Chapter NLP:I

I. Introduction to Linguistics

- ❑ Goals of Language Technology
- ❑ Examples of NLP Systems
- ❑ NLP Problems
- ❑ Lingusitic Levels & Terminology
- ❑ Historical Background

Goals of Language Technology

1. Aid humans in writing.

   Correcting mistakes, formulating and paraphrasing text, transcription.

2. Identify texts related to spoken or written requests.

   Text information retrieval, semantic text similarity, question answering.

3. Make sense of texts without reading the originals.

   Categorization, information extraction, summarization, translation.

4. Instruct, and be advised by a computer.

   Audio interfaces (e.g., dialog systems, robotics), learning and assessment.

↓

5. Converse with computers as if they were human.

   Turing test, conversational AI and chatbots, computational humor.

   What is the nature of language and its relation to (artificial) intelligence?

Examples of NLP Systems

Writing Aid: Spelling and Grammar Checking

Alan Turing

“Alan Mathison Turing (23 June 1912 – 7 June 1954) was an english mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist. Turing was highly influential in the developing of theoretical computer science, providing a formalisation of the concepts of algorithm and computatoin with the Turing machine, who can be considered a model of general-purpose computer. Turing is widely considered to been the farther of theoretical computer science and artificial intelligance. Despite these accomplishment he was ever fully recognised in his home country during his lifetime due to his homosexuality and because many of his work was covered by the Official Secrets Act.”

Can you spot any errors?

Examples of NLP Systems

Writing Aid: Spelling and Grammar Checking

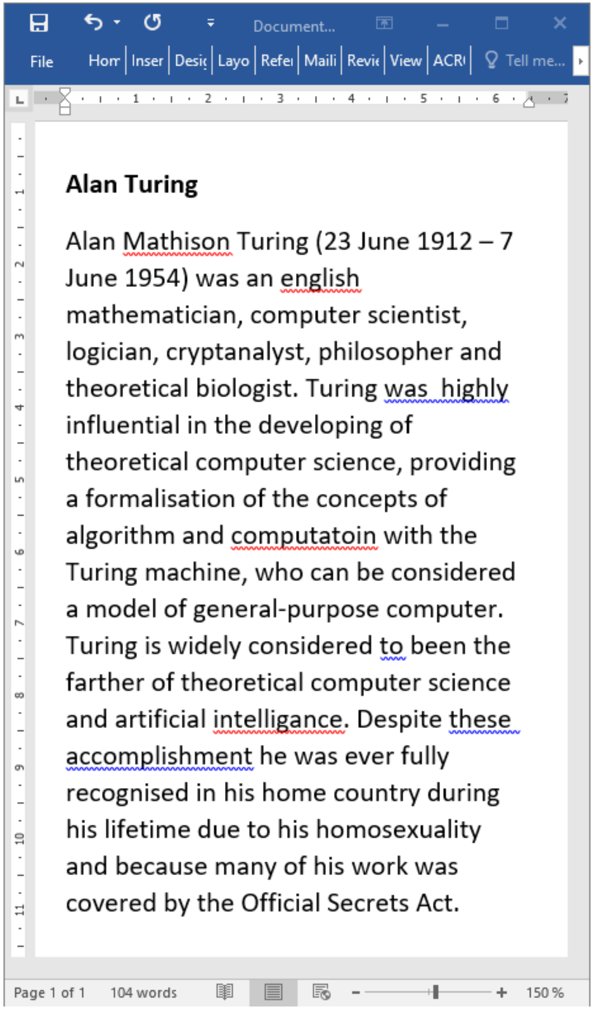
Alan Turing

“Alan Mathison Turing (23 June 1912 – 7 June 1954) was an english mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist. Turing was highly influential in the developing of theoretical computer science, providing a formalisation of the concepts of algorithm and computatoin with the Turing machine, who can be considered a model of general-purpose computer. Turing is widely considered to been the farther of theoretical computer science and artificial intelligance. Despite these accomplishment he was ever fully recognised in his home country during his lifetime due to his homosexuality and because many of his work was covered by the Official Secrets Act.”

Can you spot any errors?

Examples of NLP Systems

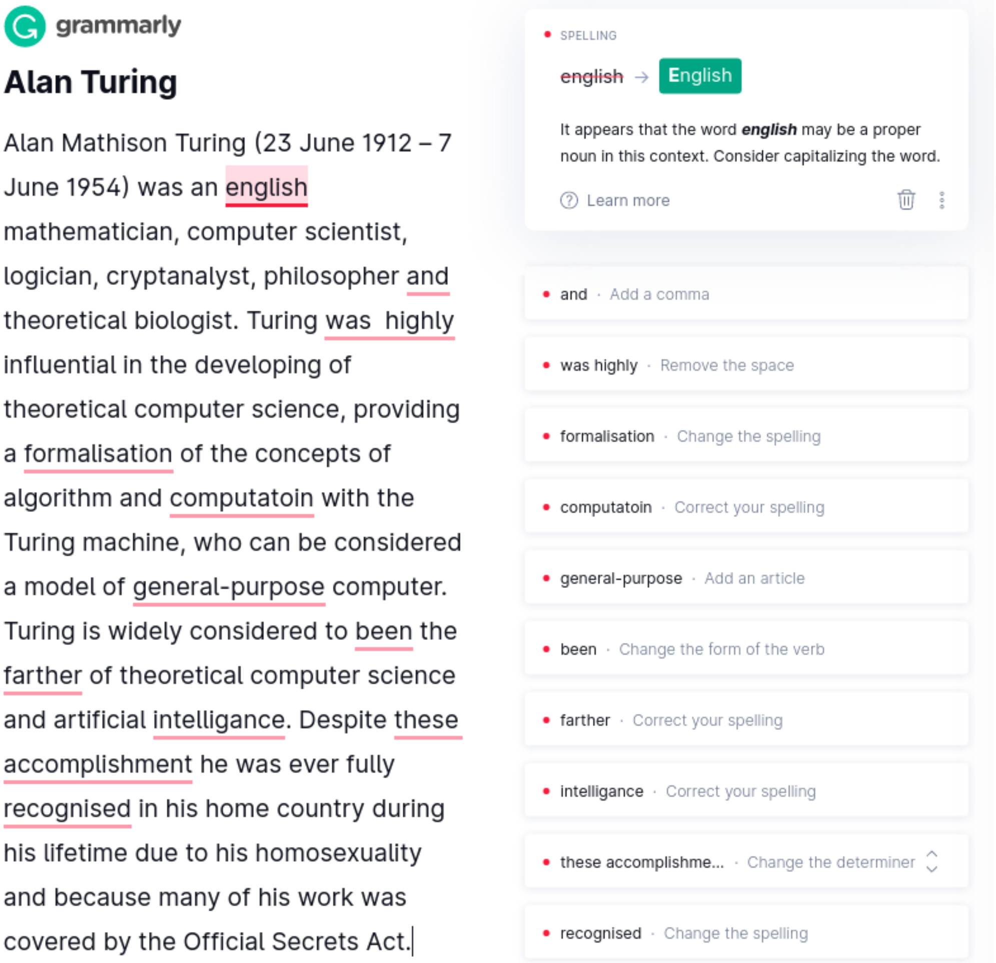
Writing Aid: Spelling and Grammar Checking



Alan Turing

Alan Mathison Turing (23 June 1912 – 7 June 1954) was an english mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist. Turing was highly influential in the developing of theoretical computer science, providing a formalisation of the concepts of algorithm and computatoin with the Turing machine, who can be considered a model of general-purpose computer. Turing is widely considered to been the farther of theoretical computer science and artificial intelligance. Despite these accomplishment he was ever fully recognised in his home country during his lifetime due to his homosexuality and because many of his work was covered by the Official Secrets Act.

