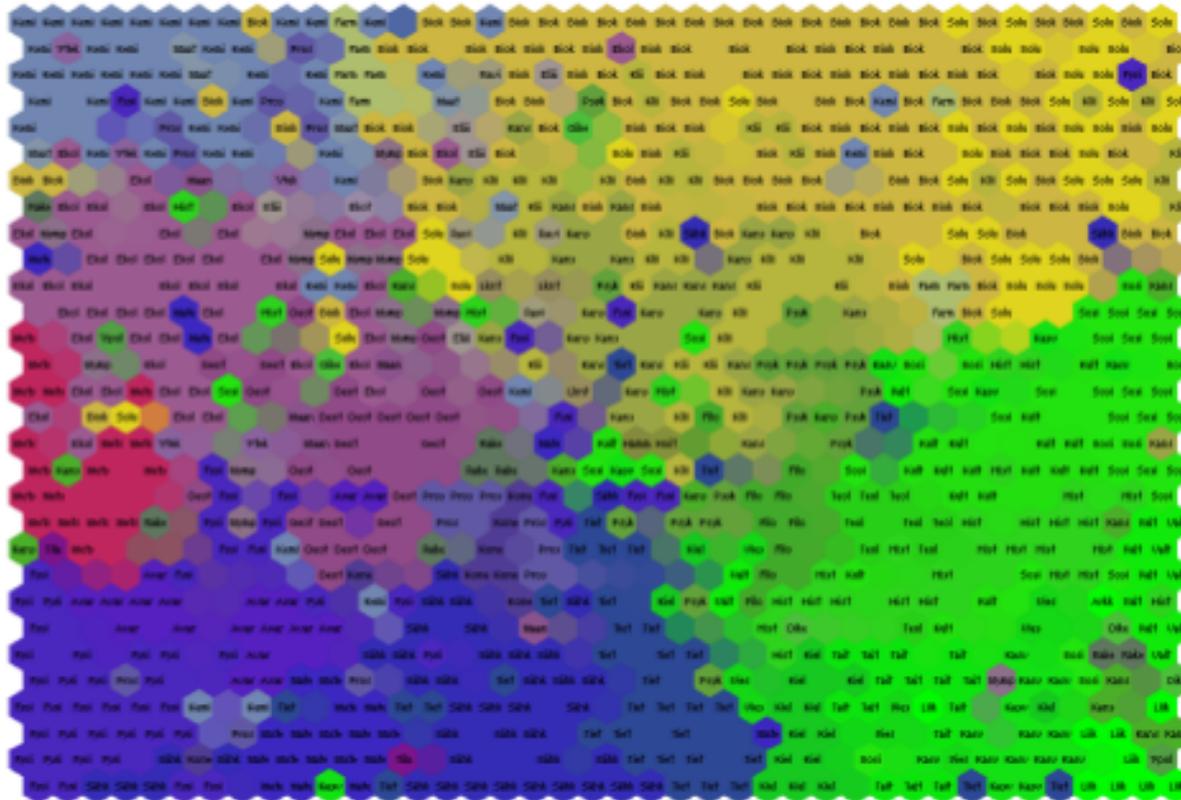


wongsableng.com



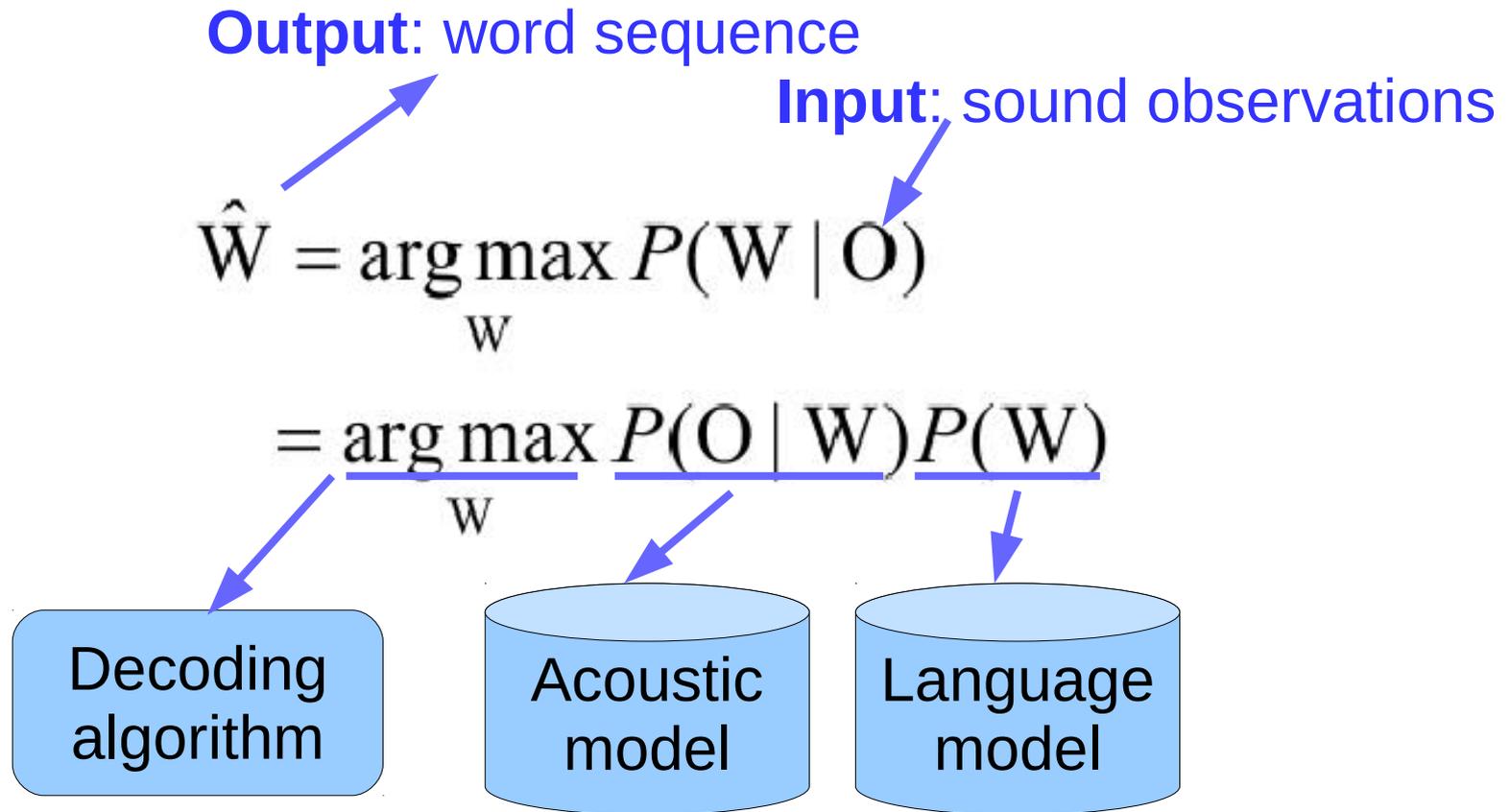
www.vicconsult.com

Text clustering and classification



WEBSOM (Honkela, Kaski, Kohonen & Lagus, 1996, etc.)

