NUI Galway
OÉ Gaillimh

Introduction to NLP

Introduction

Dr. Paul Buitelaar
Data Science Institute, NUI Galway

Learning Outcomes of this Course

Understand basic principles in automatic analysis, interpretation and transformation of textual data

Understand approaches and challenges in syntactic and semantic analysis of textual data

Understand the foundations of NLP in linguistics, statistics and machine learning

Get an insight into NLP applications, in particular information extraction, knowledge graphs and opinion mining

Awareness of ethical, legal and data privacy aspects of NLP

Learning Outcomes of This Lecture

Understand the relevance of NLP applications in society

Get an insight into levels of complexity in human language

Get an insight into state of the art directions in NLP

Understand core aspects of NLP architectures and applications

Get useful pointers to data, research and communities in the field

Overview

Organizational issues

Why NLP

Human language

Some core notions in NLP

Architectures

Applications

Useful pointers

Overview

Organizational issues

Why NLP

Human language

Some core notions in NLP

Architectures

Applications

Useful pointers

Introduction to NLP - Course Overview

01 Introduction
02 Linguistic Concepts
03 Vector Space Model & Word Embeddings
04 Language Modelling
05 Tagging & Hidden Markov Models
06 Probabilistic Parsing
07 Semantic Analysis
08 Information Extraction
09 Knowledge Graphs & Chatbots
10 Opinion Mining, Ethics & Data Privacy
11 Invited Industry Talk
12 Summary and Q&A

The Team

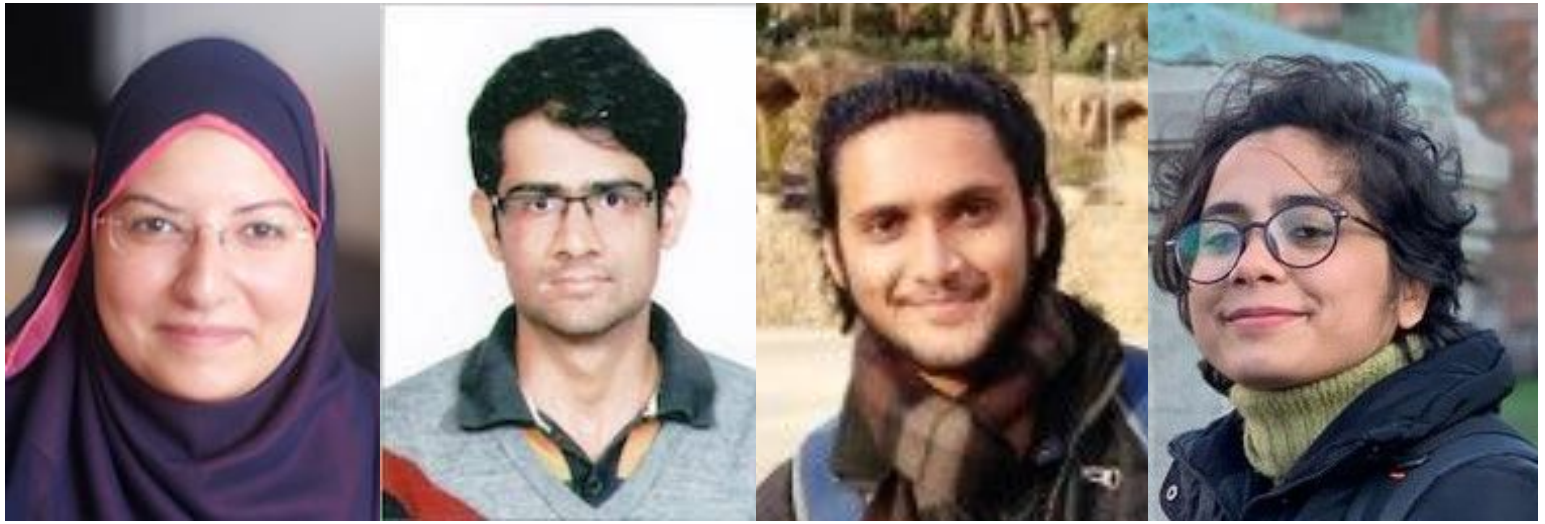
Paul Buitelaar & John McCrae - lectures

Omnia Zayed - labs

Nivranshu Pasricha - labs

Shardul Suryawanshi - labs

Priya Rani - labs



Recommended Reading

Lectures

Daniel Jurafsky, James H. Martin. *Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Pearson Prentice Hall. 3rd edition draft https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Labs

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. O'Reilly. Contents at http://www.nltk.org/book/

Other

Yoav Goldberg. *Neural network methods for natural language processing.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool -- chapters 6,7,8

Chris Manning and Hinrich Schuetze. *Foundations of Statistical Natural Language Processing.* MIT Press. Cambridge. Teaching materials at http://nlp.stanford.edu/fsnlp/

Overview

Organizational issues

Why NLP

Human language

Some core notions in NLP

Architectures

Applications

Useful pointers

What is Natural Language Processing

scientific study and technology development concerned with

algorithms, methods and systems

for the

automatic **analysis, interpretation and transformation** of human language

Why Natural Language Processing

Search Engines

- Process web queries and retrieve relevant information

Machine Translation

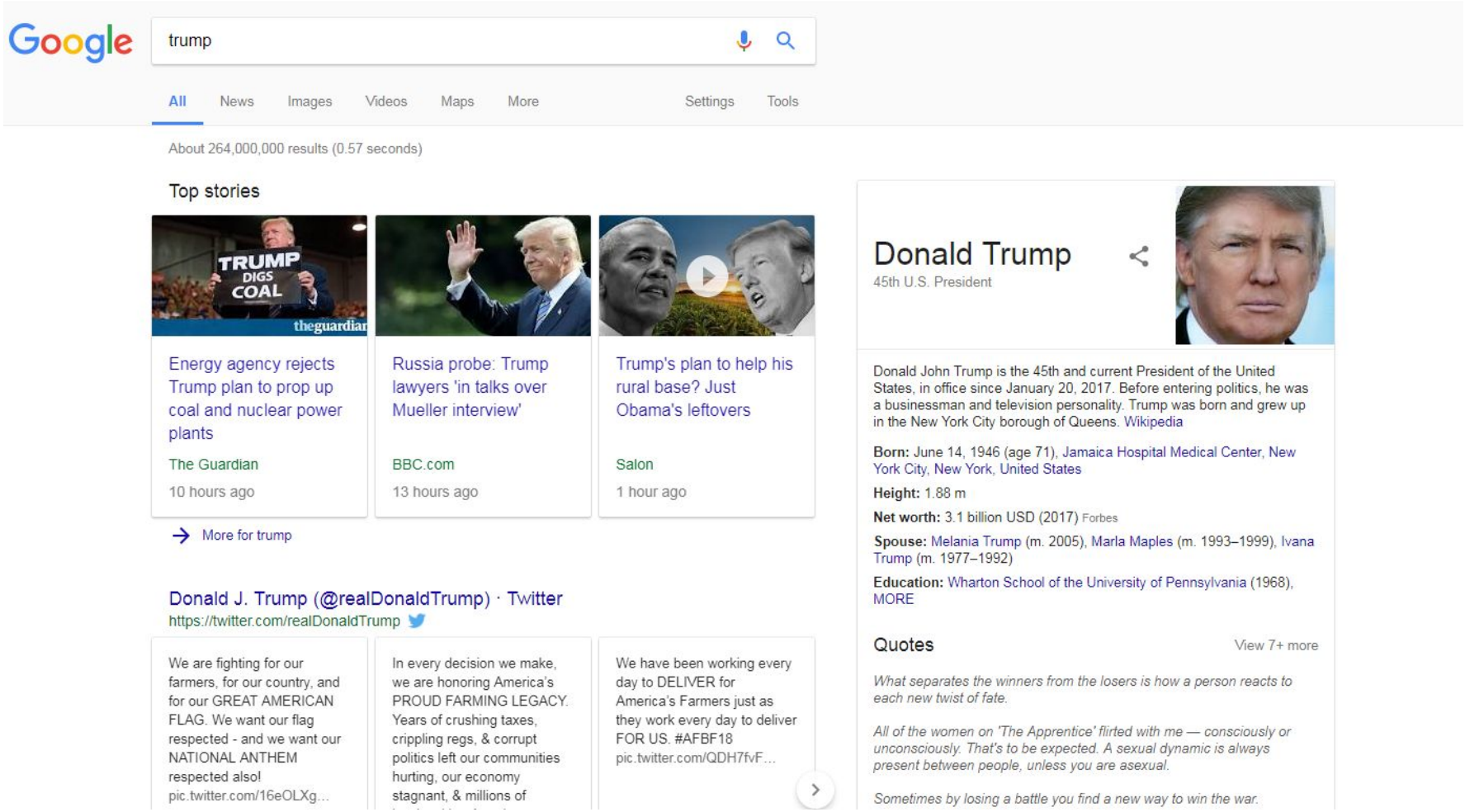
- Process input words or sentences to be translated

Dialog Systems (Chatbots / Digital Assistants)