Page 1 of 1 104 words 150 %



Alan Turing

Alan Mathison Turing (23 June 1912 – 7 June 1954) was an english mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist. Turing was highly influential in the developing of theoretical computer science, providing a formalisation of the concepts of algorithm and computatoin with the Turing machine, who can be considered a model of general-purpose computer. Turing is widely considered to been the farther of theoretical computer science and artificial intelligance. Despite these accomplishment he was ever fully recognised in his home country during his lifetime due to his homosexuality and because many of his work was covered by the Official Secrets Act.

SPELLING

~~english~~ → English

It appears that the word ***english*** may be a proper noun in this context. Consider capitalizing the word.

Learn more

- and · Add a comma
- was highly · Remove the space
- formalisation · Change the spelling
- computatoin · Correct your spelling
- general-purpose · Add an article
- been · Change the form of the verb
- farther · Correct your spelling
- intelligance · Correct your spelling
- these accomplishme... · Change the determiner
- recognised · Change the spelling

Remarks:

- The text is derived from the opening paragraph of the Alan Turing article on Wikipedia.
- Detected errors:
  - “english” should be capitalized (both)
  - “and” should be preceded by a comma; the Oxford comma (Grammarly)
  - “was highly” should only have one space between them (both)
  - “formalisation” could be switched to American English spelling (Grammarly)
  - “computatoin” is a spelling mistake (both)
  - “general-purpose” should be preceeded by the article “a” (Grammarly)
  - “to been” should be in present tense “be” (both, but Word for the wrong reason)
  - “farther” should be “father” (Grammarly)
  - “intelligance” should be “intelligence” (both)
  - “these accomplishment” should be “these accomplishments” (both)
  - “recognised” could be switched to American English spelling (Grammarly)
- False detections and undetected errors:
  - “Mathison” is correctly spelled; it is a false positive (Word)
  - “developing” should be development; it is a false negative (both)
  - “who” should be “which”; it is a false negative (both)
  - “ever” should be “never”; it is false negative (both)
  - “many” should be “much”; it is a false negative (both)

Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy

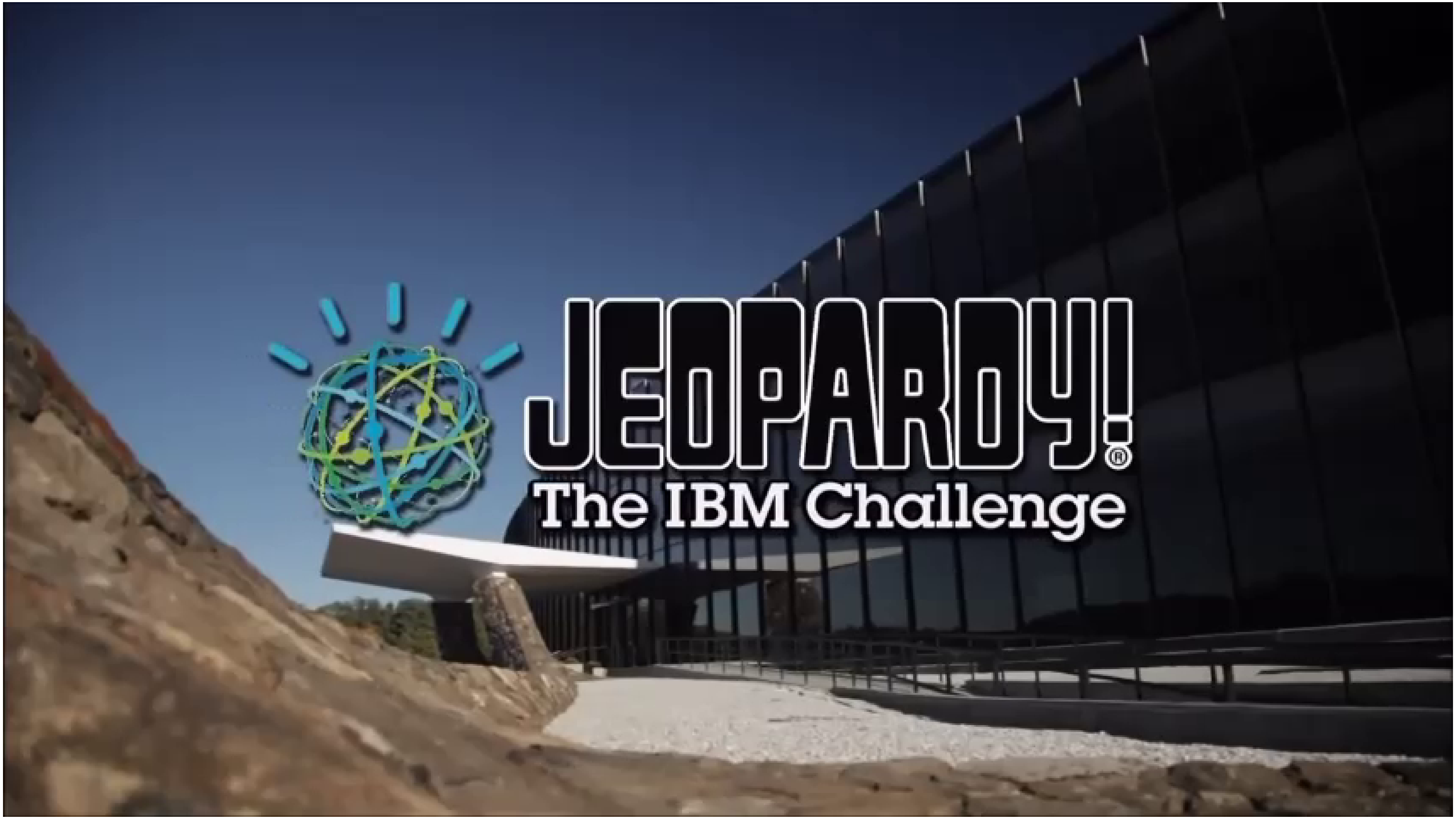
Jeopardy!