Speech recognition: large probabilistic models

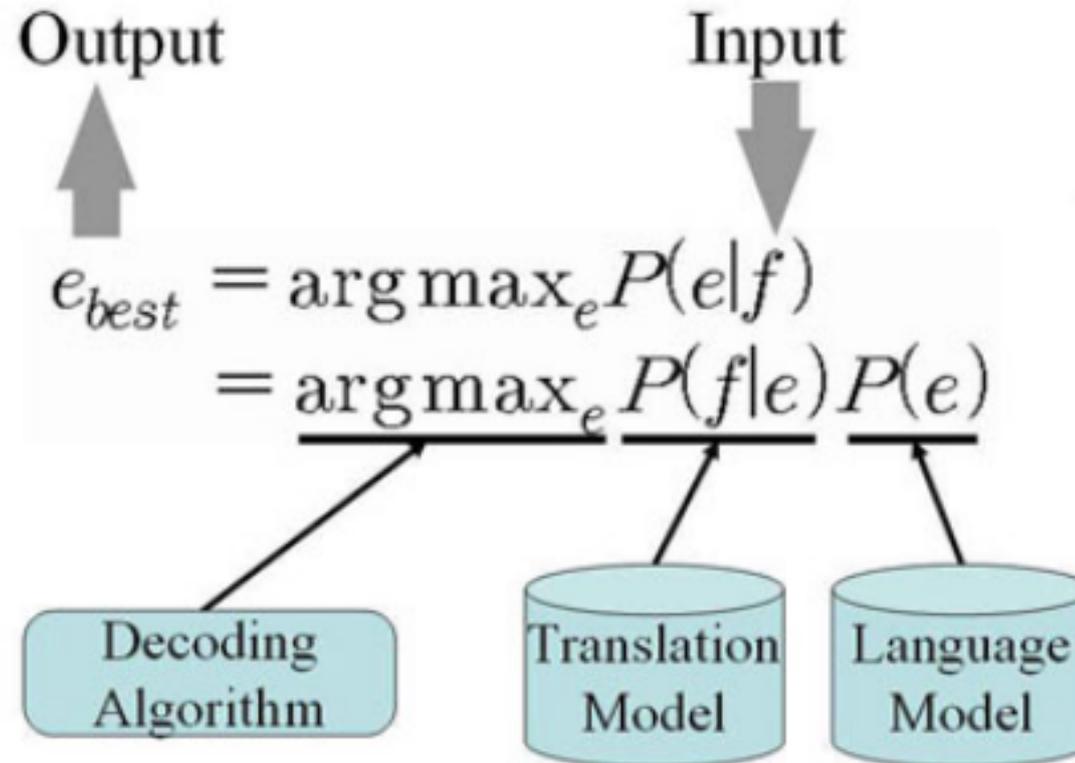


Machine translation

Google translate:

- Jos ei osaa suomen kieltä, on vaikea arvata sanojen merkityksiä.
- If you do not speak Finnish, it is difficult to guess the meanings of words.
- Wenn Sie nicht sprechen Finnisch, ist es schwierig, die Bedeutung der Worte erraten.
- Если вы не говорите по фински, трудно угадать смысл слов.
- 如果你不会讲芬兰语，很难猜测词的含义。