- Process speech and interpret it into task-specific representations and actions

Other applications: Opinion Mining, Knowledge Graphs

Search Engines



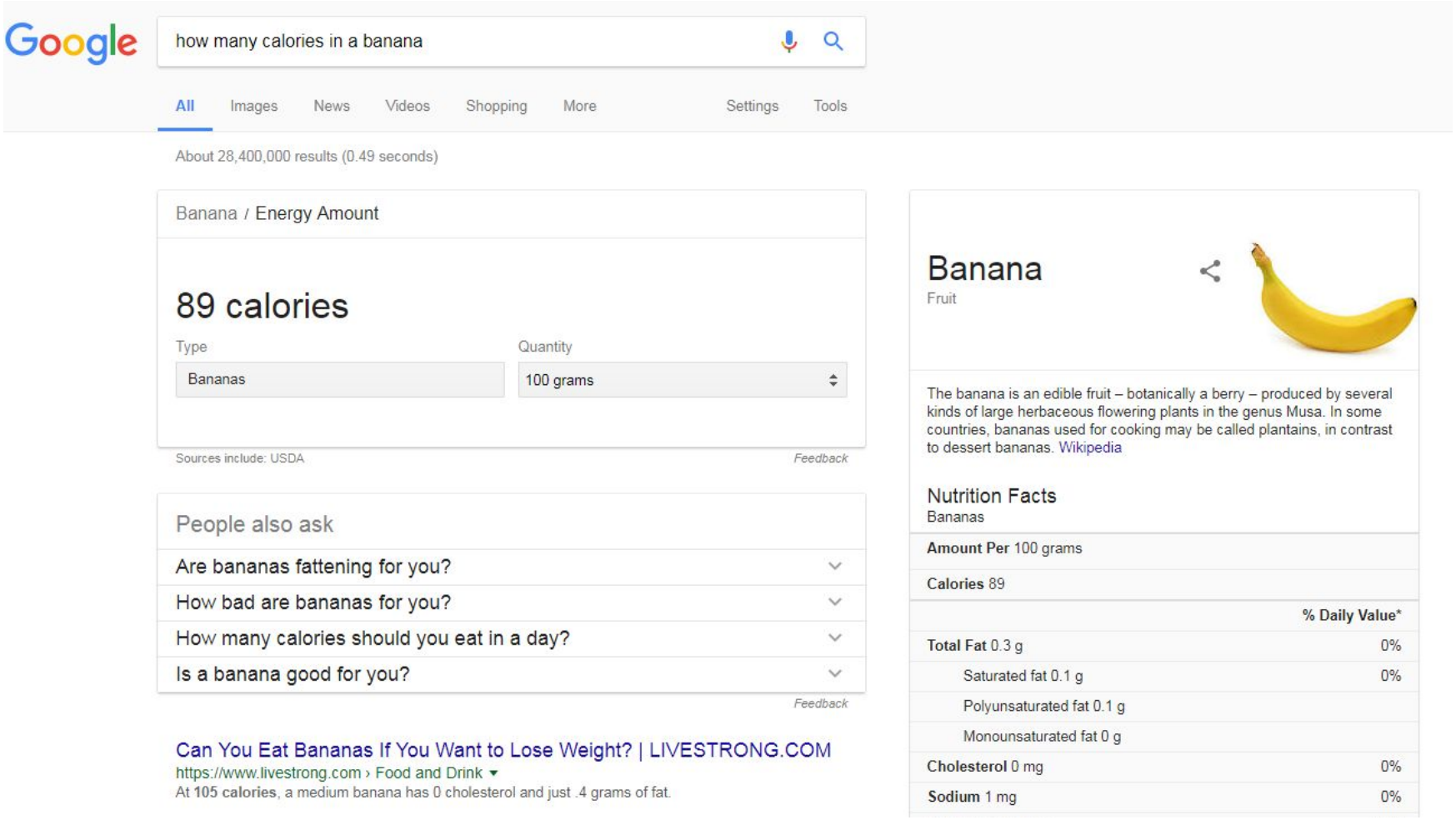
Search Engines

US Presidents list - PresidentsUSA.net
https://www.presidentsusa.net/presvplist.html
Listing of **Presidents in order** and their terms in office and Vice **Presidents** of the **United States**.
Political Parties of the Presidents · Thomas Jefferson · Donald Trump · George Bush

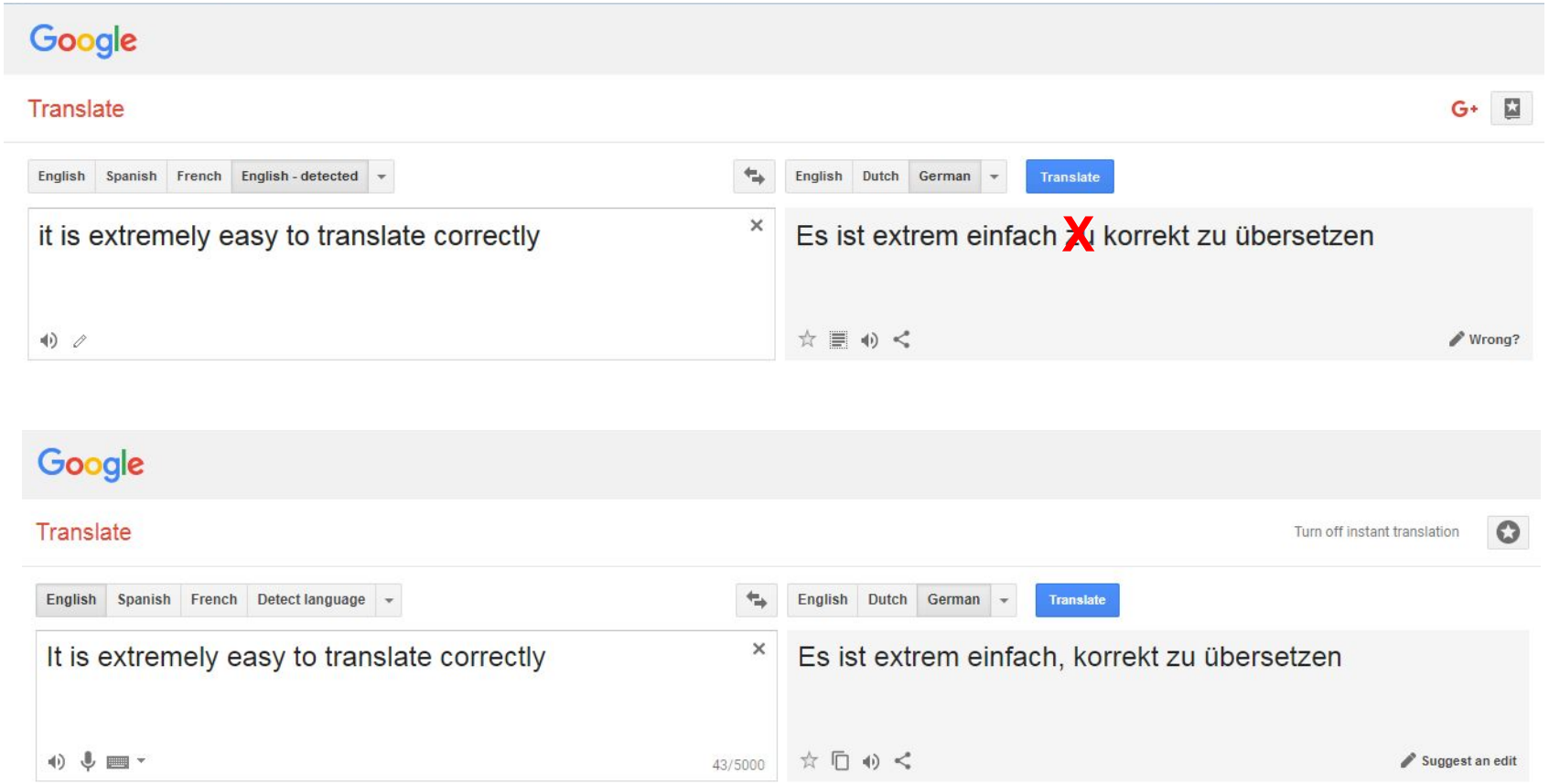
List of Presidents of the United States - Wikipedia
https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States
Of the elected presidents, four died in office of natural causes (William Henry Harrison, **Zachary Taylor**, **Warren G. Harding**, and **Franklin D. Roosevelt**), four were assassinated (**Abraham Lincoln**, **James A. Garfield**, **William McKinley** and **John F. Kennedy**), and one resigned (**Richard Nixon**).
List of Presidents of the United ... · Lifespan timeline of Presidents ...

Presidents of the United States (POTUS)
www.ipl.org/div/potus/
They are listed at the bottom of the page. **George Washington**, 1789-1797. **John Adams**, 1797-1801. **Thomas Jefferson**, 1801-1809. **James Madison**, 1809-1817. **James Monroe**, 1817-1825. **John Quincy Adams**, 1825-1829. **Andrew Jackson**, 1829-1837. **Martin Van Buren**, 1837-1841.