- American television quiz show running since the 1960s
- several general knowledge topics (e.g. history, literature, popular culture) at different dollar values
- participants presented with *clues in the form of answers*
- must formulate their *responses in the form of questions*

- between the 1960s and 2011 several returning champions; among others, Rutter and Jennings
- 2011: Rutter and Jennings vs. 200 million pages of content + AI (structured and unstructured, including full 2011 Wikipedia; ca. 4Tb of storage)

Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy (continued)

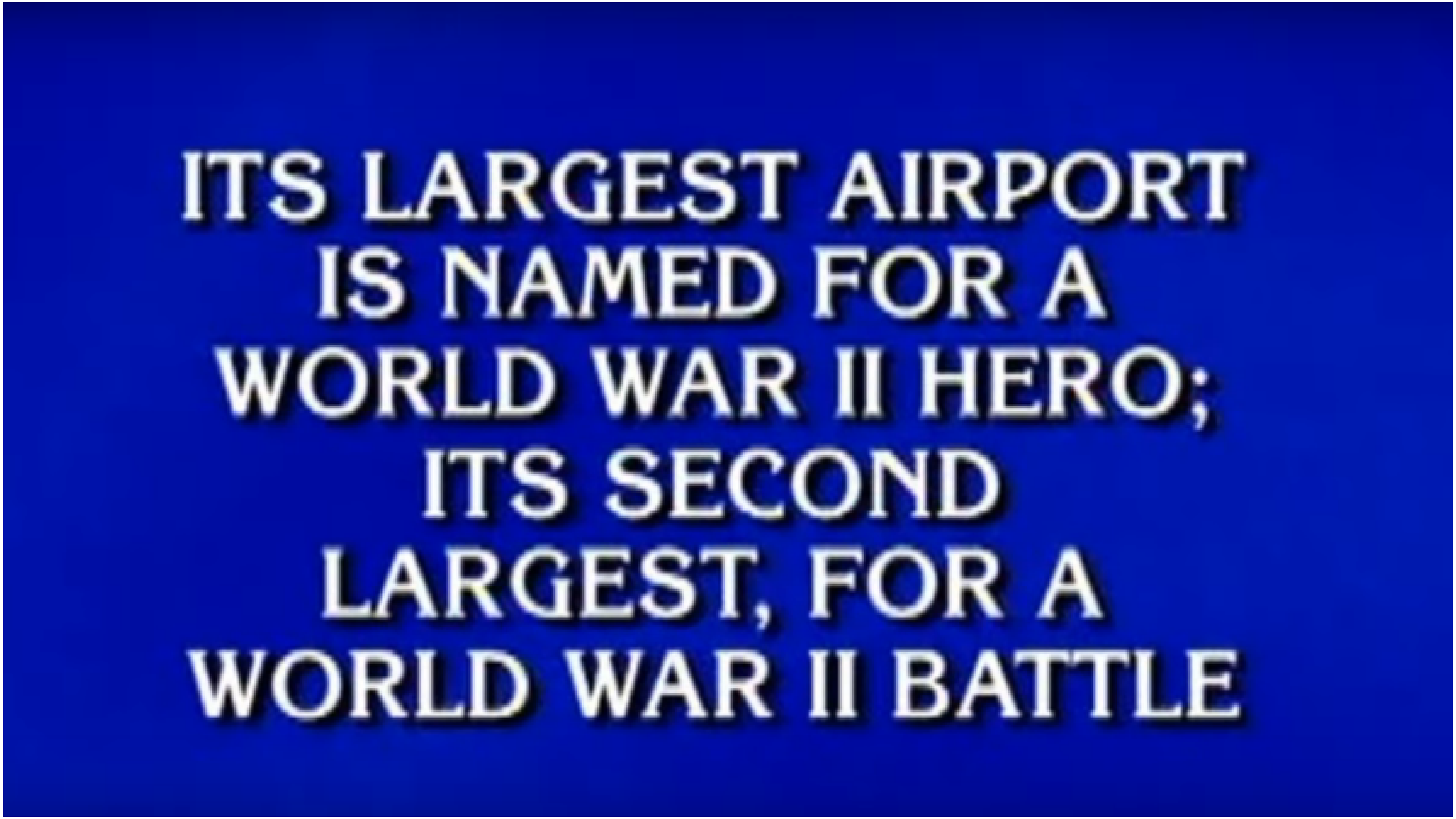


Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy (continued)

Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy (continued)



ITS LARGEST AIRPORT IS NAMED FOR A WORLD WAR II HERO; ITS SECOND LARGEST, FOR A WORLD WAR II BATTLE

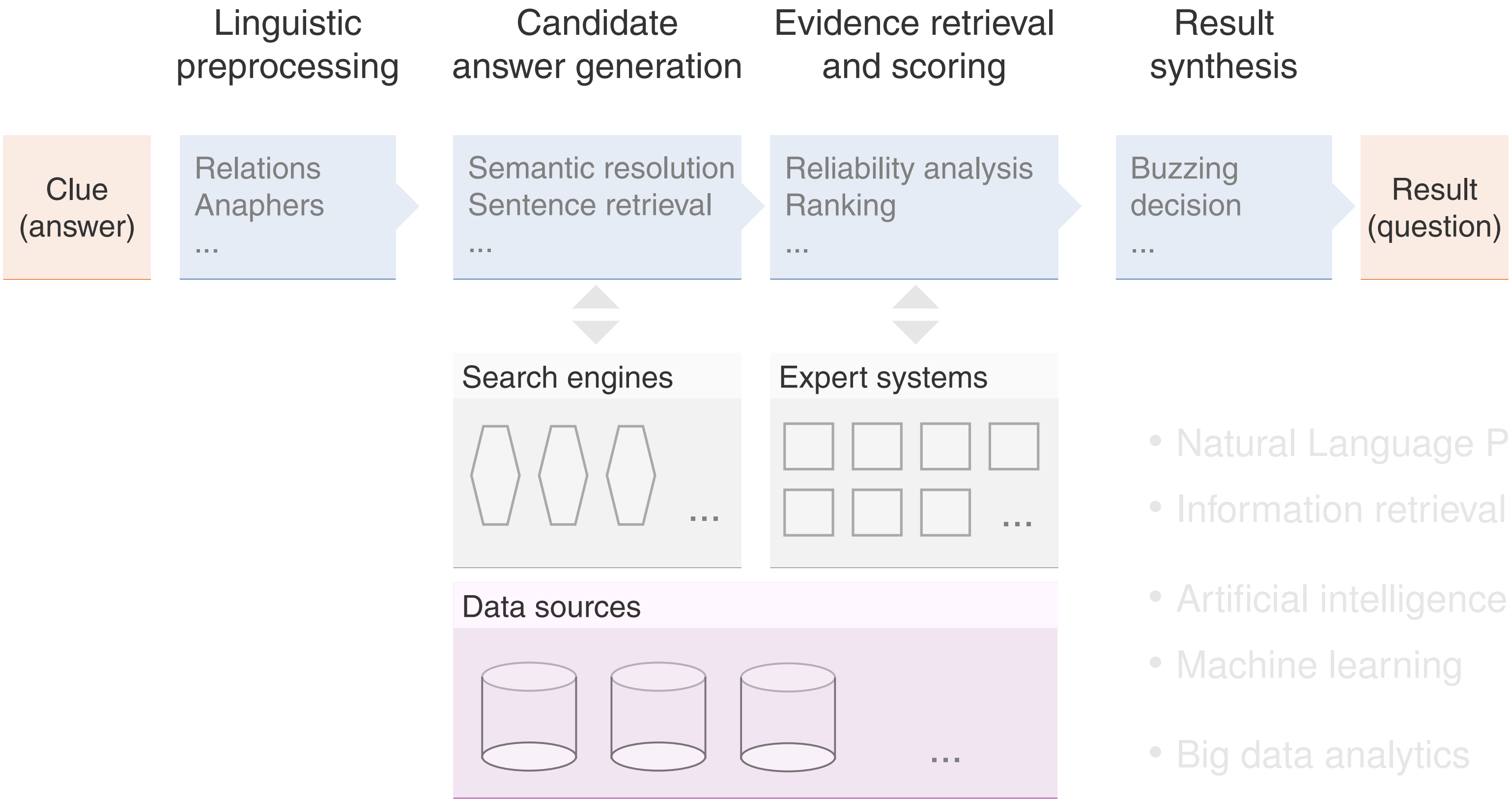
[IBM Watson at Jeopardy: Chicago, Toronto]