Machine translation: large probabilistic models

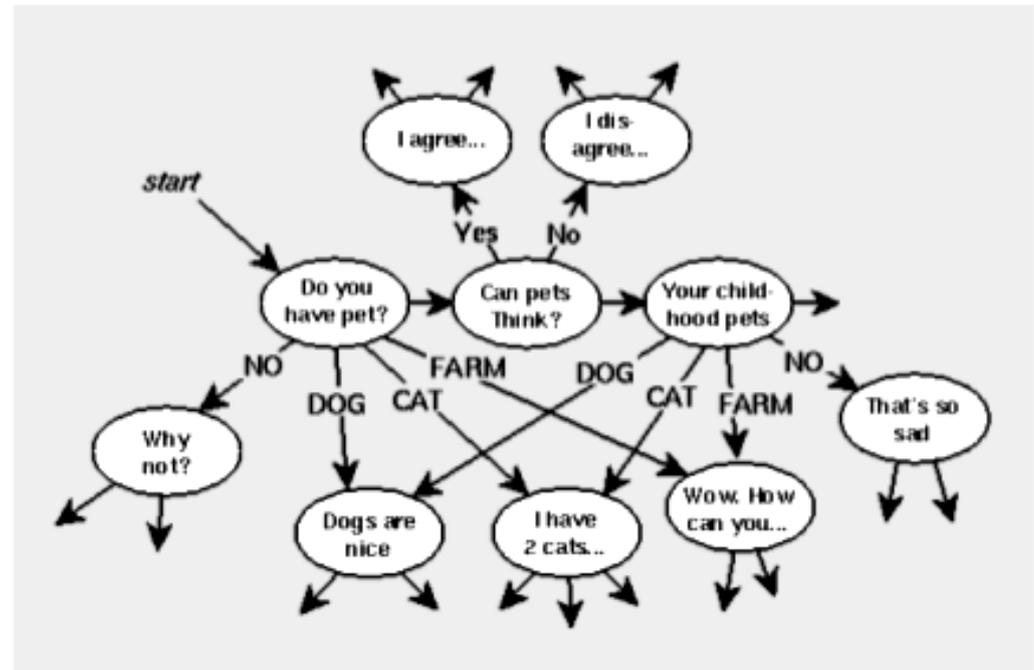


isoft.postech.ac.kr

Natural language interfaces

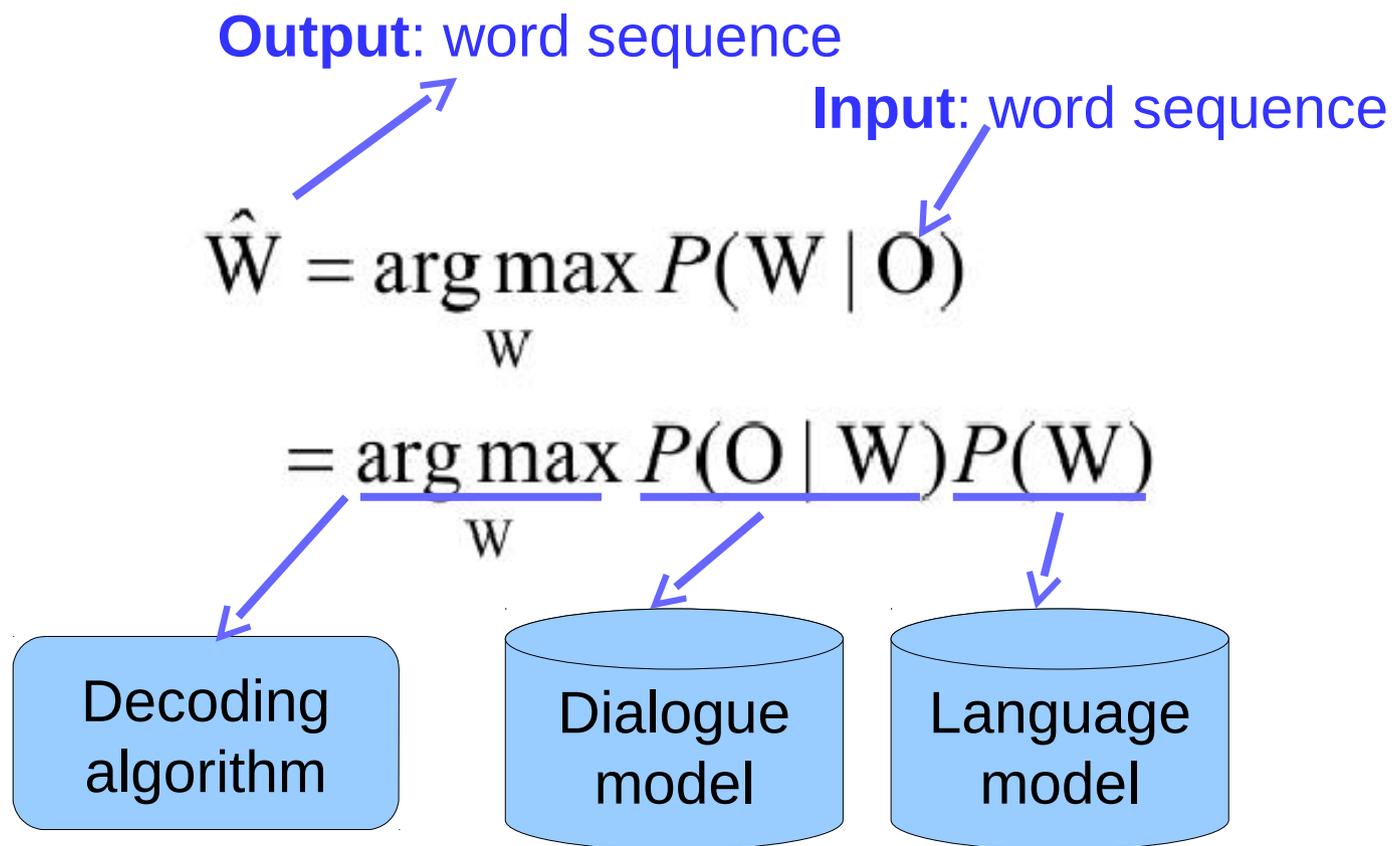


www.zabaware.com



robot-club.com

Dialogue generation: a large probabilistic models point of view



Discussion

Discuss 10 mins in breakout groups and make some notes:

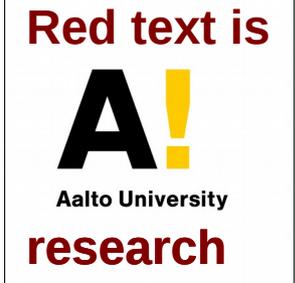
1. Introduce yourself to the group
2. What kind of Natural Language Processing applications have you used?
3. What is working well? What does not work?
4. What kind of future applications would be useful in your daily life?

To receive an activity point, submit your notes (photo, text or pdf):

- In MyCourses => Lectures => Lecture 1 exercise return box
- Hint: You can directly write notes in the text box while you discuss

More about how to earn activity points and how they affect course grading will be discussed at the end of this lecture.

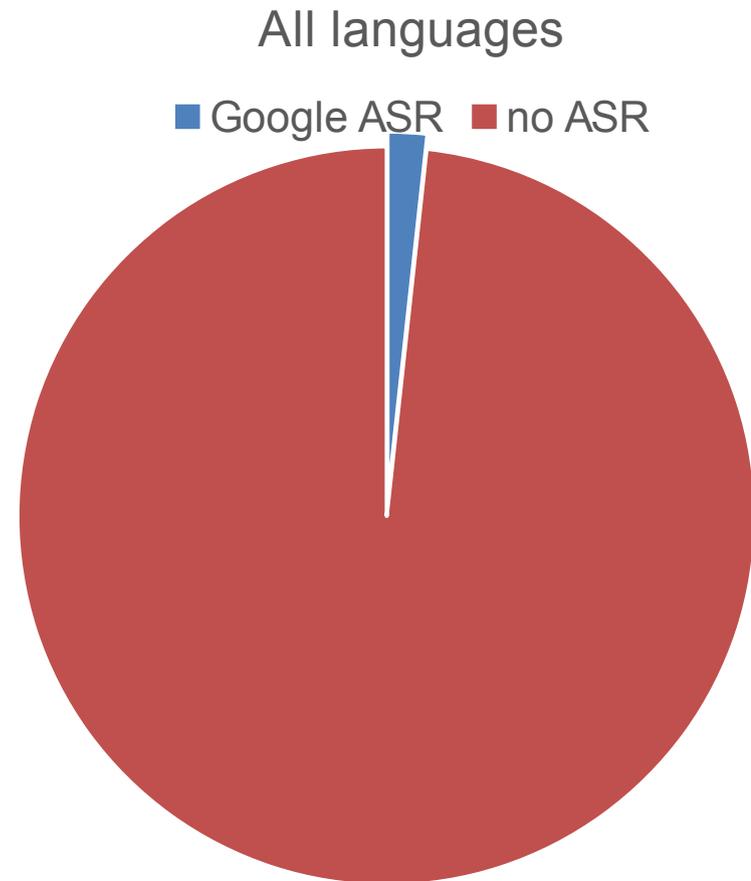
More application areas



- Information retrieval
- Text clustering and classification
- Automatic speech recognition
- Natural language interfaces
- Statistical machine translation
- Topic detection
- Sentiment analysis
- Word sense disambiguation
- Syntactic parsing
- **Text generation**
- **Image, audio and video description**
- **Text-to-speech synthesis**
- ...