Search Engines

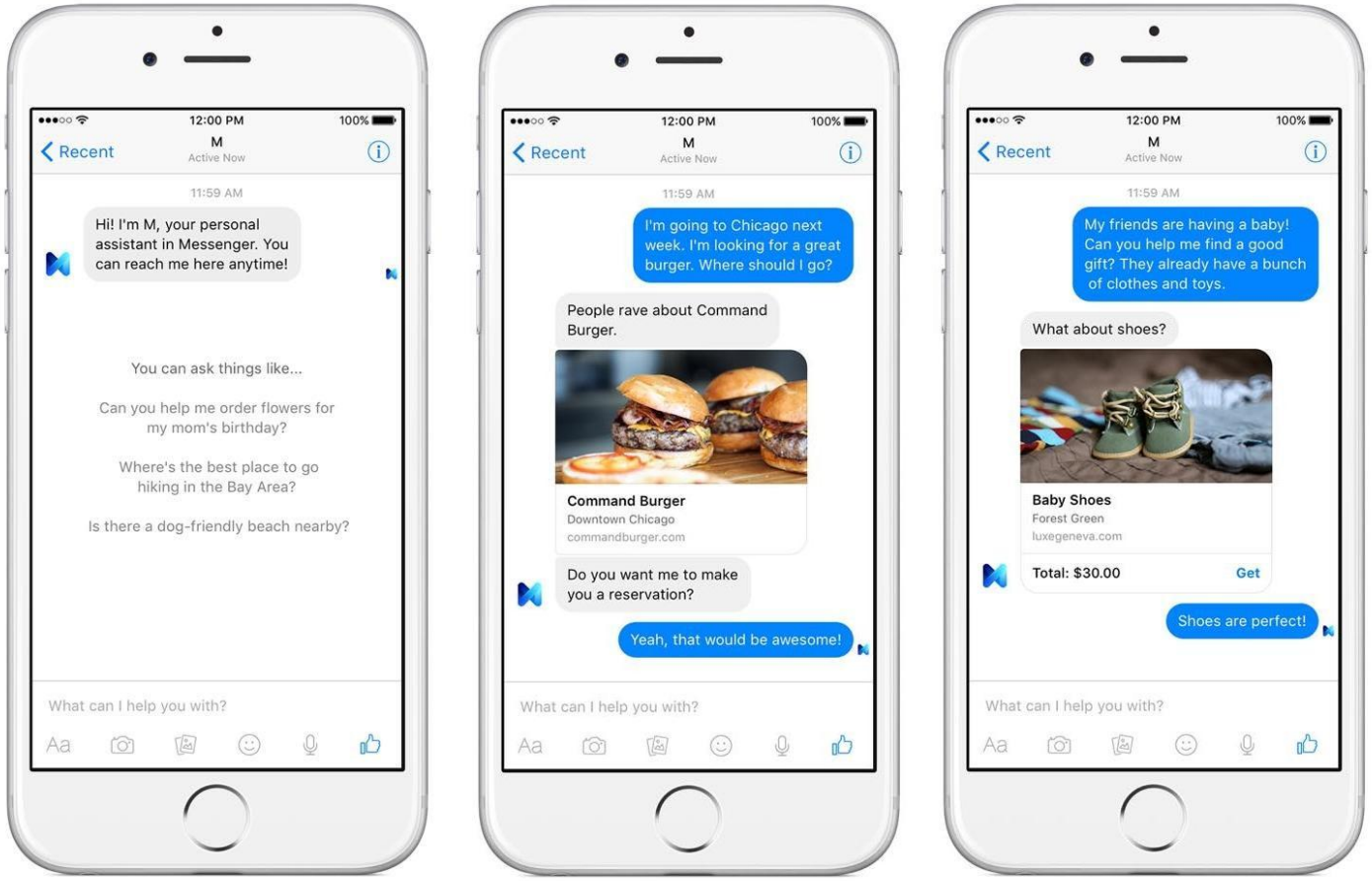


Machine Translation

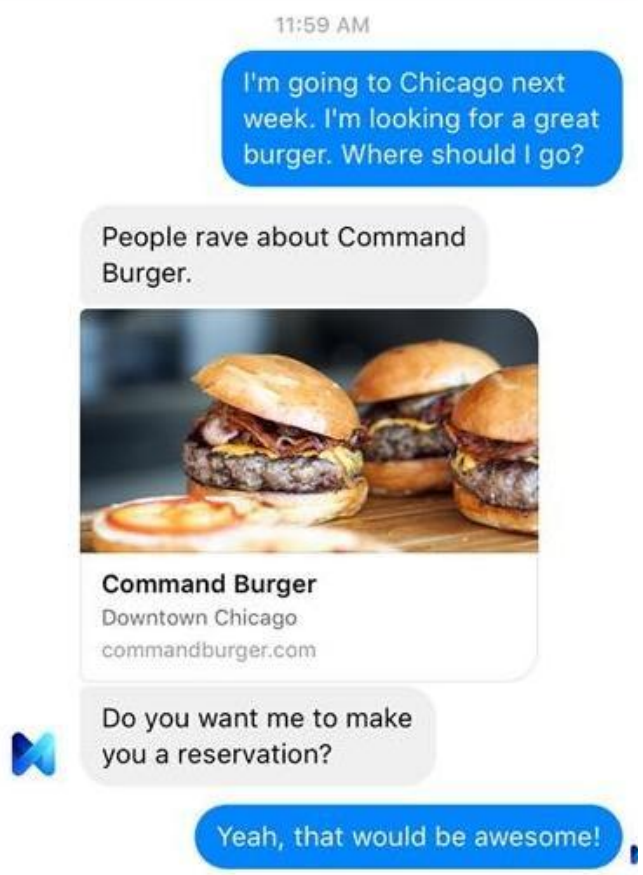


Chatbots

Chatbots - Dialog Systems



Chatbots - Dialog Systems

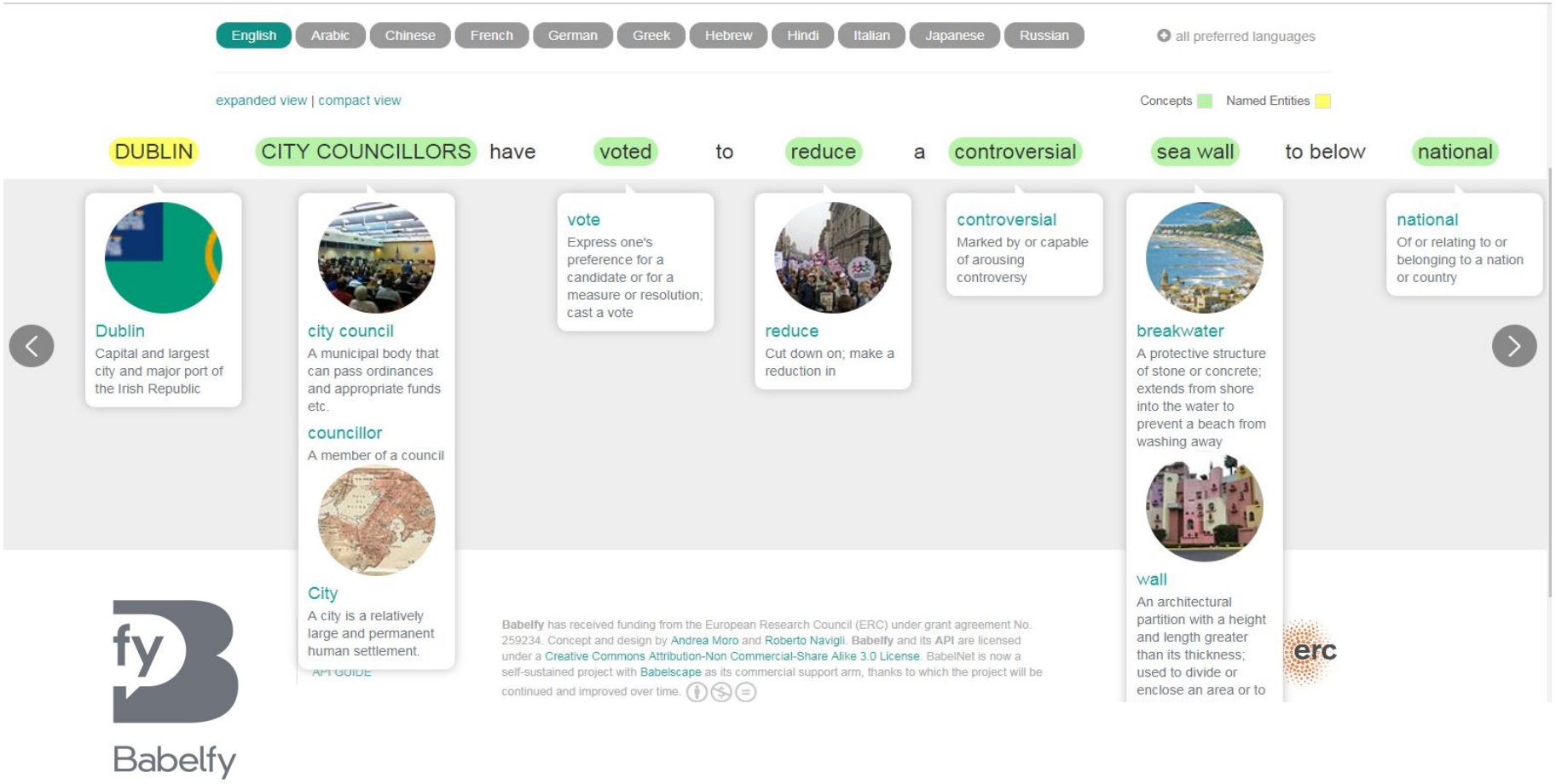


Opinion Mining

- Text analytics is now a requirement for large companies
- Multiple sources of unstructured data
- "Voice of the Customer" on steroids
- Ability to understand why NPS metrics are going up or down
- Ability to respond quickly to service issues
- Compelling ROI for recent deployments