Remarks:

- Why did Watson think Toronto was in the U.S.A.?
  - mindmatters.ai
  - ibm.com

Examples of NLP Systems

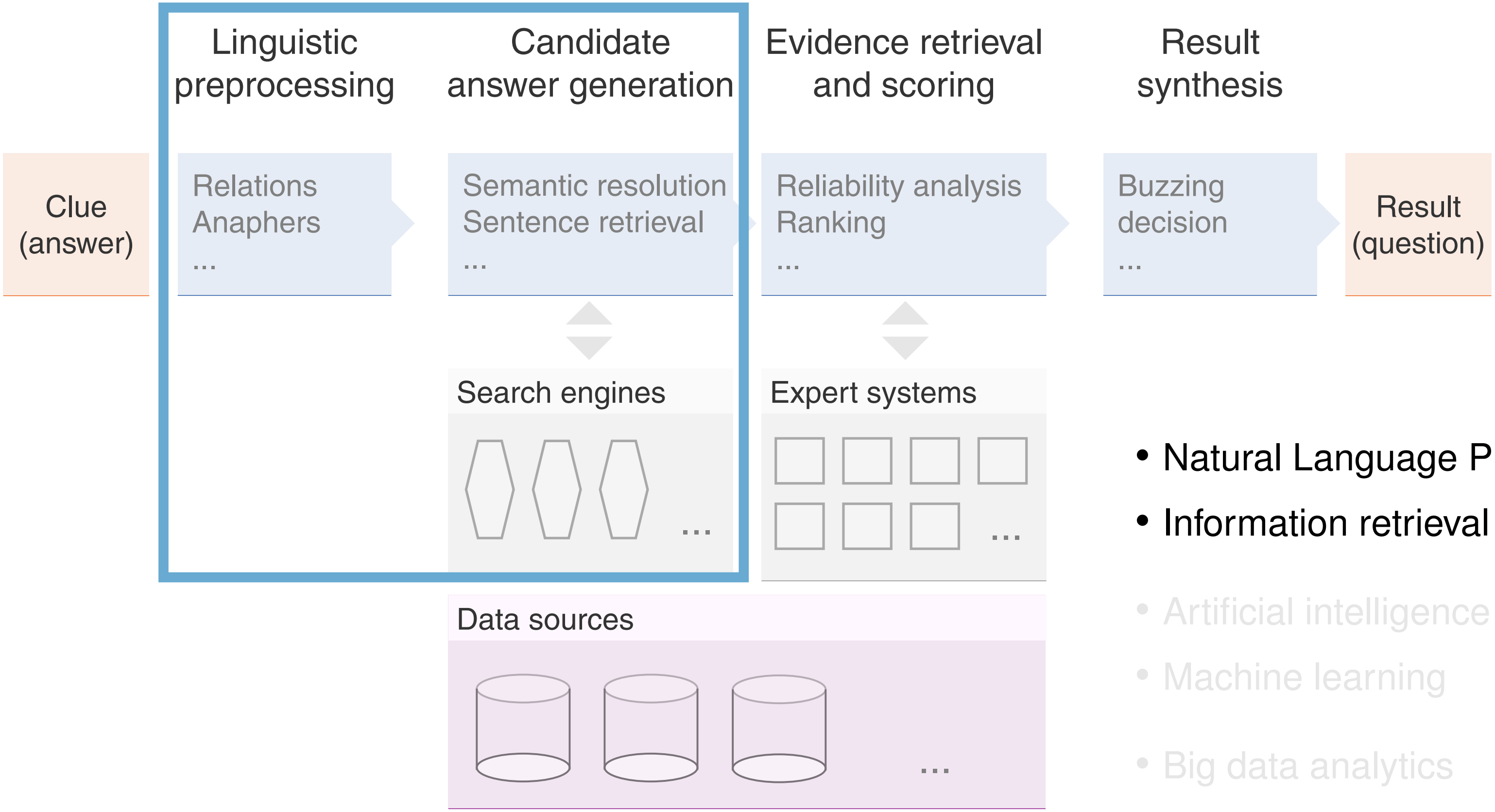
Question Answering: IBM Watson at Jeopardy (continued)



- Natural Language Processing
- Information retrieval
- Artificial intelligence
- Machine learning
- Big data analytics