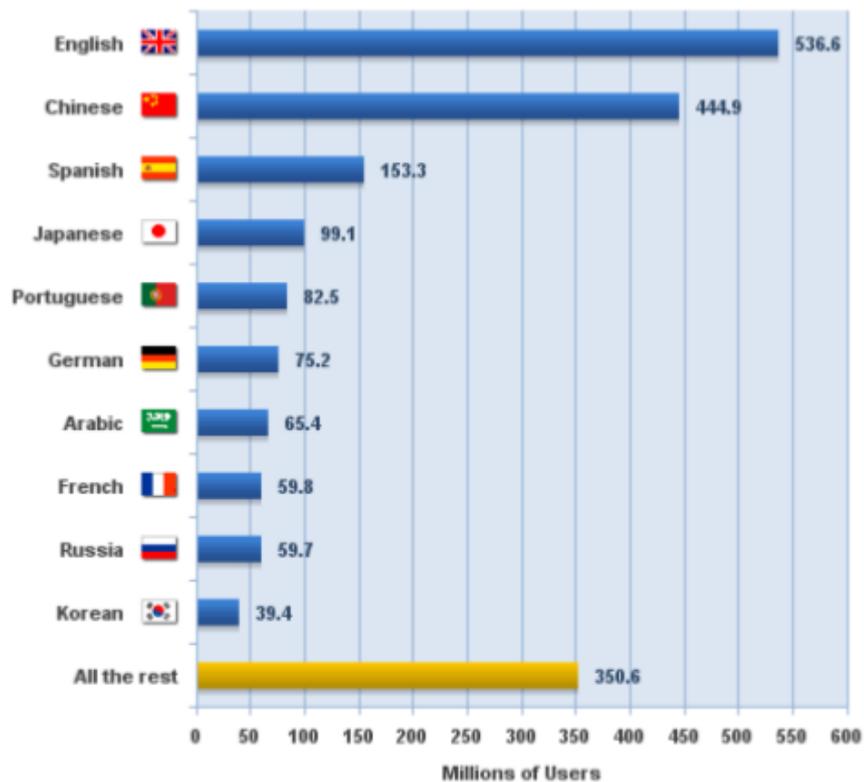
Complexity of natural languages

- 6000+ languages, many dialects
- Each has many words
- Each word is understood slightly differently by each speaker
- Large variety of sentence structures

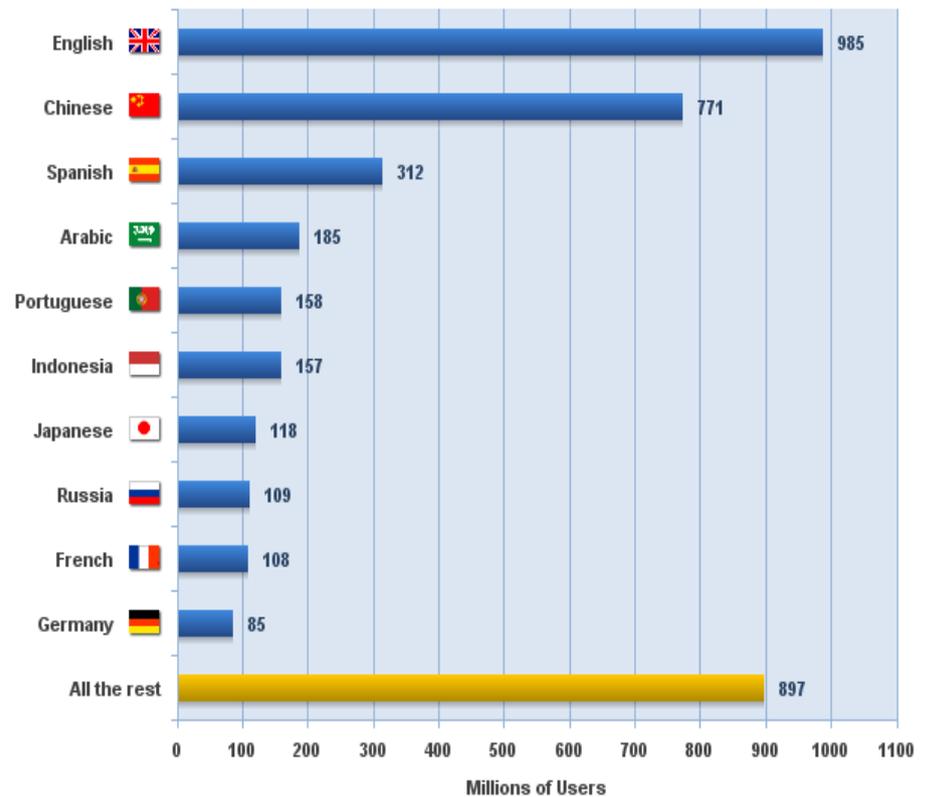


Languages in the internet

Top Ten Languages in the Internet
2010 - in millions of users



Top Ten Languages in the Internet
in Millions of users - June 2017



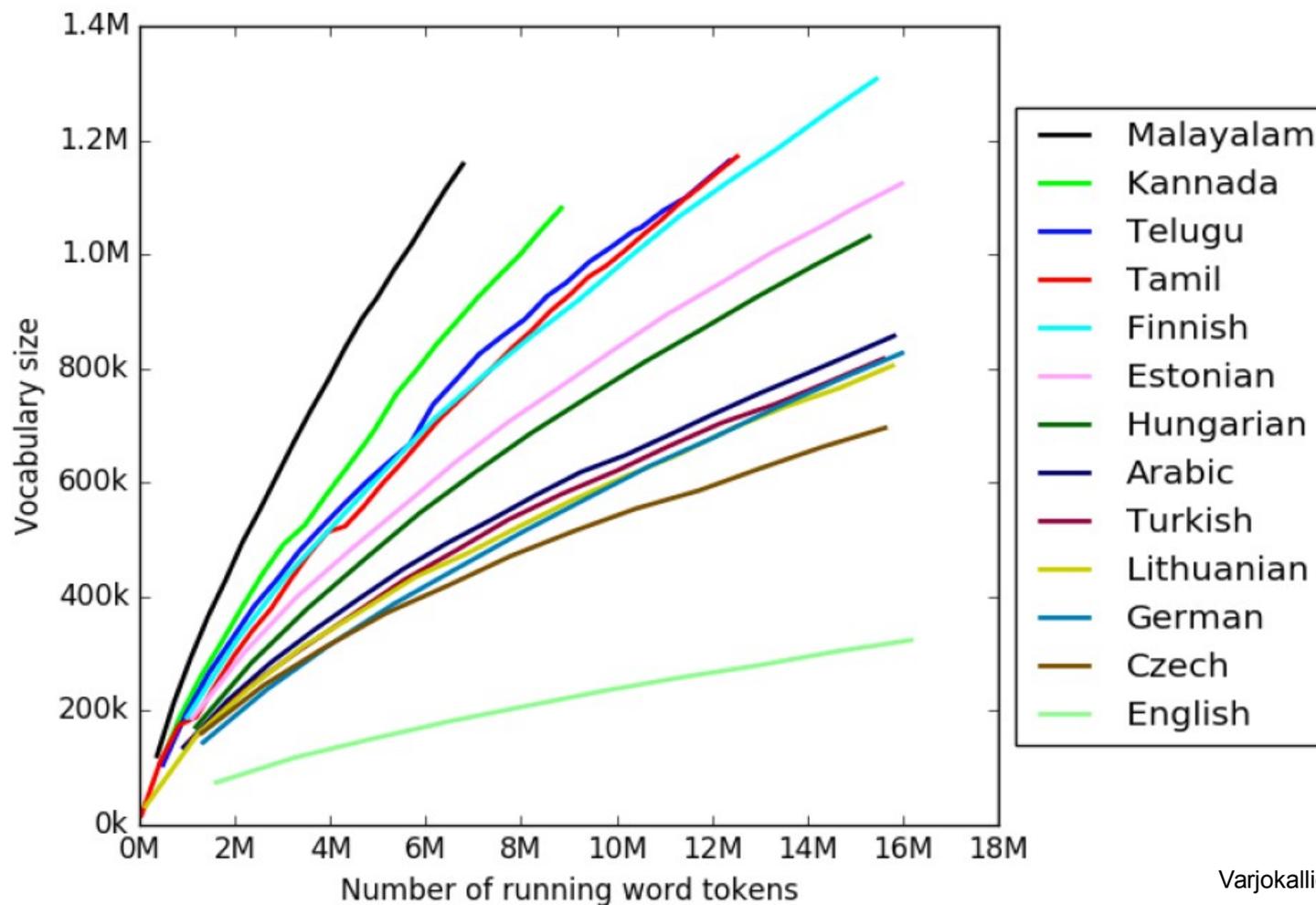
www.internetworldstats.com

EU languages

Table 2. An example phrase from each of the EU languages.