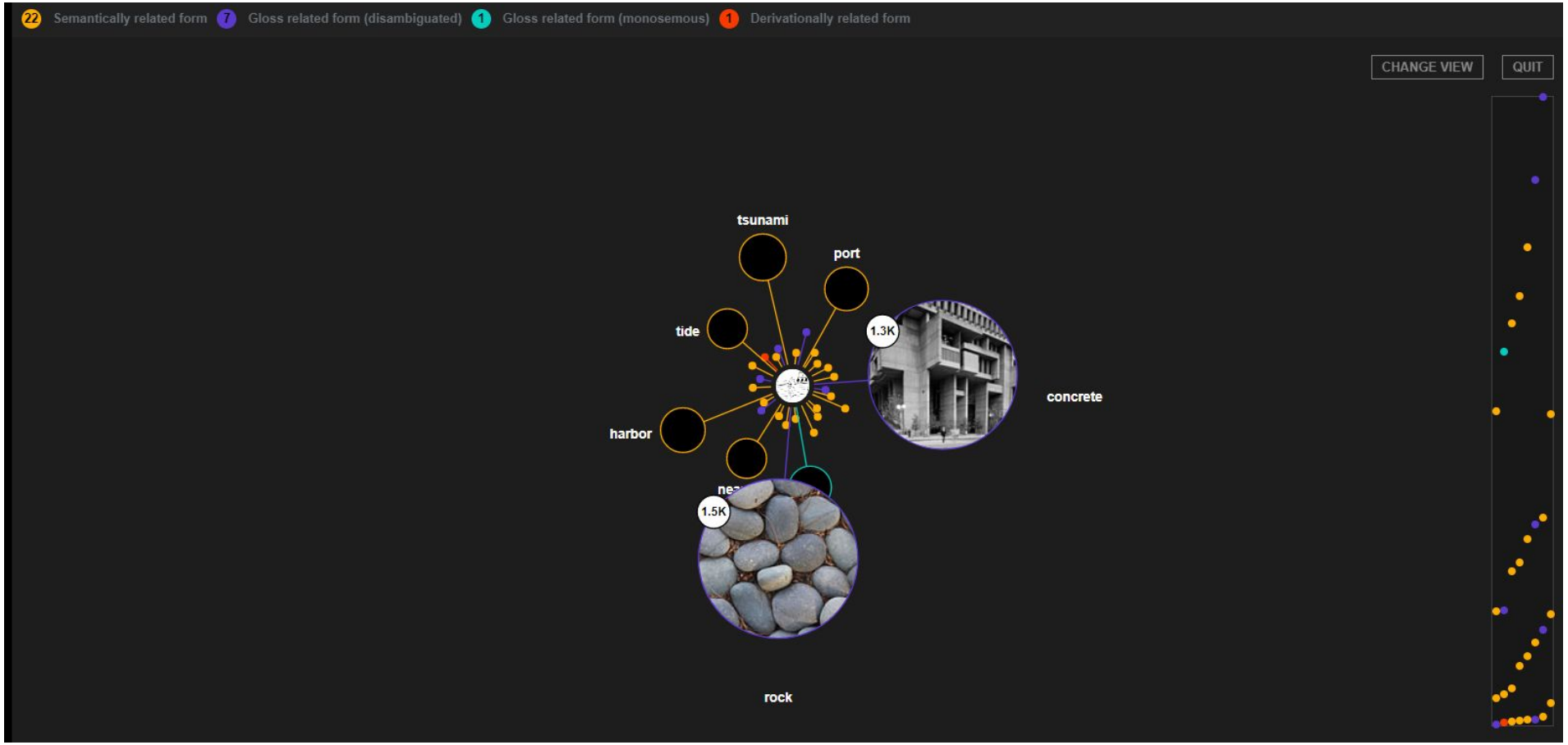
Opinion Mining

Knowledge Graphs



http://babelfy.org/

Knowledge Graphs



Overview

Organizational issues

Why NLP

Human language

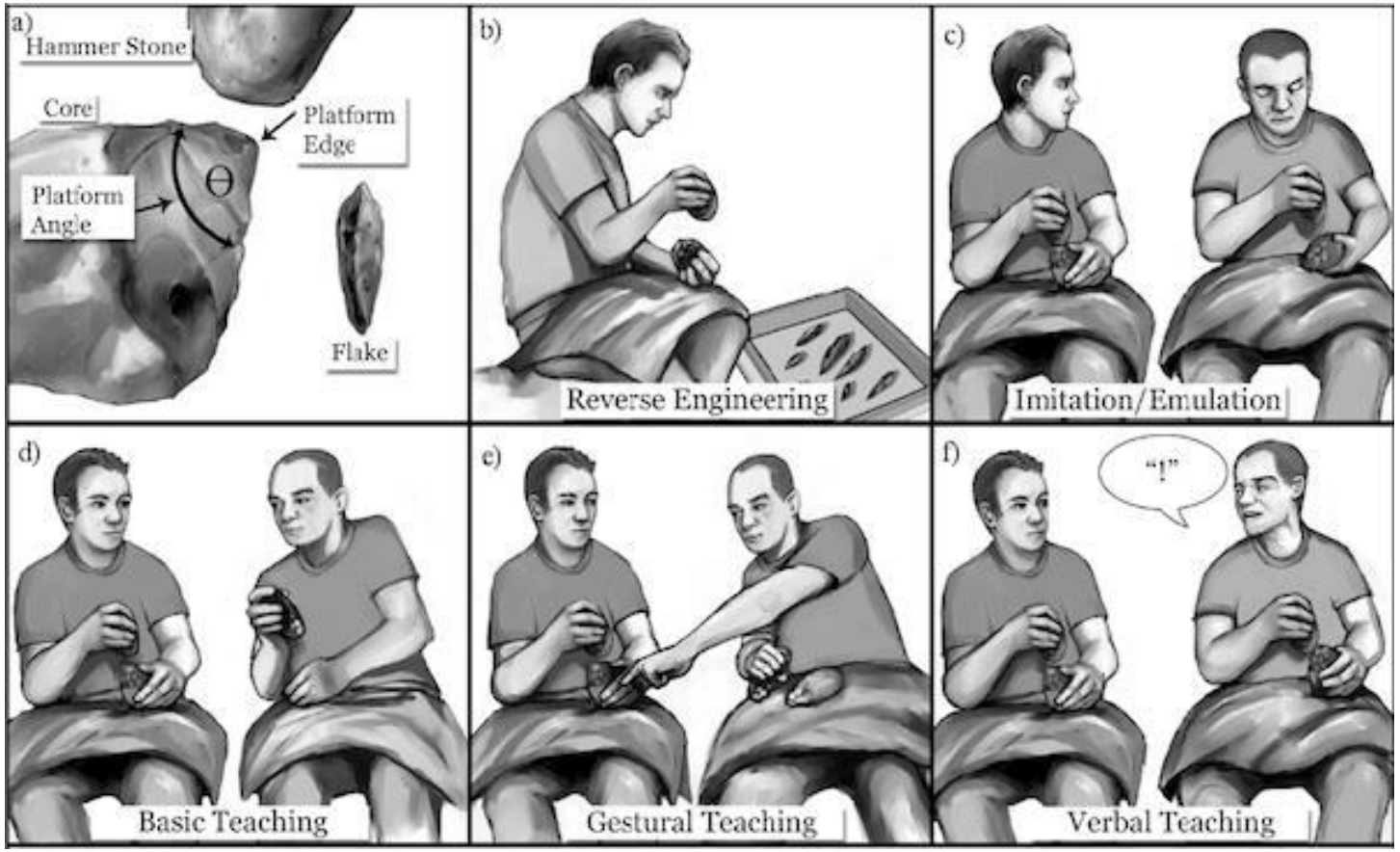
Some core notions in NLP

Architectures

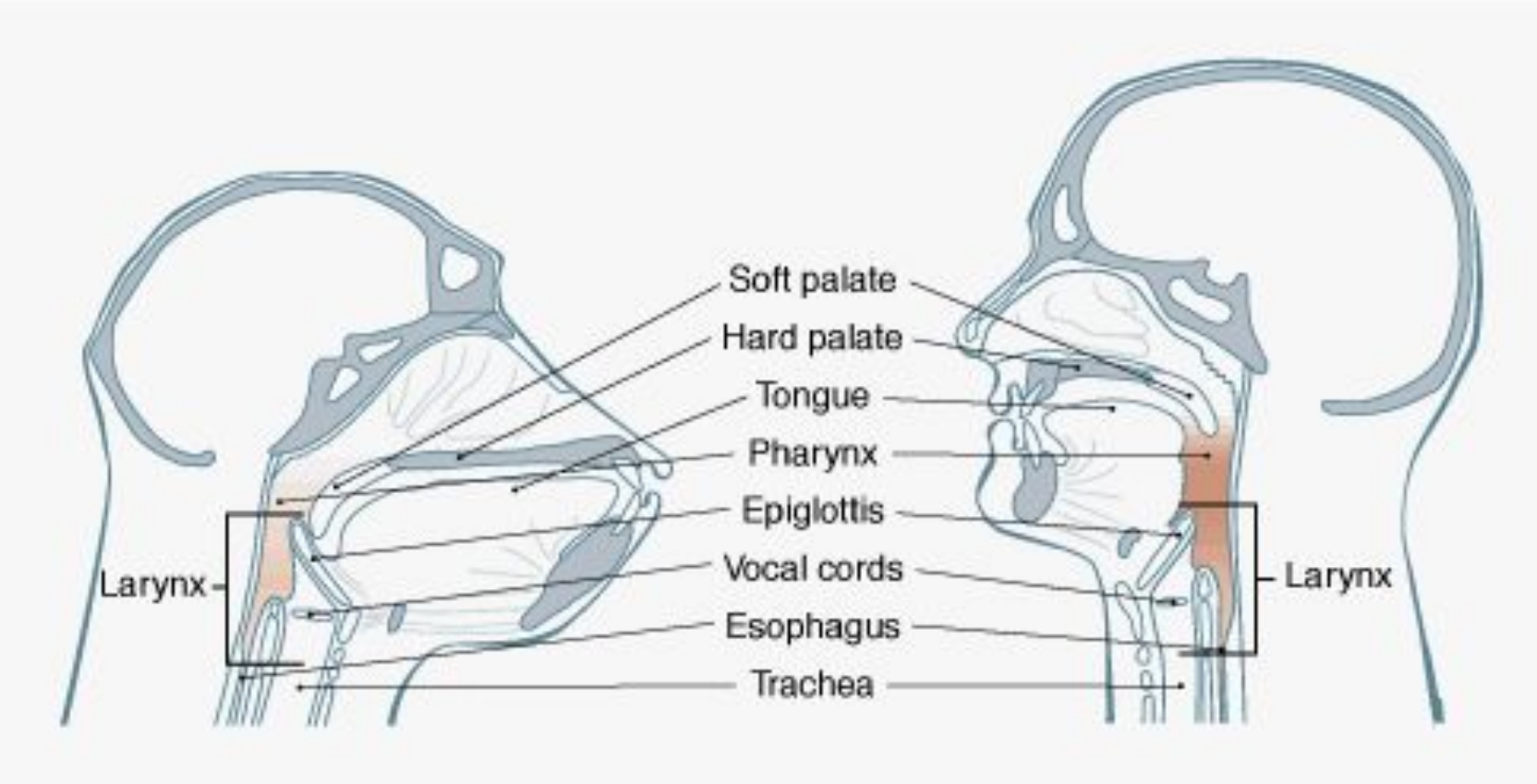
Applications

Useful pointers

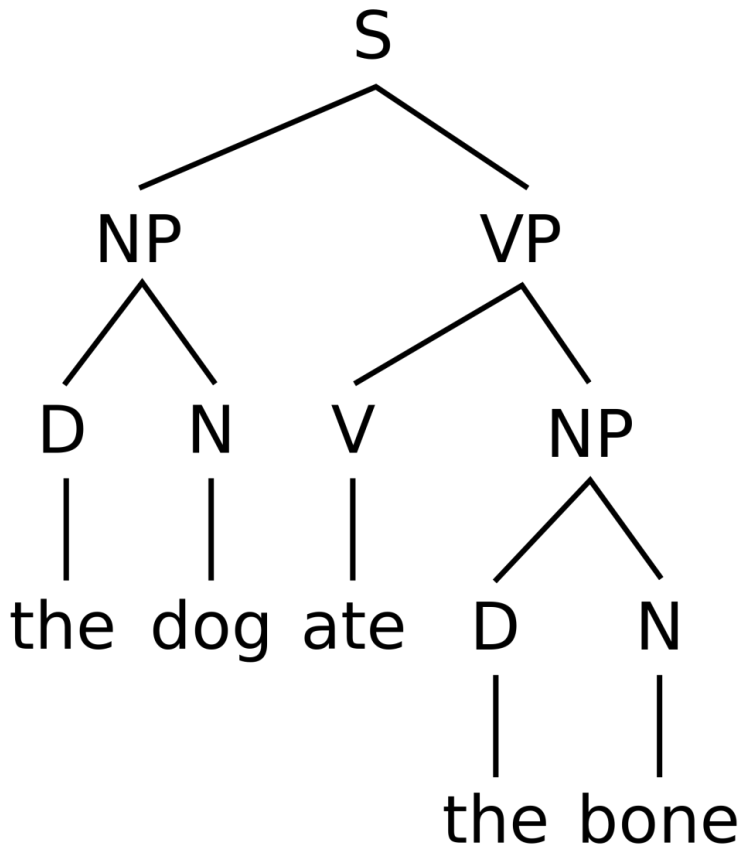
Human Language as a Social Tool



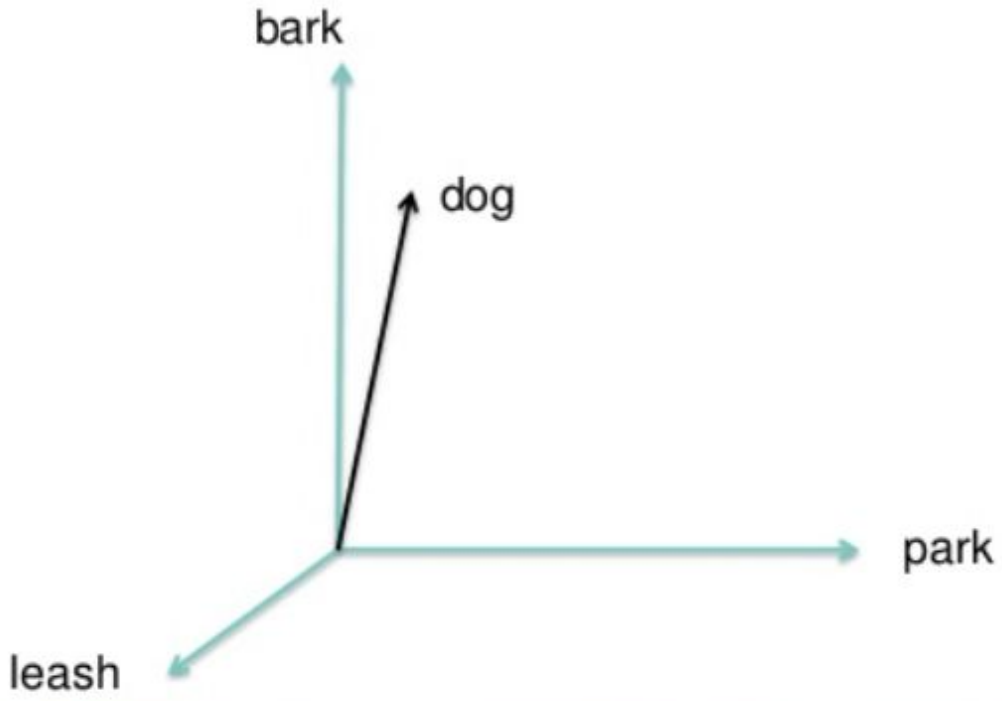
Human Language as a Biological System



Human Language as a Logical System



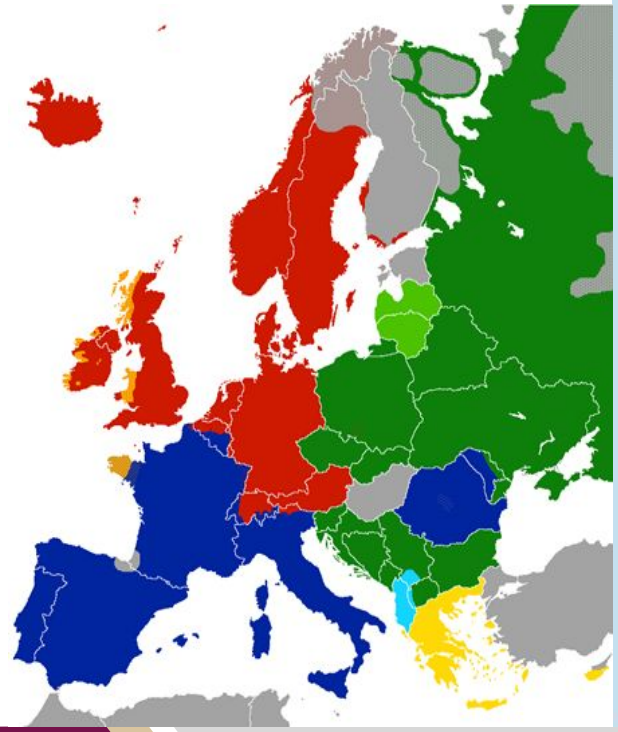
Human Language as a Mathematical System



Human Language and Politics

Language Families

NLP applications may be easier to transfer within language families



Overview

Organizational issues

Why NLP

Human language

Some core notions in NLP

Architectures

Applications

Useful pointers

Ambiguity

ambiguity is at the core of human language complexity