Examples of NLP Systems

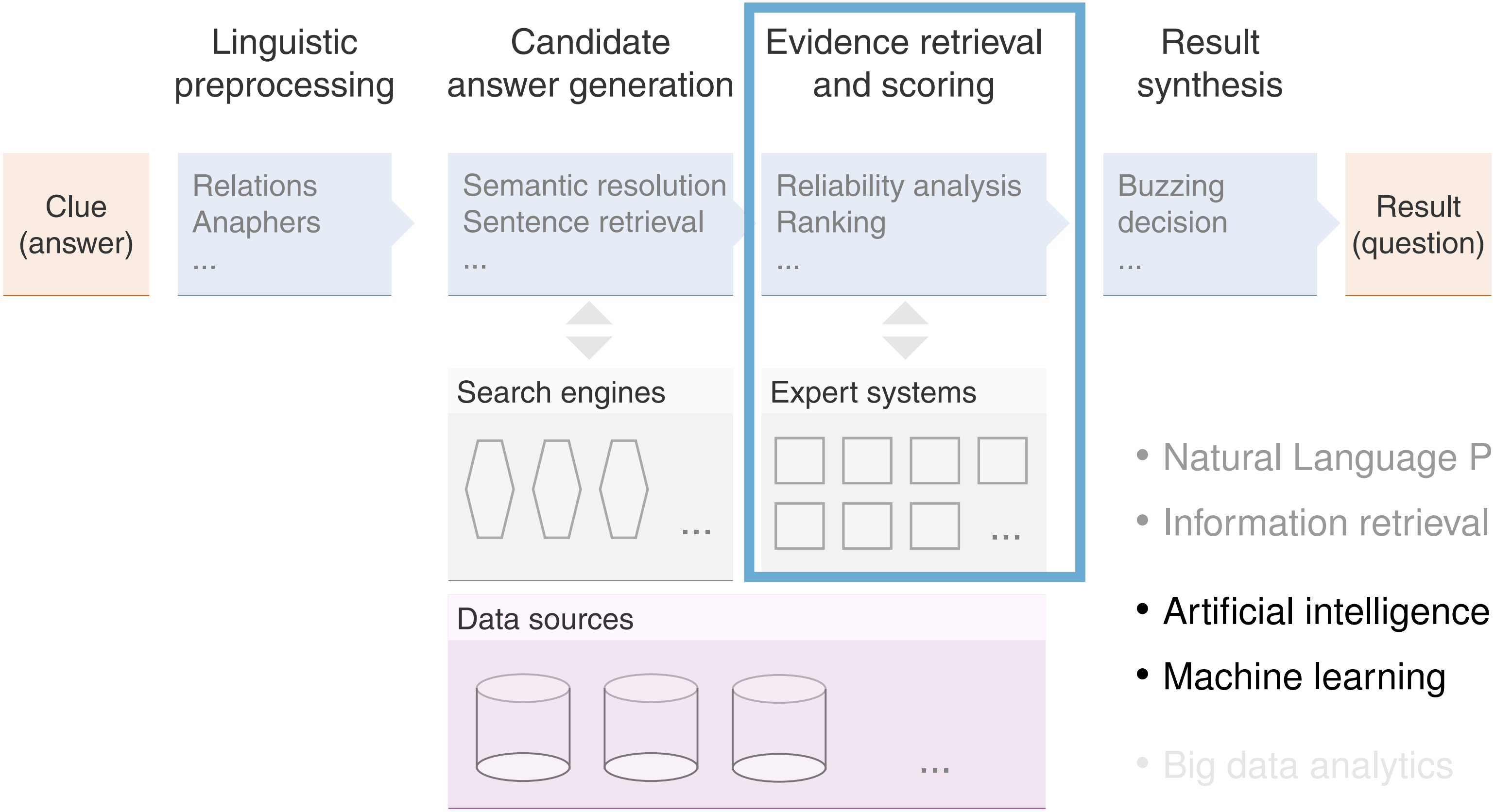
Question Answering: IBM Watson at Jeopardy (continued)



- Natural Language Processing
- Information retrieval
- Artificial intelligence
- Machine learning
- Big data analytics

Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy (continued)



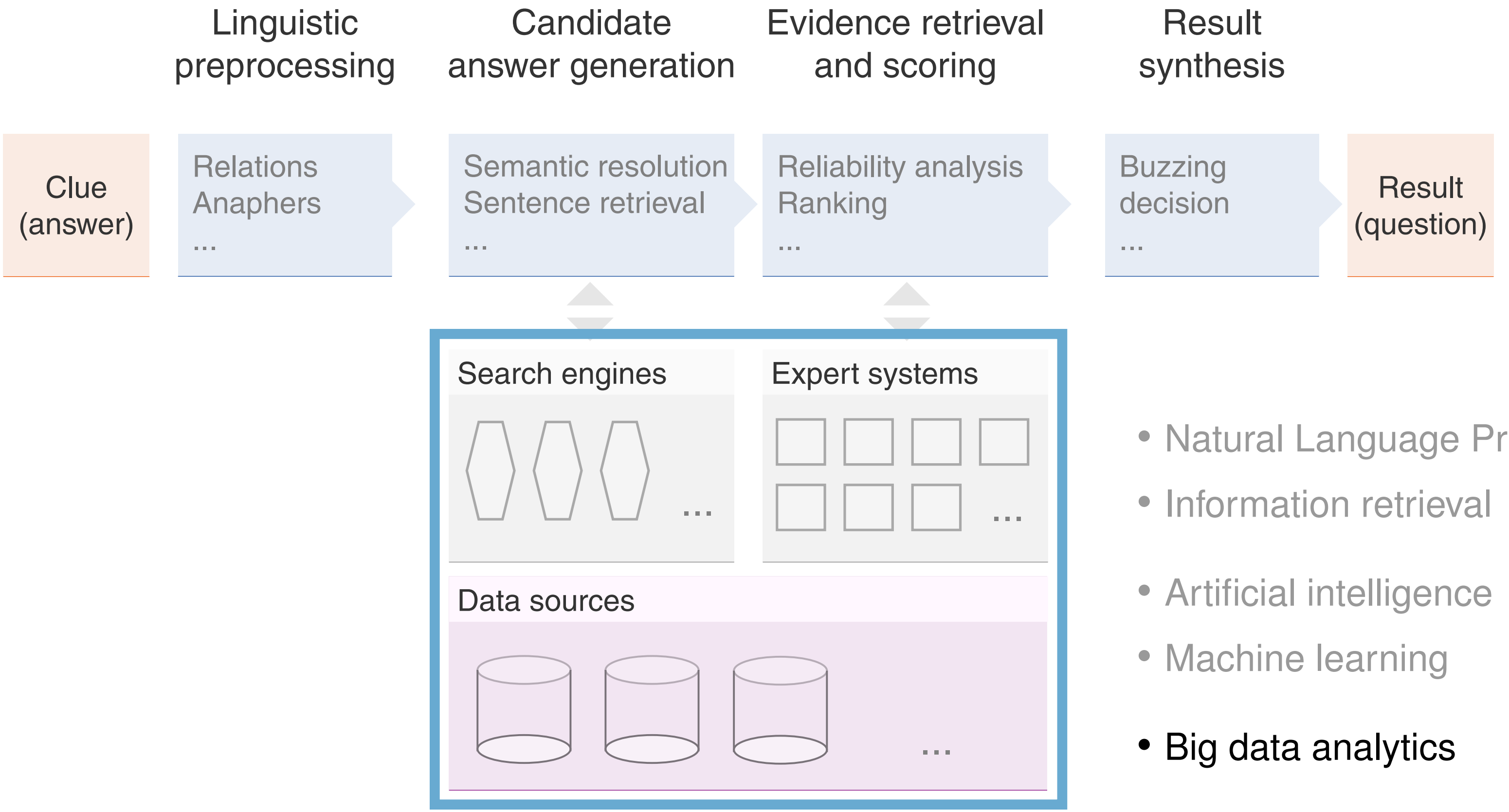
- Natural Language Processing
- Information retrieval

- Artificial intelligence
- Machine learning

- Big data analytics

Examples of NLP Systems

Question Answering: IBM Watson at Jeopardy (continued)



- Natural Language Processing
- Information retrieval
- Artificial intelligence
- Machine learning
- Big data analytics