cs	Smlouva o ústavě pro evropu
da	Traktat om en forfatning for europa
de	Vertrag über eine Verfassung für Europa
el	Συνθήκη για τη θέσπιση Συνταγμάτος Ευρώπης
en	Treaty establishing a Constitution for Europe
es	Tratado por el que se establece una constitución para Europa
et	Euroopa põhiseaduse leping
fi	Sopimus euroopan perustuslaista
fr	Traité établissant une Constitution pour l'Europe
ga	Conradh ag bunú Bunreachta don eoraip
hu	Szerződés európai alkotmány létrehozásáról
it	Trattato che adotta una Costituzione per l'Europa
lt	Sutartis dėl Konstitucijos Europai
lv	Līgums par konstitūciju eiropai
mt	Trattat Li Jistabbilixxi kostituzzjoni għall-Ewropa
nl	Verdrag tot vaststelling van een grondwet voor europa
pl	Traktat ustanawiają Konstytucję dla europy
pt	Tratado que estabelece uma Constituição para a Europa
sl	Zmluva o ústave pre Európu
sk	Pogodba o ustavi za evropo
sv	Fördrag om upprättande av en konstitution för europa

Effect of morphology: vocabulary size as function of corpus size



Varjokallio, Kurimo, Virpioja (2016)

Challenges of segmentation

- Modeling morphology -- segmenting words
 - istua "to sit", istuutua "to sit down",
 - Istun "I sit", istahdan "I sit down for a while"
 - istahtaisin "I would sit down for a while"
 - istahtaisinko? "should I sit down for a while?"
 - istahtaisinkohan? "I wonder if I should sit down for a while?"
- Where are the word boundaries?

Hello World

周公吐哺

Challenge of modeling syntax

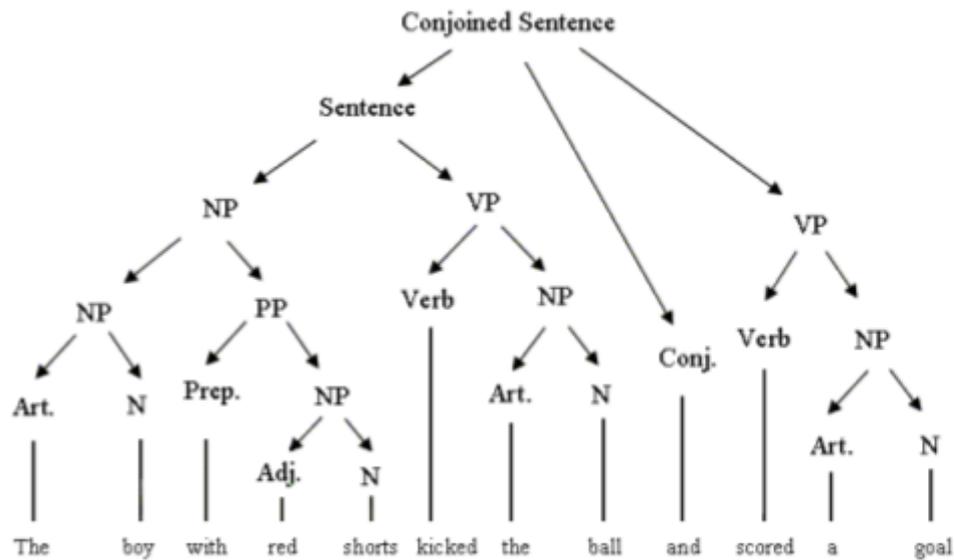


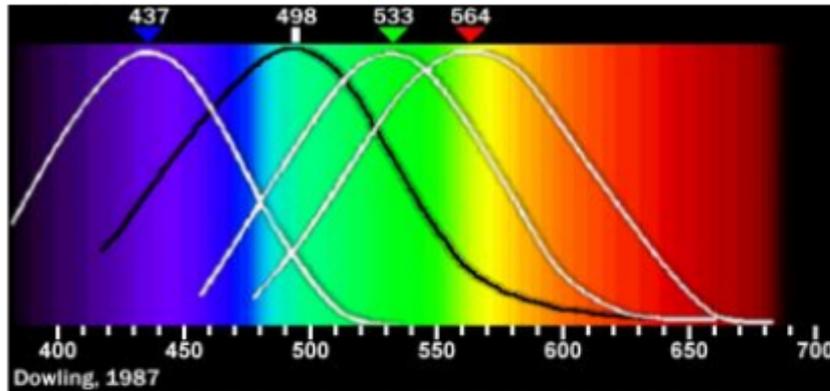
Figure 1.1.3.

“White House”
versus
“white house”

Challenges of natural language

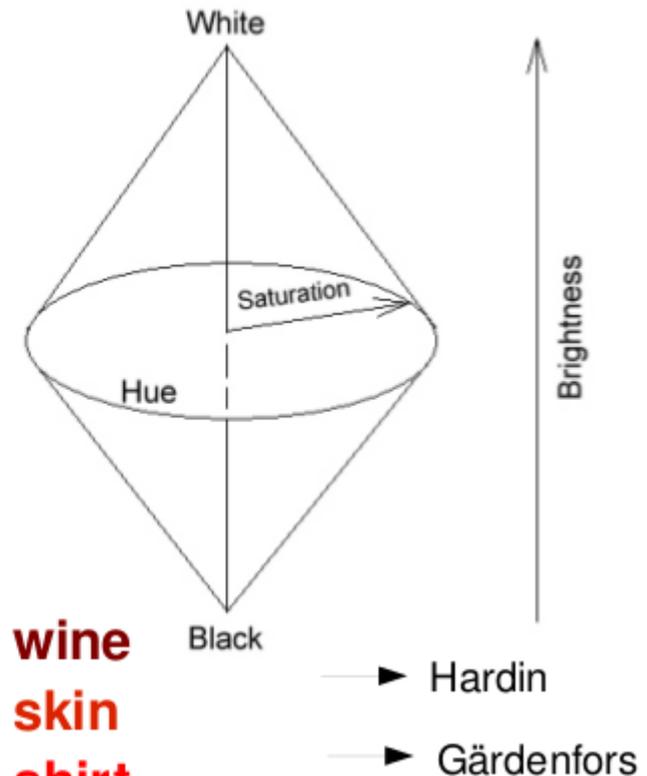
- Understanding the meaning of words is subjective:
 - learning language through individual life paths
 - end up having different ways of understanding and producing language
- Many words have several meanings:
 - E.g. “play”, “game”, “window”
- Sentences have several interpretations:
 - E.g. “Big children and adults saw a man with a telescope”