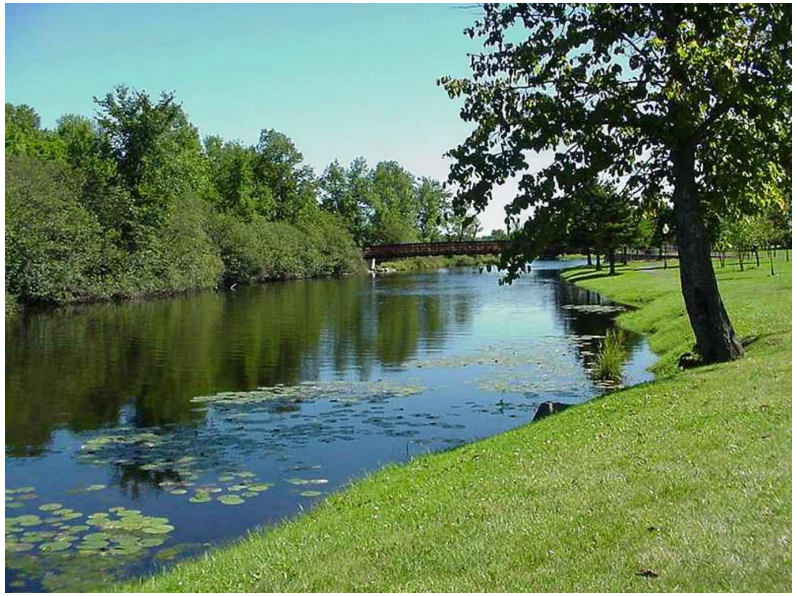
human language is

non-deterministic: *one word/phrase expressing different meanings*

redundant: *different words/phrases expressing the same meaning*

therefore **ambiguous**

Lexical Ambiguity



*savings **bank*** *vs.* *river **bank***

Syntactic Ambiguity

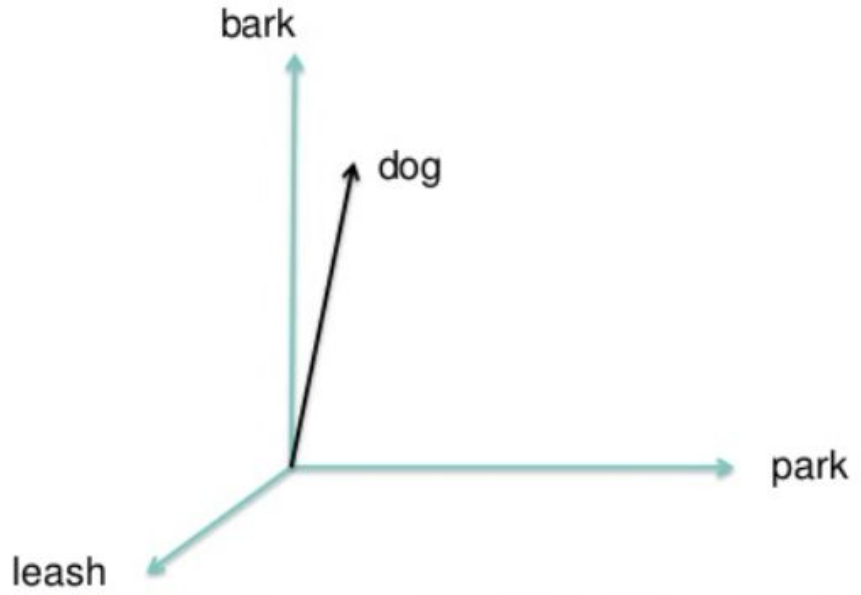
astronomers saw stars with telescopes

Vector Space Model & Word Embeddings

Distributional Semantics: words are similar in meaning if they share similar contexts

Represent words using **vectors (embeddings)**

Based on **language modeling**



Language Modeling

Frequency and co-occurrence are central to language modelling

Example: n-grams from the Corpus of Contemporary American English

frequency	word1	word2	word3
31891	much	of	the
13261	much	of	a
8000	much	more	than
7396	much	as	i
5650	much	the	same
5633	much	of	it
4229	much	better	than
4191	much	as	the