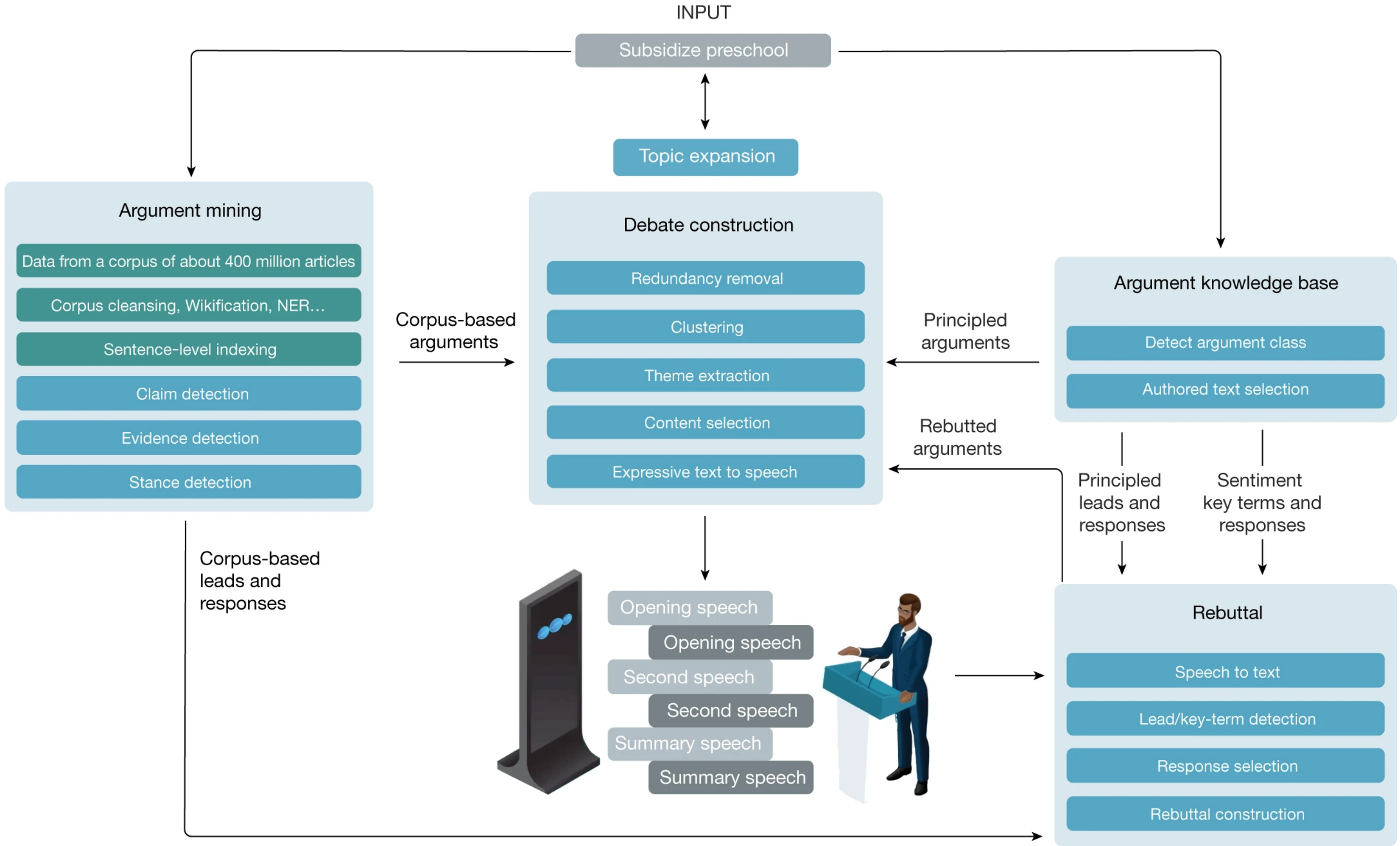
NLP Problems

IBM Debater

Debater – Uses structure from language to participate in a full live debate with expert human debaters.

NLP Problems

IBM Debater



INPUT

Subsidize preschool

Topic expansion

Argument mining

- Data from a corpus of about 400 million articles
- Corpus cleansing, Wikification, NER…
- Sentence-level indexing
- Claim detection
- Evidence detection
- Stance detection

Corpus-based arguments

Debate construction

- Redundancy removal
- Clustering
- Theme extraction
- Content selection
- Expressive text to speech

Principled arguments

Rebutted arguments

Argument knowledge base

- Detect argument class
- Authored text selection

Principled leads and responses

Sentiment key terms and responses

Corpus-based leads and responses

Opening speech

Opening speech

Second speech

Second speech

Summary speech

Summary speech

Rebuttal

- Speech to text
- Lead/key-term detection
- Response selection
- Rebuttal construction

Source: [Nature Article]

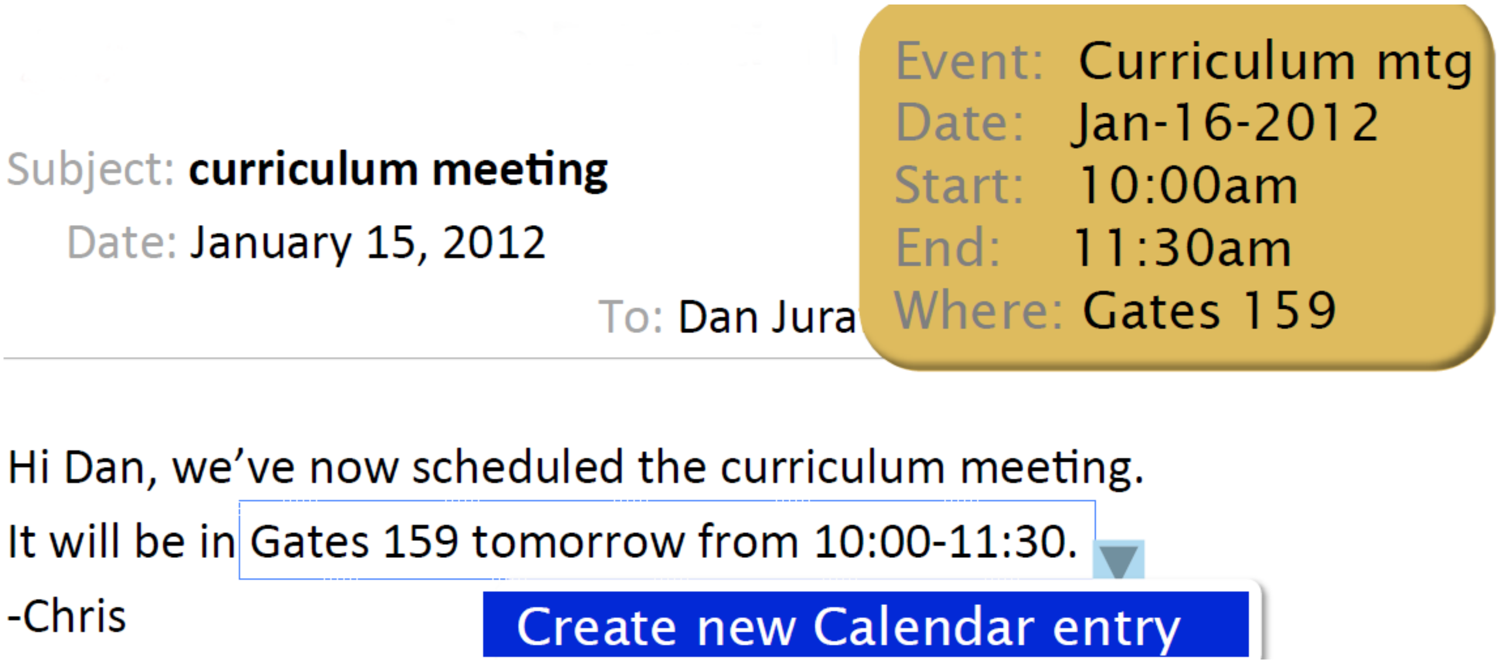
NLP Problems

Search Engines

Search Engines – Apply all sorts of NLP and Machine Learning to extract structure