Example: color naming

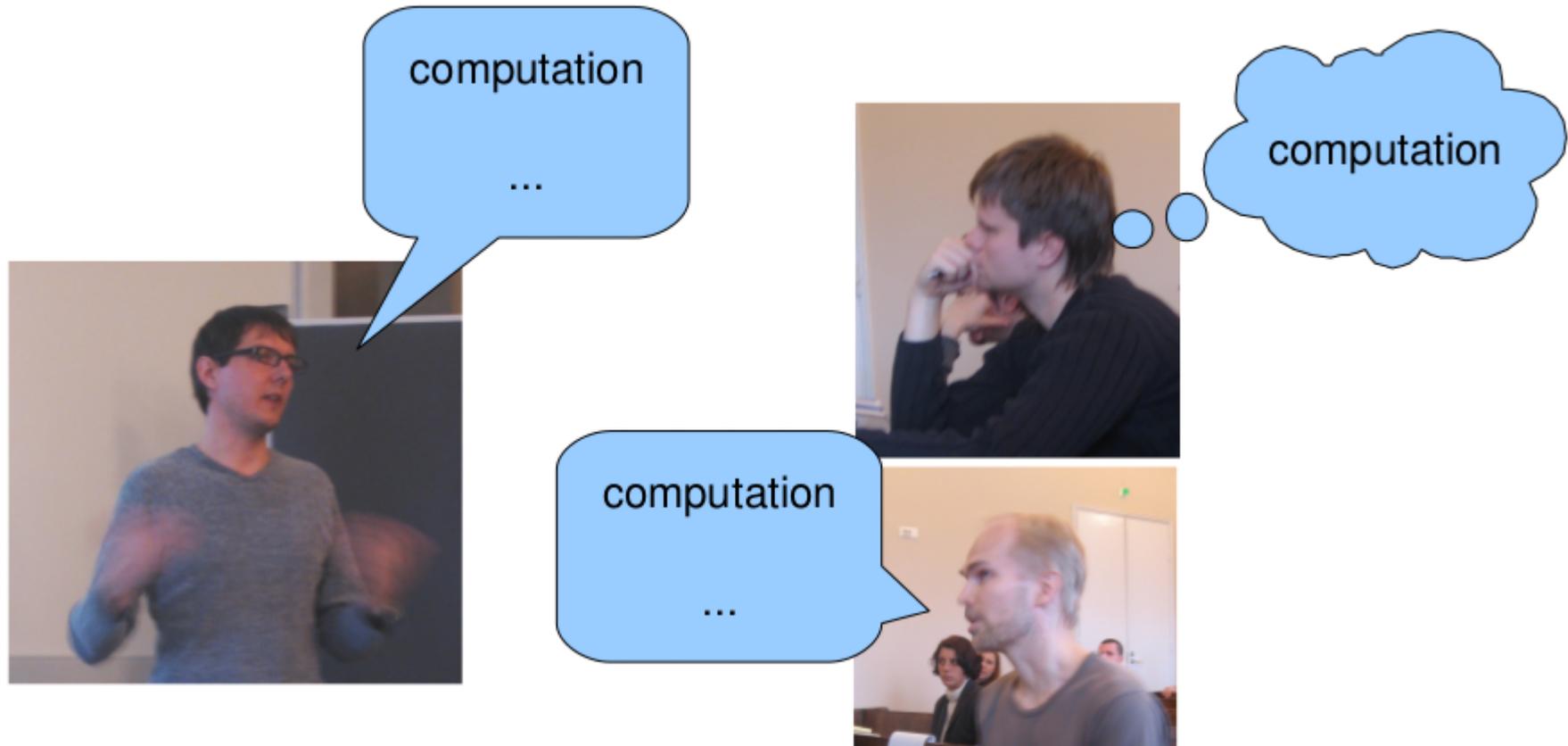


Human vision: rods, cones,...
Physical reasons for color
Contextuality of naming

red wine
red skin
red shirt



Complex concepts: e.g. concept of computation



Different cultural contexts



?

Shakespeare's sonnet:

“Shall I compare thee to a summer's day?”

?

Challenge of encoding world knowledge

- For good performance, world knowledge is needed
 - Quantitatively this is challenging
 - Qualitatively there are also many problems (mapping between language and the world is complex, cf. examples above)
- Note: world is essentially dynamic, continuous and multimodal, symbolic systems are not

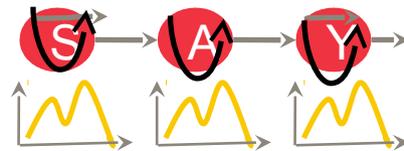
Corpus-based methods

- Corpora are large collections of text
 - Annotated: add knowledge about words or structure into corpus
 - Or just plain text
- Statistical information on
 - Distribution of words and parts of words
 - Structure
 - Word similarity
- Allow us to build models and **test** hypotheses
- Allow us to explore
- Choose the best models based on statistics

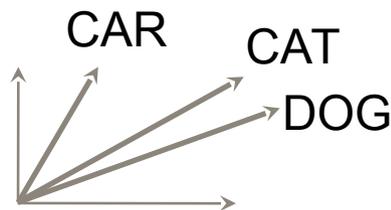
Natural language processing

METHODS

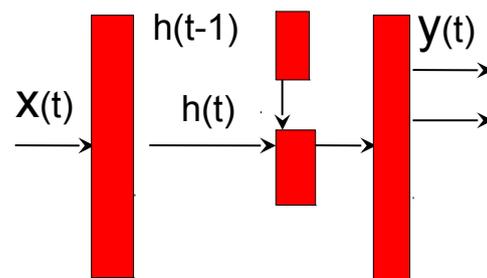
Hidden Markov model



Vector space model



Recurrent neural network



TOOLS

- Speech-to-text
- Text-to-speech
- Machine translation
- Information retrieval
- Named entity recognition
- Sentence parsing
- Topic detection

Natural language modeling: basic tasks

Red text is



Aalto University

research

Word level

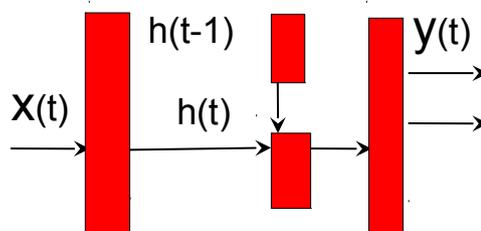
1. Vector space models
2. Text preprocessing
3. Bag of words models
4. Modeling morphology

Sentence level

1. Part-of-speech tagging
2. Named entity recognition
3. Statistical language models
4. Neural language models

A recent revolution in the language modeling approach

- Split language into tokens
 - Vector space modeling, embedding
 - Representation learning
 - Deep & recurrent learning
 - Sequence to sequence mapping
- => artificial intelligence