https://www.english-corpora.org/coca/
https://www.ngrams.info/

Overview

Organizational issues

Why NLP

Human language

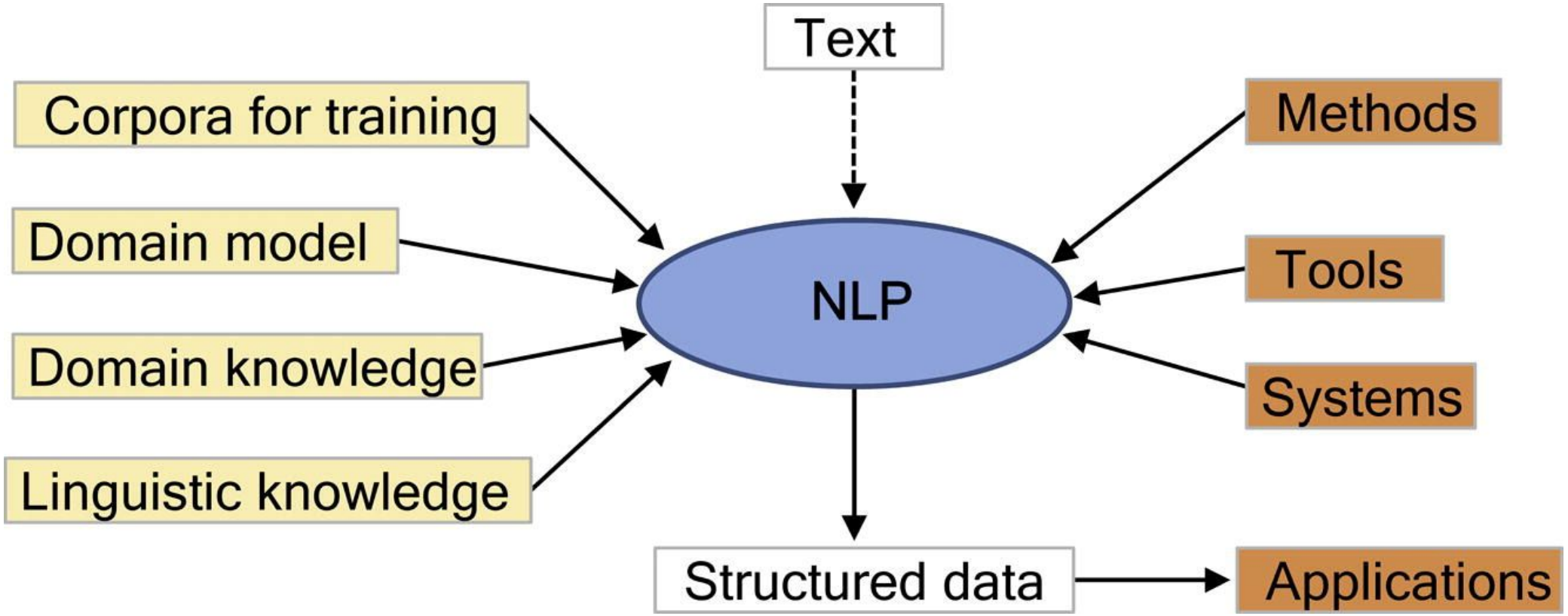
Some common notions in NLP

Architectures

Applications

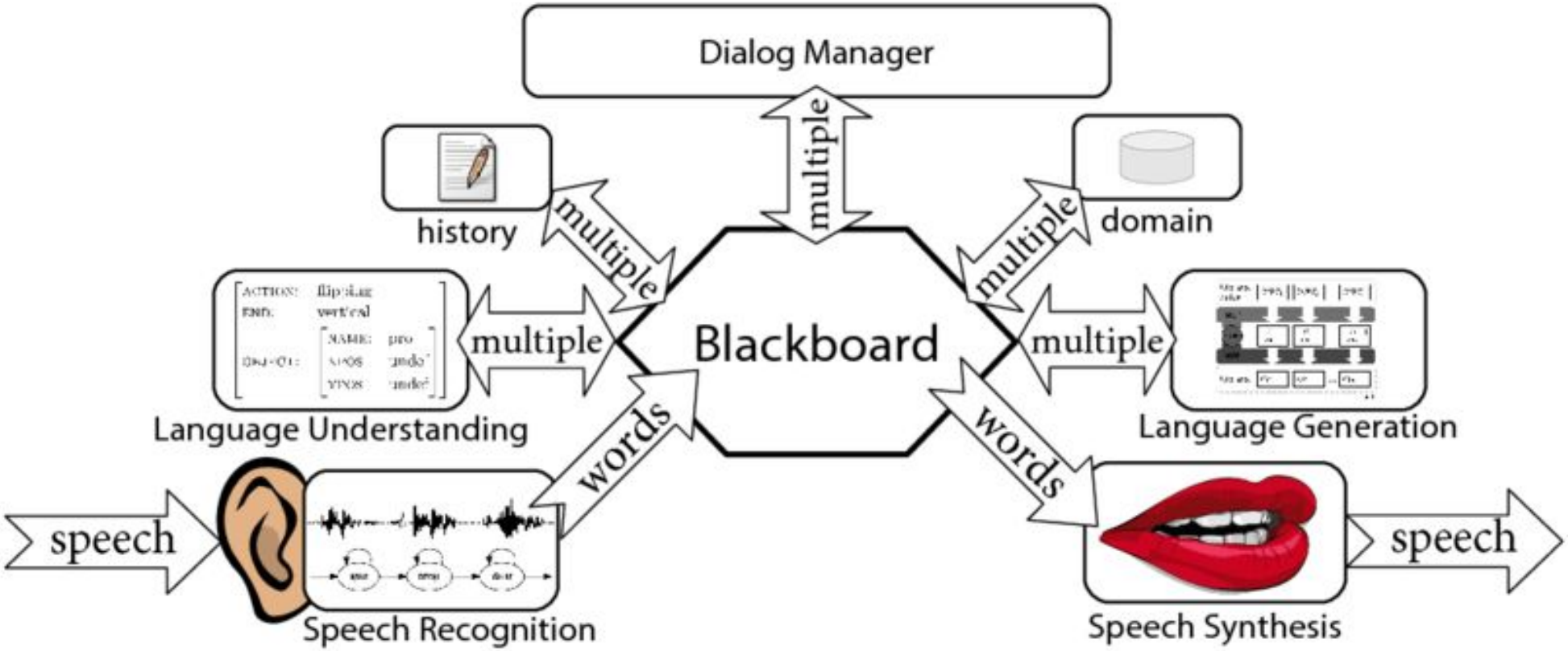
Useful pointers

NLP Architecture - General Overview



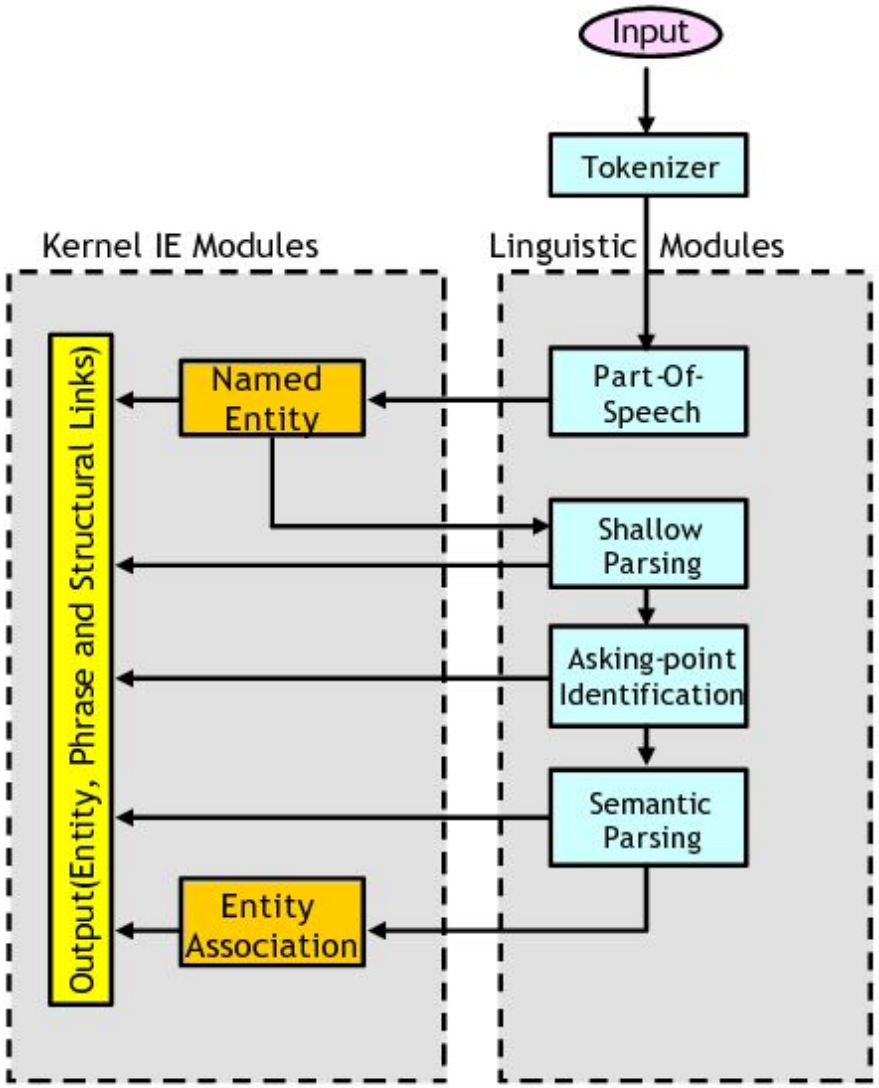
Friedman, C., Rindflesch, T.C. and Corn, M., 2013. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of biomedical informatics*, *46*(5), pp.765-773.

Blackboard Architecture



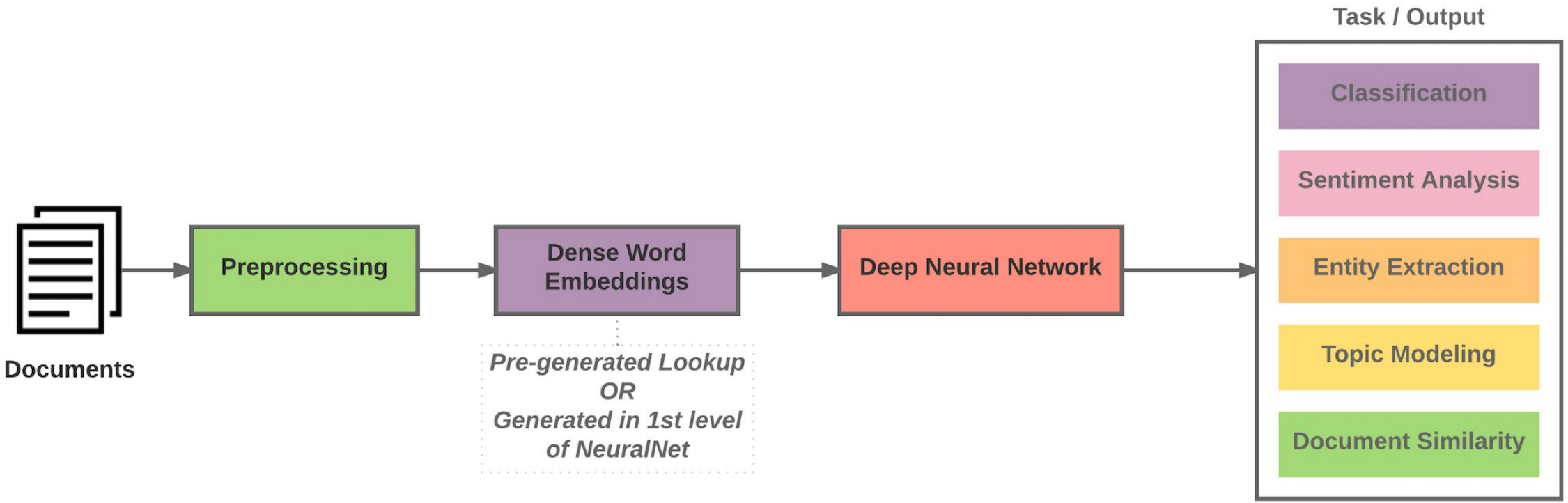
T. Baumann. Incremental Spoken Dialogue Processing: Architecture and Lower-level Components. PhD thesis, Universität Bielefeld, Germany, 2013.

Pipeline Architecture



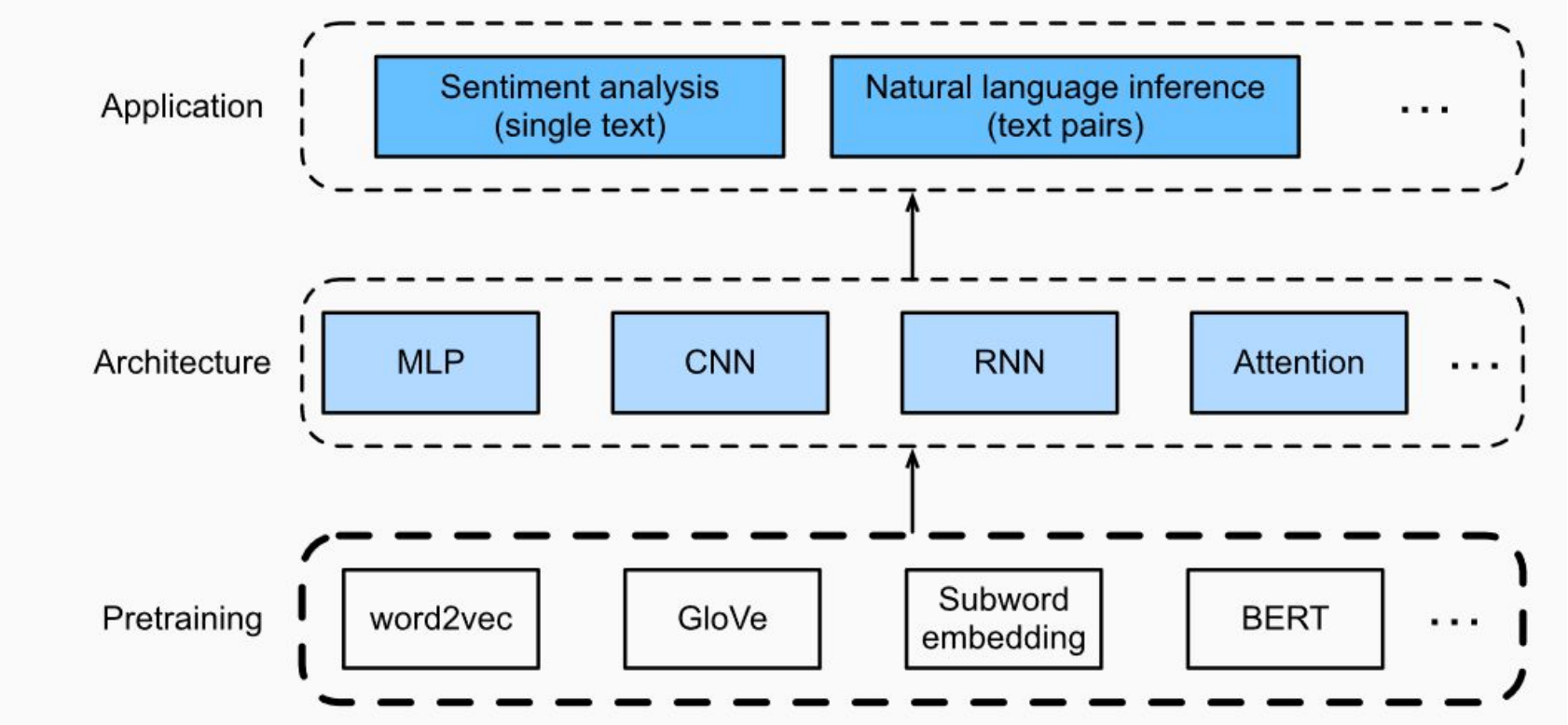
https://www.researchgate.net/figure/InfoXtract-NLP-IE-System-Architecture_fig1_228960112

Deep Learning Architecture ‘End-to-End’



http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/

Deep Learning - Pre-Trained Models



https://d2l.ai/chapter_natural-language-processing-pretraining/index.htm |

Overview

Organizational issues

Why NLP

Human language

Some common notions in NLP

Architectures

Applications

Useful pointers

NLP Applications

Classification such as Sentiment Analysis

Retrieval such as Question Answering

Natural Language **Generation** such as Dialog Systems (Chatbots)

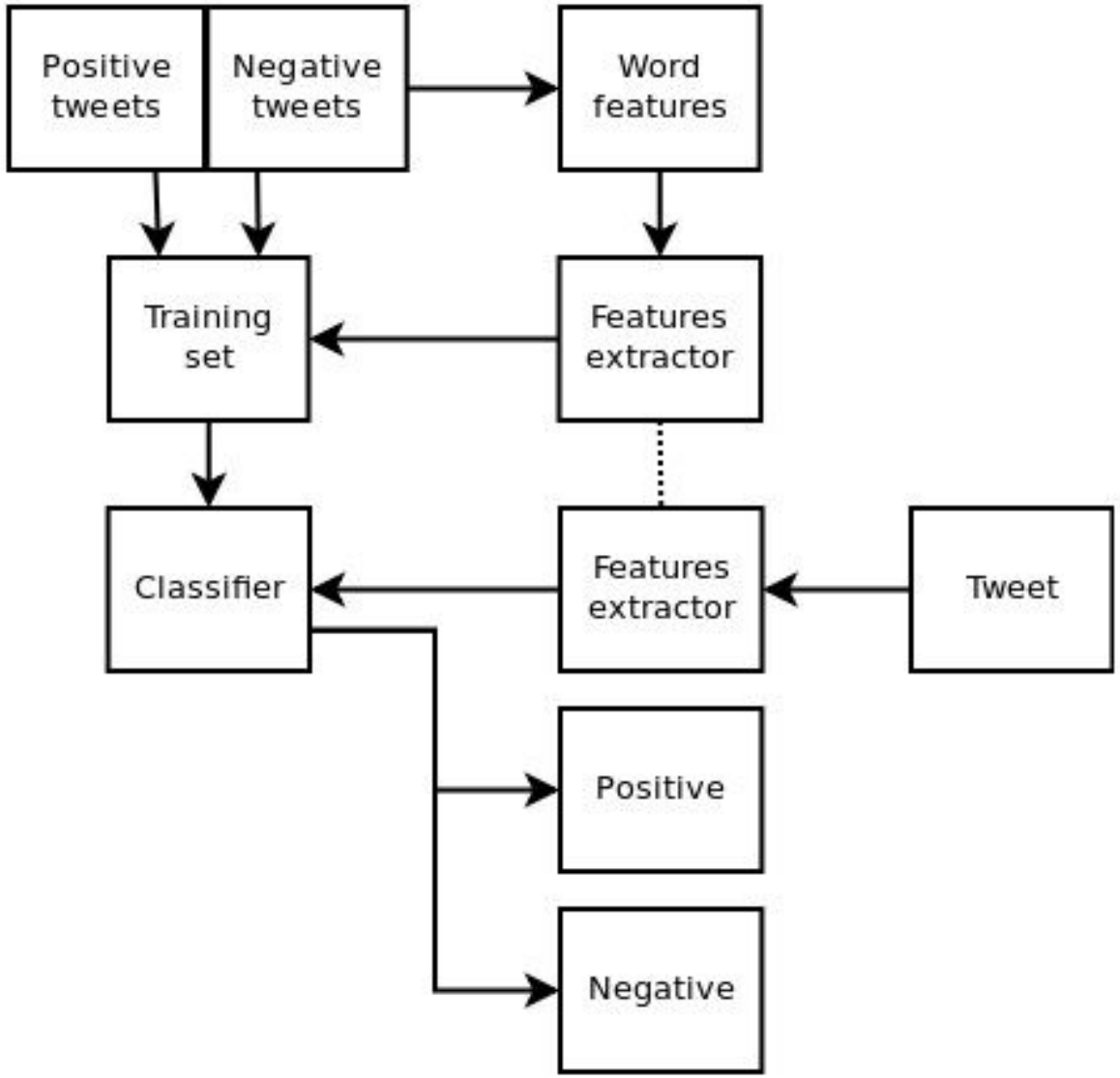
Transformation such as Machine Translation

NLP Applications - Classification

Many NLP tasks can be formulated as a **classification** problem, e.g.