NLP Problems

Information Extraction (IE)



Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.
It will be in Gates 159 tomorrow from 10:00-11:30.
-Chris

NLP Problems

Review Analysis

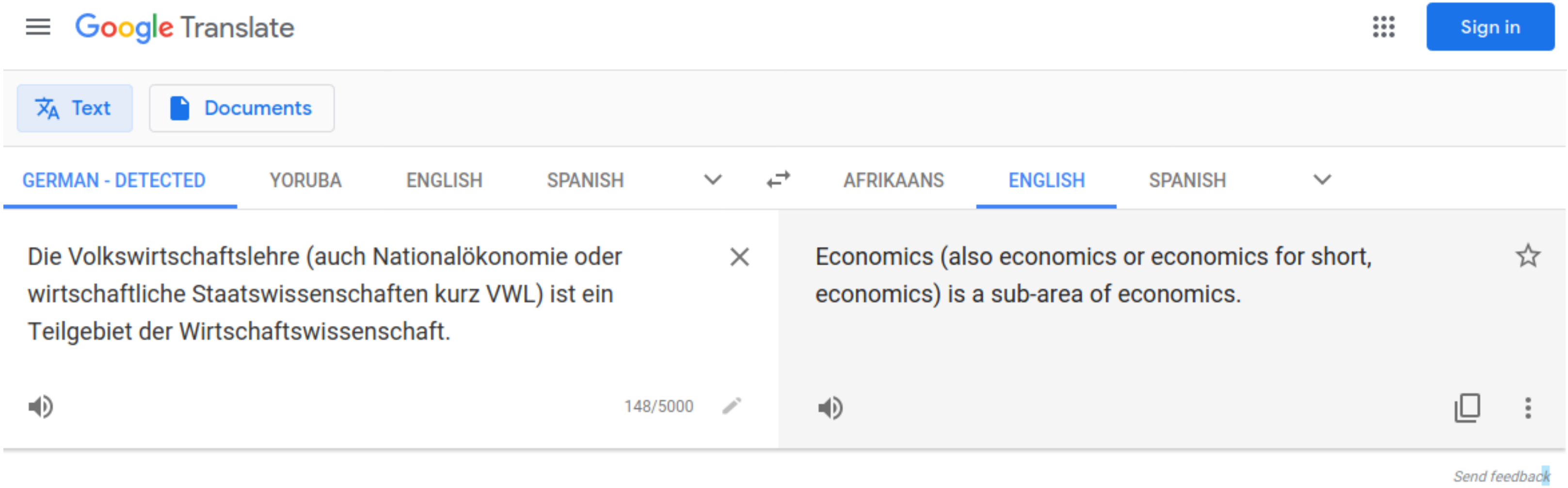
Attributes: zoom, affordability, size and weight, flash, ease of use

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

NLP Problems

Machine Translation (MT)



Die Volkswirtschaftslehre (auch Nationalökonomie oder wirtschaftliche Staatswissenschaften kurz VWL) ist ein Teilgebiet der Wirtschaftswissenschaft.

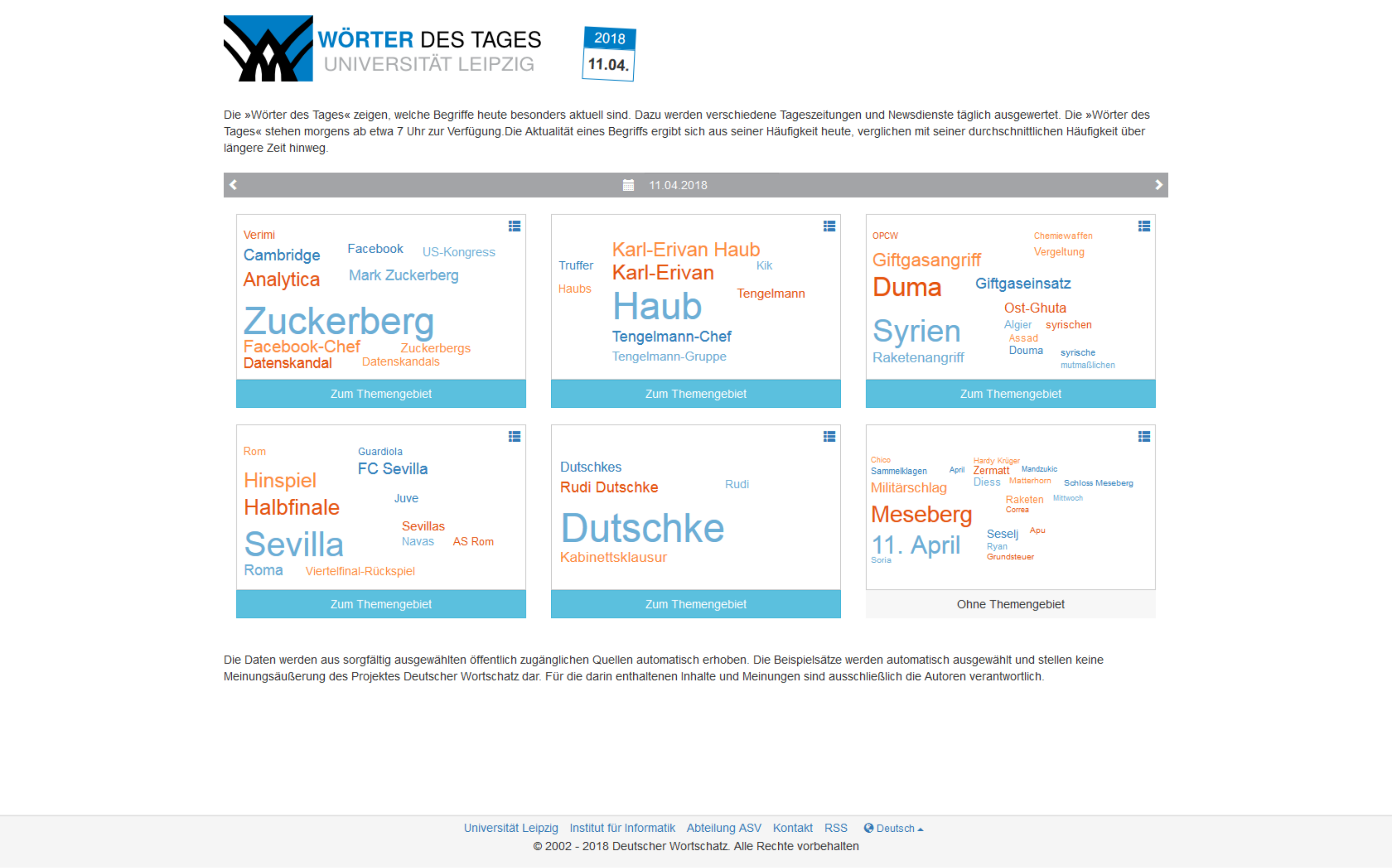
Economics (also economics or economics for short, economics) is a sub-area of economics.