Read more

- Manning & Schütze: Foundations of Statistical Natural language processing
 - Chapter 1: Introduction
 - Chapter 2: Probability and Information Theory basics

Part II: Course details

- 1.Goals
- 2.Materials and tools
- 3.Lectures
- 4.Exercises
- 5.Course project
- 6.Grading
- 7.Submission DLs

1. Goals

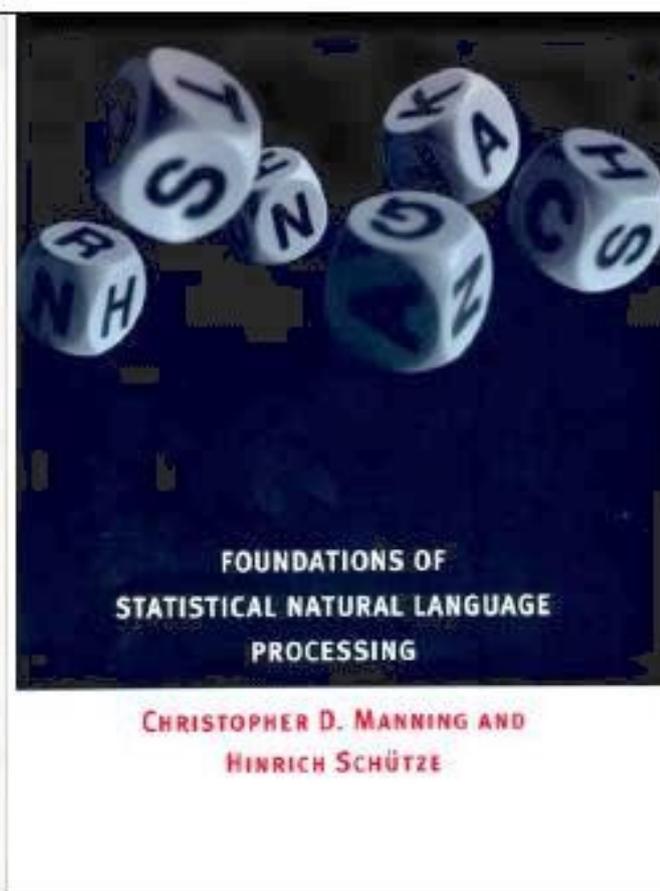
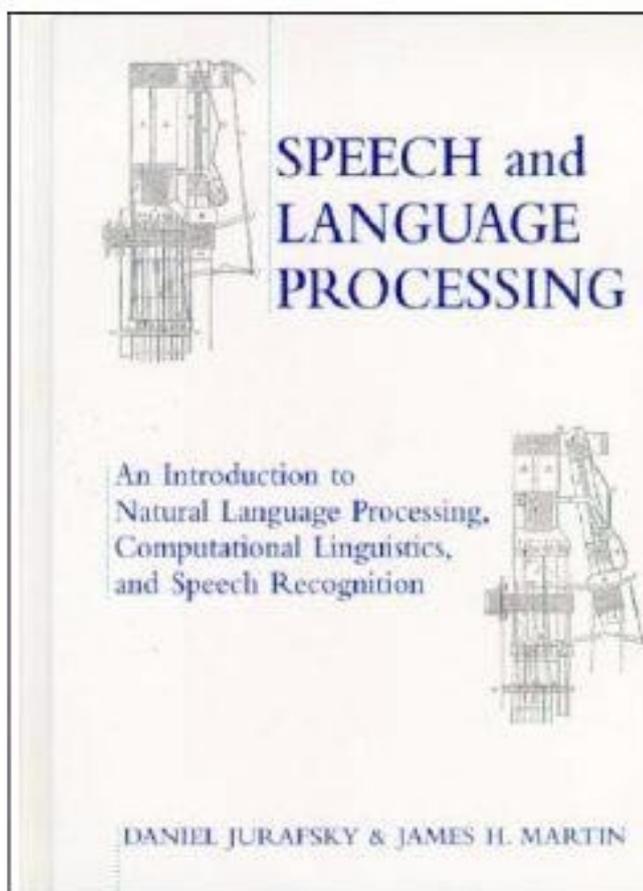
- To learn **how statistical and adaptive methods** are used in information retrieval, machine translation, text mining, speech processing and related areas **to process natural language data**
- To learn how to apply the basic methods and techniques for clustering, classification, generation and recognition **by natural language modeling**

Course personnel

- Responsible professor & lecturer: *Mikko Kurimo*
- Assistant & exercises: *Ekaterina (Katja) Voskoboinik, Aku Rouhe, Dejan Porjazovski*
- Project work: *all the above + research group members*
- Visiting lecturers: *Mathias Creutz, Janne Pylkkönen, Tiina Lindh-Knuutila, Stig-Arne Grönroos, Jaakko Väyrynen, Mittul Singh*

2. Materials: Text books

1. **C. Manning, H. Schütze.** Foundations of Statistical Natural Language Processing. MIT Press, 1999. <http://nlp.stanford.edu/fsnlp/>
2. **D. Jurafsky, J. H. Martin.** Speech and Language Processing (3rd ed. Draft, 2020) <http://web.stanford.edu/~jurafsky/slp3/>



2. Other materials and tools

- **MyCourses:** Lecture slides and recordings, assignments and info for project work
- **Slack:** For questions and discussions before and after the lectures and exercise sessions, project groups can make their own public or private channels
- **Internet:** Plenty of books, articles, demos and course material from all over the world

3. Lectures

- **12 lectures: January 12 – March 30**
- Tue 12:15 – 14:00 in Zoom
- Visiting experts who have PhD and industrial experience in their topic. Typically they are the best experts available in Finland.
- Slides and other material provided by lecturers
- All lectures will be recorded and available for the course participants
- Active attendance to the weekly lectures, studying the material and taking the exam corresponds to 2 cr.
- Participation to the lectures is not mandatory, but recommended for reaching the learning outcomes of the course.

Lectures in the course (changes possible)

1. 12 Jan Introduction & Project groups / Mikko Kurimo
2. 19 jan Statistical language models / Mikko Kurimo
3. 26 jan Word2vec / Tiina Lindh-Knuutila
4. 02 feb Sentence level processing / Mikko Kurimo
5. 09 feb Speech recognition / Janne Pylkkönen
6. 16 feb Chatbots and dialogue agents / Mikko Kurimo
7. 23 feb Exam week, no lecture
8. 02 mar Statistical machine translation / Jaakko Väyrynen
9. 09 mar Morpheme-level processing / Mathias Creutz
10. 16 mar Neural language modeling and BERT / Mittul Singh
11. 23 mar Neural machine translation / Stig-Arne Grönroos
12. 30 mar Societal impacts and course conclusion / Krista Lagus, Mikko

See Mycourses
for updates

4. Exercises

- **10 exercise sessions:** Thu 14:15 – 16:00 in Zoom
- Participation to the weekly exercise sessions and submitting the home exercises corresponds to 1 cr.
- The participation to the Thursday sessions is not mandatory, but highly recommended. The assistance for the home exercises is only available during these sessions.
- The DL for submitting the home exercises is before the next Tuesday lecture. Larger exercises can have two weeks.