Sentiment Analysis:	classification of a sentence as Pos or Neg
Part of Speech Tagging:	classification of a word as a Noun, Verb, Adj,...
Word Sense Disambiguation:	classification of ‘bank’ as Sense-1 or Sense-2

Sentiment Analysis



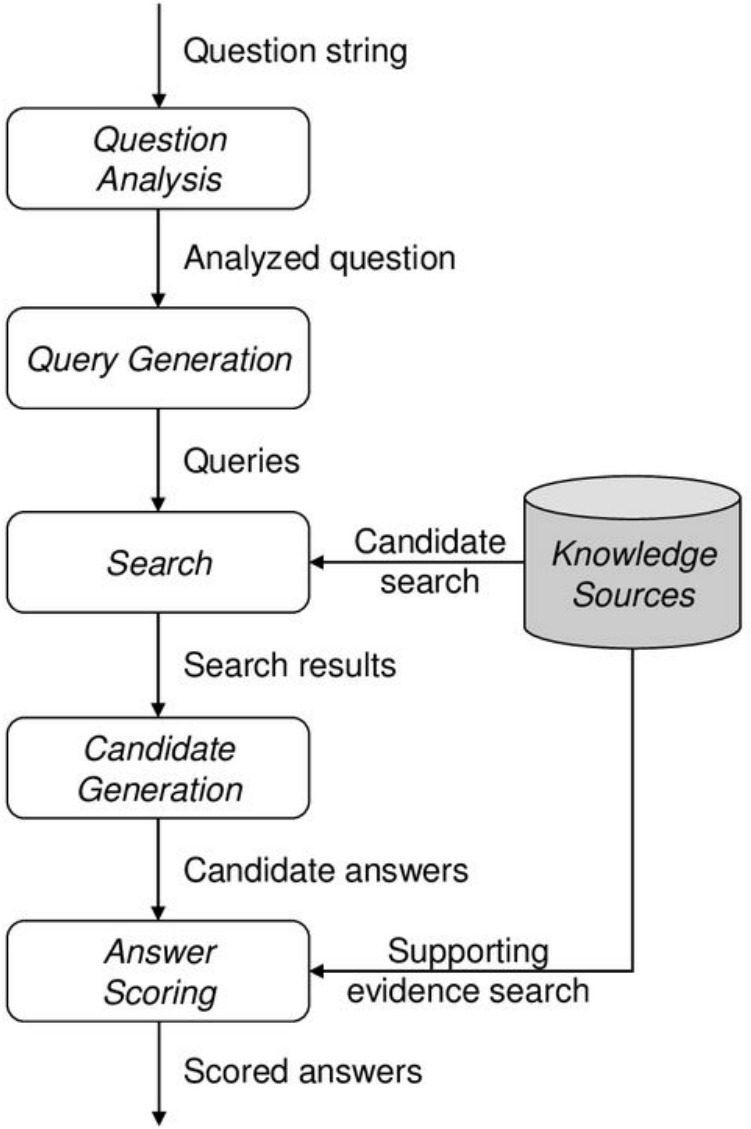
https://www.pantechsolutions.net/twitter-sentiment-analysis-using-machine-learning-algorithms-on-python

NLP Applications - Retrieval

NLP tasks based on **retrieval**, e.g.

- Question Answering

Question Answering



"Which computer scientist invented the smiley?"

Answer type: computer scientist
Keywords: invented, smiley

"computer scientist" invented smiley

The two original text smileys were invented on **September 19, 1982** by **Scott E. Fahlman** at **Carnegie Mellon**.

- September 19, 1982
- Scott E. Fahlman
- Carnegie Mellon

Answer	Score
Scott E. Fahlman	0.9
Carnegie Mellon	0.4
September 19, 1982	0.3

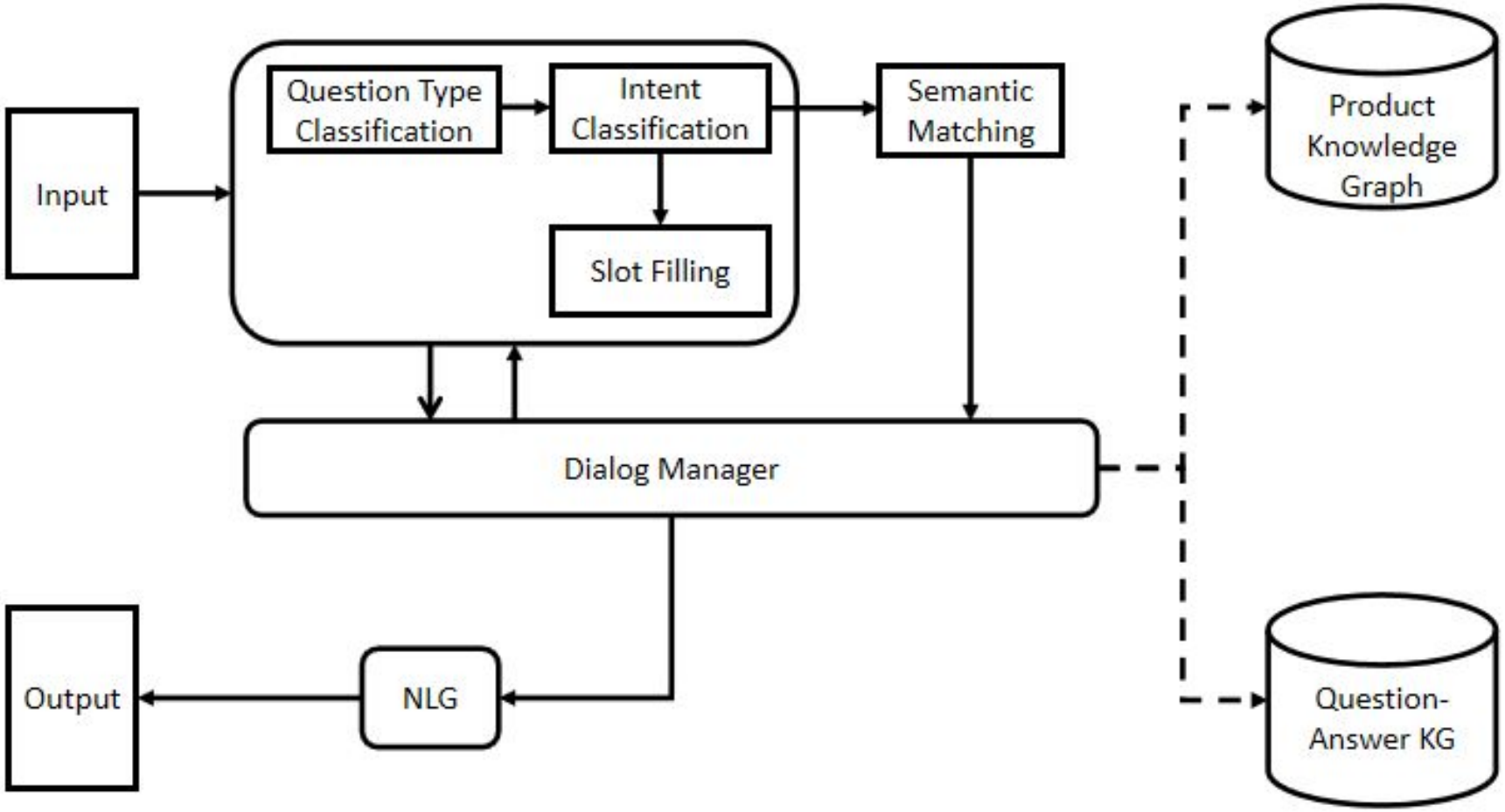
Schlaefer, N., 2011. *Statistical Source Expansion for Question Answering* (Doctoral dissertation, IBM TJ Watson Research Center).

NLP Applications - Generation

NLP tasks based on **generation**, e.g.

- Dialog Systems (Chatbots)
- Natural Language Generation from Data
- Abstractive Document Summarization

Dialog Systems (Chatbots)



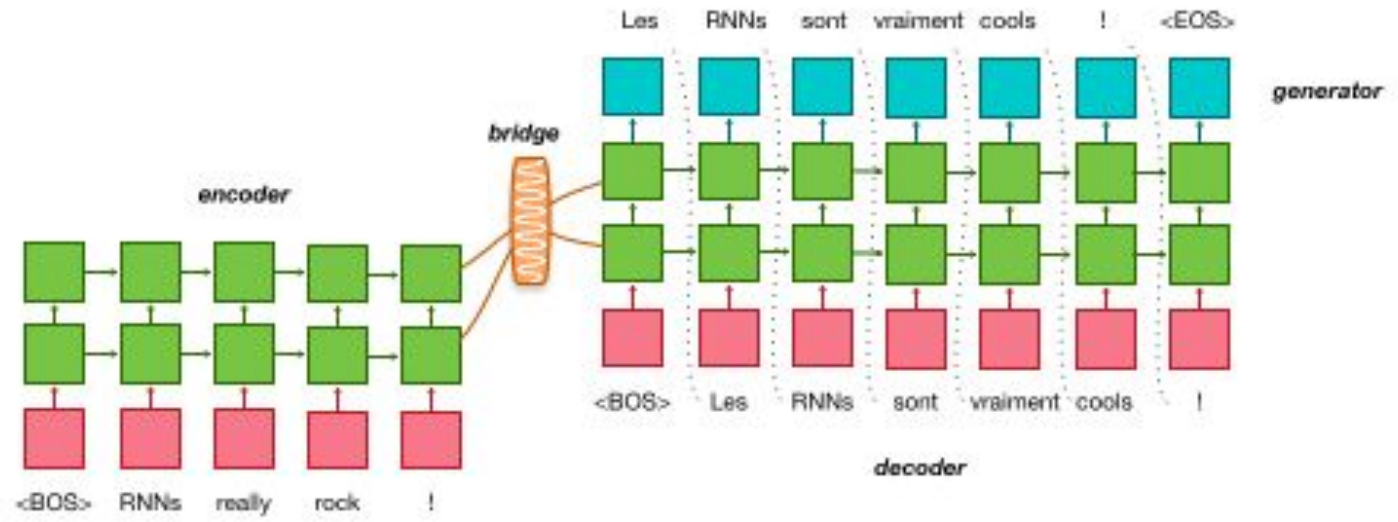
Zhao, G., Zhao, J., Li, Y., Alt, C., Schwarzenberg, R., Hennig, L., Schaffer, S., Schmeier, S., Hu, C. and Xu, F., 2019. MOLI: Smart Conversation Agent for Mobile Customer Service. *Information*, *10*(2), p.63.

NLP Applications - Transformation

NLP tasks based on **transformation**, e.g.

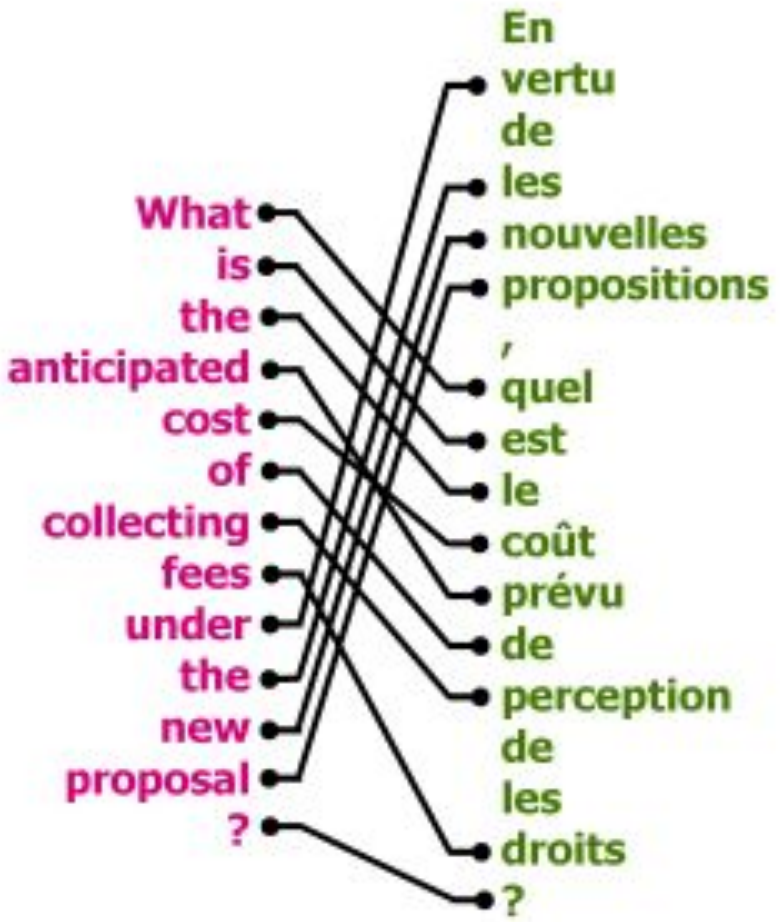
- Machine Translation

Neural Machine Translation



https://opennmt.net/OpenNMT/applications/

Bilingual Corpus



Overview

Organizational issues

Why NLP

Human language

Some core notions in NLP

Architectures

Applications

Useful pointers

NLP Publications - Conferences

Association for Computational Linguistics

COLING 2020 | The 28th International Conference on Computational Linguistics

NLP Publications - Journals

Transactions of the Association for Computational Linguistics

HOME ABOUT LOGIN REGISTER SEARCH CURRENT VOLUME ALL VOLUMES ANNOUNCEMENTS

Home > Volumes > Vol 3 (2015)

Vol 3 (2015)

Table of Contents

Reasoning about Quantities in Natural Language Subhro Roy, Tim Vieira, Dan Roth	PDF (PRESENTED AT NAACL 2015)	1-13
Cross-Document Co-Reference Resolution using Sample-Based Clustering with Knowledge Enrichment Sourav Dutta, Gerhard Weikum	PDF	15-28
Efficient Inference and Structured Learning for Semantic Role Labeling Oscar Täckström, Kuzman Ganchev, Dipanjan Das	PDF (PRESENTED AT ACL 2015)	29-41
SPRITE: Generalizing Topic Models with Structured Priors Michael J. Paul, Mark Dredze	PDF (PRESENTED AT NAACL 2015)	43-57
A Sense-Topic Model for Word Sense Induction with Unsupervised Data Enrichment Jing Wang, Mohit Bansal, Kevin Gimpel, Brian D. Ziebart, Clement T. Yu	PDF (PRESENTED AT NAACL 2015)	59-71
Which Step Do I Take First? Troubleshooting with Bayesian Models Annie Louis, Mirella Lapata	PDF	73-85
Gappy Pattern Matching on GPUs for On-Demand Extraction of Hierarchical Translation Grammars Hua He, Jimmy Lin, Adam Lopez	PDF (PRESENTED AT NAACL 2015)	87-100
Erratum: "Exploring Compositional Architectures and Word Vector Representations for Prepositional Phrase Attachment" Yonatan Belinkov, Tao Lei, Regina Barzilay, Amir Globerson	ERRATUM PDF ORIGINAL PAPER	101-101
A Bayesian Model of Grounded Color Semantics Brian McMahan, Matthew Stone	PDF (PRESENTED AT NAACL 2015)	103-115
Exploiting Parallel News Streams for Unsupervised Event Extraction Congle Zhang, Stephen Soderland, Daniel S. Weld	PDF (PRESENTED AT ACL 2015)	117-129
Unsupervised Declarative Knowledge Induction for Constraint-Based Learning of Information Structure in Scientific Documents Yufan Guo, Roi Reichart, Anna Korhonen	PDF (PRESENTED AT NAACL 2015)	131-143
Entity Disambiguation with Web Links Andrew Chisholm, Ben Hachey	PDF (PRESENTED AT NAACL 2015)	145-156
An Unsupervised Method for Uncovering Morphological Chains Karthik Narasimhan, Regina Barzilay, Tommi Jaakkola	PDF (PRESENTED AT ACL 2015)	157-167

March 2005 Volume 31, Number 1

Computational Linguistics

ACL Lifetime Achievement Award

Some Points in a Time — Karen Spärck Jones — 1

Squibs and Discussions

Real versus Template-Based Natural Language Generation: A False Opposition? — Kees van Deemter, Emiel Krahmer, and Mariët Theune — 15

Articles

Discriminative Reranking for Natural Language Parsing — Michael Collins and Terry Koo — 25

The Proposition Bank: An Annotated Corpus of Semantic Roles — Martha Palmer, Daniel Gildea, and Paul Kingsbury — 71

Clustering Syntactic Positions with Similar Semantic Requirements — Pabla Gamallo, Alexandre Agustini, and Gabriel P. Lopes — 107

Book Reviews

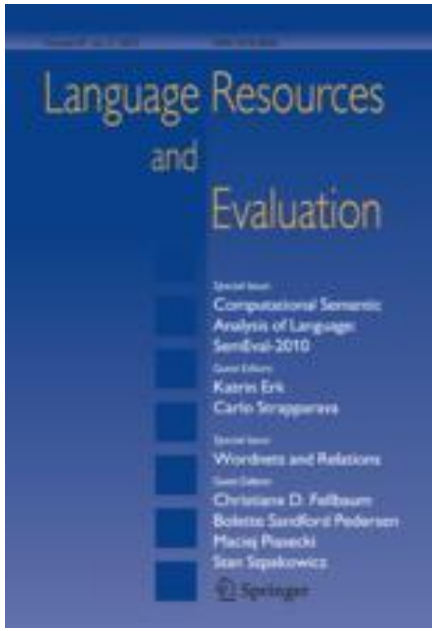
Ontological Semantics by Sergei Nirenburg and Victor Raskin — Reviewed by John F. Sowa — 147

Argumentation Machines: New Frontiers in Argumentation and Computation edited by Chris Reed and Timothy J. Norman — Reviewed by Ingrid Zukerman — 153

Briefly Noted — 157

Publications Received — 159

Published Quarterly by the MIT Press for the Association of Computational Linguistics



Language Resources and Evaluation

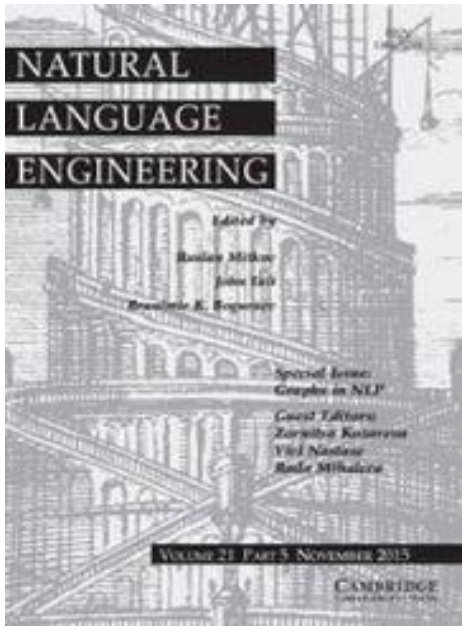
Special Issue: Computational Semantic Analysis of Language: SemEval-2010

Guest Editors: Katrin Erk, Carlo Strapparava

Special Issue: Wordnets and Relations

Guest Editors: Christiane D. Fellbaum, Bolette Sandford Pedersen, Maciej Piasecki, Stan Szpakowicz

Springer



NATURAL LANGUAGE ENGINEERING

Edited by

Special Issue: Graphs in NLP

Guest Editors: Zornitsa Kozareva, Yifei Nastase, Rada Mihalcea

VOLUME 21 PART 5 NOVEMBER 2015

CAMBRIDGE

NLP Publications - Ranking

Google Scholar

Categories > Engineering & Computer Science > **Computational Linguistics**

	Publication	h5-index	h5-median
1.	Meeting of the Association for Computational Linguistics (ACL)	106	168
2.	Conference on Empirical Methods in Natural Language Processing (EMNLP)	88	157
3.	Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)	61	104
4.	International Workshop on Semantic Evaluation	49	89
5.	Transactions of the Association for Computational Linguistics	47	80
6.	International Conference on Language Resources and Evaluation (LREC)	45	71
7.	International Conference on Computational Linguistics (COLING)	41	68
8.	Conference of the European Chapter of the Association for Computational Linguistics (EACL)	36	62
9.	Computer Speech & Language	36	48
10.	Conference on Computational Natural Language Learning (CoNLL)	34	51
11.	Workshop on Machine Translation	27	45
12.	IEEE Spoken Language Technology Workshop (SLT)	26	43
13.	Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)	26	39
14.	Computational Linguistics	25	51
15.	Language Resources and Evaluation	23	34

NLP Forums - Mailing Lists

Corpora List

http://clu.uni.no/icame/corpora/

Linguist List

http://www.linguistlist.org/

ACL - Professional Association for NLP

Association for Computational Linguistics (ACL)

http://www.aclweb.org/

ACL Anthology: archive of NLP conference publications
https://www.aclweb.org/anthology/

Lab of this Week

Intro to NLTK - labs will use NLTK in Python

https://www.nltk.org/

Advanced Topics in NLP - CT5121

Optional Semester 2 course by Dr. John McCrae & Dr. Mihael Arčan

- Word Embeddings
- Recurrent Neural Network (RNN) & Long Short-Term Memory (LSTM)
- Machine Translation
- Textual Similarity
- Topic Models
- Linguistic Linked Open Data
- Lexicography and Under-resourced Languages

QA