5. Course project

- Word2vec, BERT or your own NLP project
- Learn to use word embeddings and apply pre-trained models like BERT
- Word2vec and BERT tutorials and materials available online, e.g. in Stanford
- Course project material available at Aalto machines at
 - `/work/courses/unix/T/ELEC/E5550`

BERT example task

- Try pre-trained BERT models for General Language Understanding Evaluation (GLUE) tasks
 - Tasks like sentiment analysis, document classification etc.
 - Various pre-trained models are available
 - Bert-base-uncased
 - Bert-base-cased etc
 - Analyse for your task which model performs best
- Use pre-trained models for your own task
- Experiments with dataset size and sequence length, batch size, learning rate
- Try different BERT-like models on your task

Available BERT models

- First, read the tutorial on what is BERT:
 - <https://blog.usejournal.com/part1-bert-for-advance-nlp-with-transformers-in-pytorch-357579d63512>
- Next, read the tutorial on how to use use BERT:
 - <https://medium.com/@aniruddha.choudhury94/part-2-bert-fine-tuning-tutorial-with-pytorch-for-text-classification-on-the-corpus-of-linguistic-18057ce330e1>
 - This webpage will help you install and run (“finetune + evaluate”) bert on a specific task
- Many BERT-like models are available:
 - Like multilingual BERT or GPT or XLM
 - https://huggingface.co/transformers/v2.2.0/pretrained_models.html

Available BERT datasets

- General language understanding evaluation (GLUE) datasets
 - <https://gluebenchmark.com/tasks>
 - Download using:
<https://gist.github.com/W4ngatang/60c2bdb54d156a41194446737ce03e2e>
- Stanford question answering dataset (SQuaD)
 - <https://huggingface.co/transformers/v2.2.0/examples.html#squad>
- Finnish datasets:
 - <https://github.com/TurkuNLP/FinBERT>
- Your own dataset?

word2vec example task

1. Try existing word2vec embeddings

- Embeddings in various languages available

- Analyze: Are the nearest neighbors semantically meaningful? How? Why not?

2. Build your own model:
English, Finnish, or any other language you are fluent with

3. Experiment with parameters:
Window size, CBOW vs. skipgram, downsampling

4. Evaluate the model with evaluation set(s) of your choice:
Nearest neighbor, noun categorization, Analogical reasoning

5. Evaluate the non-English model with a similar task, translate part of the evaluation set

6. EXTRA: Compare results between languages

Available embeddings

- Available embeddings online in different languages
 - English: Google news (at course project folder)
 - Finnish: lemmatized: Finnish internet (at course project folder)
 - French: <http://fauconnier.github.io/#data>
 - Arabic: <https://github.com/bakrianoo/aravec>
 - German:
<https://devmount.github.io/GermanWordEmbeddings/>
 - MultilingualPolyglot:
<https://sites.google.com/site/rmyeid/projects/polyglot>
 - Let me know what else you have found!

Available corpora for word2vec

- English Wikipedia text corpus (2008, preprocessed)
- Finnish Wikipedia text corpus (2008, automatically lemmatized, preprocessed)
- EXTRA: Find a corpus in a language you are fluent in

Word2vec evaluation possibilities

- At course folder “eval”
- Sanity check: use the word list and check the nearest neighbor for each word: Are they semantically or syntactically meaningful?
- Concrete nouns categorization task data set (ESLLI 2008), Translated also to Finnish
- Concrete vs. abstract noun categorization task (ESLLI 2008), Translated also to Finnish
- Google Analogical reasoning task (choose a subset)
- New Finnish SimLexdataset (<http://www.aclweb.org/anthology/W17-0228>)

word2vec programming

- Minimal programming skills needed
- Word2vec available in Python in Gensim package
 - available on Aalto machines
 - Can be also installed on your own computer
- For usage, see: <https://radimrehurek.com/gensim/index.html>
- original Word2vec package available in C
 - <https://code.google.com/archive/p/word2vec>
- For analysis, use whatever you want: Python, Matlab, R, CLUTO...

Available word2vec code

- There are very good resources online
 - Gensim-package and tutorials
 - Tensorflow
 - Other tutorials
- You can reuse programming examples found online but you must reference your sources
 - If you reuse or modify code, make a note who is the original author, and where you found the code, and list your sources in “References” section of your report.
 - For example, Tensorflow code samples are under Apache 2.0 License
 - Return the program code you've used to run your experiments as an appendix of your report

Gensim at Aalto machines

- Python3 virtual environment installed in course folder with name gensim
- Use it in the virtual environment with command “source /work/courses/unix/T/ELEC/E5550/gensim/bin/activate”
- deactivate the environment with command “deactivate”
- In addition, gensim should be available on Aalto Ubuntu via “module load anaconda”

Course project grading

- See Mycourses for the requirements of an acceptable project report
- Peer grading will be performed for some parts to get more feedback, but that is separate from the final project grade
- Excellent projects typically include additional work such as
 - Exceptional analysis of the data
 - Application of the method to a task or several
 - Algorithm development
 - Own data set(s) (with preprocessing etc to make them usable)

6. Course grading

- 20% of the grade comes from the **exam**. The exam will be organized at the end of the course in April. For those who can not participate in it, there will be a second exam in Autumn. Exams passed in previous years are still valid for completing the course.
- 40% of the grade is from the weekly **home exercises and lecture activities**. The lecture activities may include pen&paper tasks, quizzes, discussions. To get the points return your solutions during the lecture or on the day after, at the latest.
- 40% of the grade is from the **project work**. It depends on experiments, literature study, short (video) presentation and final report. Course projects accepted in previous years are still a valid for completing the course.
- The course includes a mandatory **entrance survey**. The purpose of the test is to filter the students who aim at doing the project work and completing the course. It will also be used to find out the expectations, preferences and background skills of the students by self-evaluation.

7. Submission DLs

- The submission date of the first home exercise is **Monday January 18**
- The submission date of the entrance test is **Tuesday January 19**
- The submission date of each weekly home exercise is **by Monday** in the following week (or 2 weeks for some).
- The project group should select and register their topic **by 4 February**.
- The DL for the final project report is **30 April**.

See Mycourses
for updates

How to achieve the learning goals (and pass the course)?

- *Participate actively in each lecture, read the corresponding material and ask questions* to learn the basics, take part in discussions, complete the lecture exercises
- *Participate actively in each exercise session after each lecture* to learn how to solve the problems, in practice
- Complete the home exercises in time
- *Participate actively in project work* to learn to apply your knowledge
- *Prepare well for the examination*

Questions?

- Responsible professor & lecturer: *Mikko Kurimo*
- Projects & exercises: *Ekaterina Voskoboinik, Aku Rouhe, Dejan Porjazovski*
- Emails: *firstname.lastname@aalto.fi*
- Home page:
<https://mycourses.aalto.fi/course/view.php?id=28539#section-0>