First sentence of the Wikipedia article on “Volkswirtschaftslehre”.

See also twitter.com/hashtag/googletranslatefails

NLP Problems

Knowledge and Information Management



Die »Wörter des Tages« zeigen, welche Begriffe heute besonders aktuell sind. Dazu werden verschiedene Tageszeitungen und Newsdienste täglich ausgewertet. Die »Wörter des Tages« stehen morgens ab etwa 7 Uhr zur Verfügung.Die Aktualität eines Begriffs ergibt sich aus seiner Häufigkeit heute, verglichen mit seiner durchschnittlichen Häufigkeit über längere Zeit hinweg.

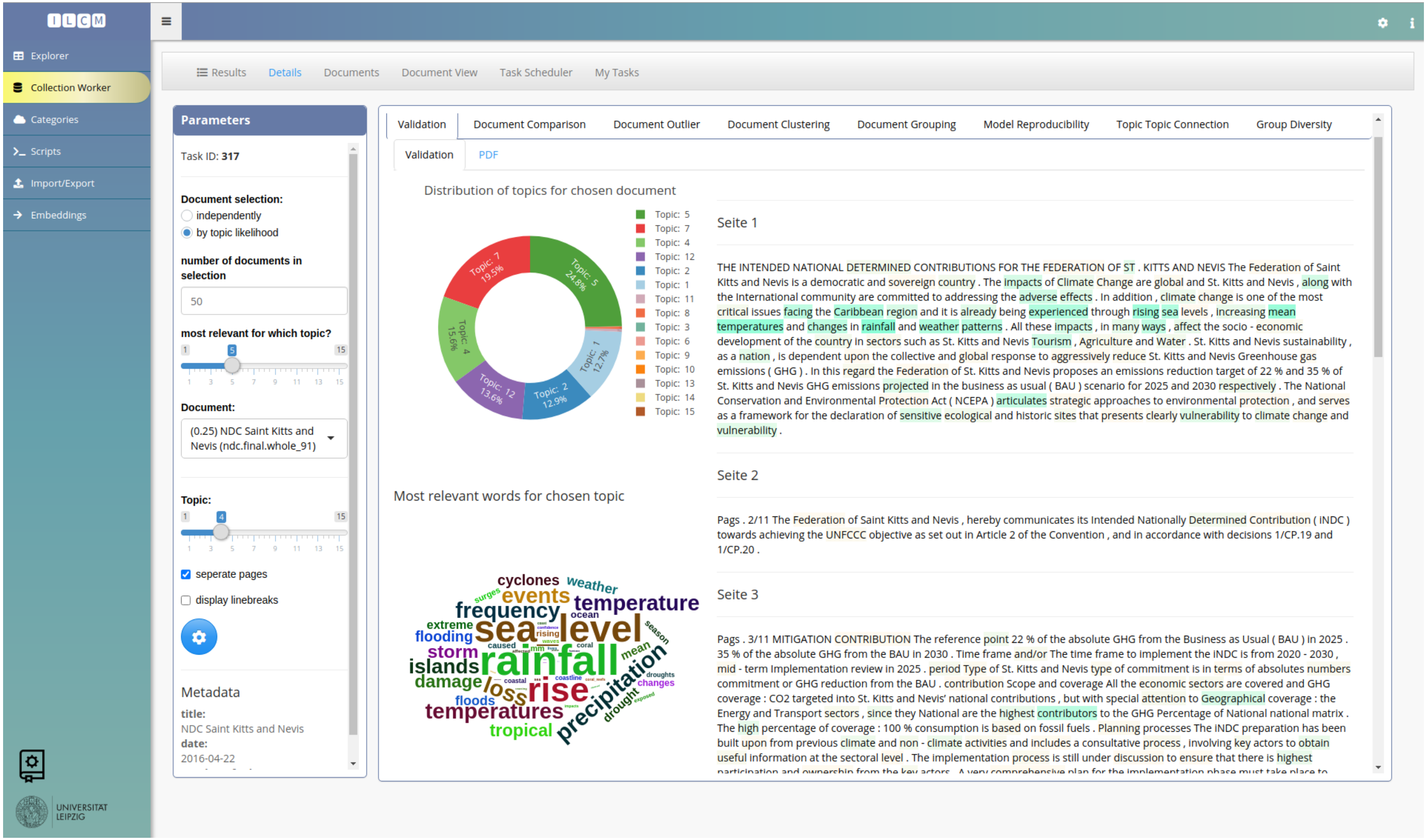
Die Daten werden aus sorgfältig ausgewählten öffentlich zugänglichen Quellen automatisch erhoben. Die Beispielsätze werden automatisch ausgewählt und stellen keine Meinungsäußerung des Projektes Deutscher Wortschatz dar. Für die darin enthaltenen Inhalte und Meinungen sind ausschließlich die Autoren verantwortlich.

Universität Leipzig Institut für Informatik Abteilung ASV Kontakt RSS Deutsch

© 2002 - 2018 Deutscher Wortschatz. Alle Rechte vorbehalten

NLP Problems

Knowledge and Information Management



Parameters

Task ID: **317**

Document selection:
○ independently
● by topic likelihood

number of documents in selection
50

most relevant for which topic?
5

Document:
(0.25) NDC Saint Kitts and Nevis (ndc.final.whole_91)

Topic:
4

☑ seperate pages
□ display linebreaks

Metadata

title:
NDC Saint Kitts and Nevis
date:
2016-04-22

Distribution of topics for chosen document

Most relevant words for chosen topic

Seite 1

THE INTENDED NATIONAL DETERMINED CONTRIBUTIONS FOR THE FEDERATION OF ST . KITTS AND NEVIS The Federation of Saint Kitts and Nevis is a democratic and sovereign country . The impacts of Climate Change are global and St. Kitts and Nevis , along with the International community are committed to addressing the adverse effects . In addition , climate change is one of the most critical issues facing the Caribbean region and it is already being experienced through rising sea levels , increasing mean temperatures and changes in rainfall and weather patterns . All these impacts , in many ways , affect the socio - economic development of the country in sectors such as St. Kitts and Nevis Tourism , Agriculture and Water . St. Kitts and Nevis sustainability , as a nation , is dependent upon the collective and global response to aggressively reduce St. Kitts and Nevis Greenhouse gas emissions (GHG) . In this regard the Federation of St. Kitts and Nevis proposes an emissions reduction target of 22 % and 35 % of St. Kitts and Nevis GHG emissions projected in the business as usual (BAU) scenario for 2025 and 2030 respectively . The National Conservation and Environmental Protection Act (NCEPA) articulates strategic approaches to environmental protection , and serves as a framework for the declaration of sensitive ecological and historic sites that presents clearly vulnerability to climate change and vulnerability .

Seite 2

Pags . 2/11 The Federation of Saint Kitts and Nevis , hereby communicates its Intended Nationally Determined Contribution (INDC) towards achieving the UNFCCC objective as set out in Article 2 of the Convention , and in accordance with decisions 1/CP.19 and 1/CP.20 .

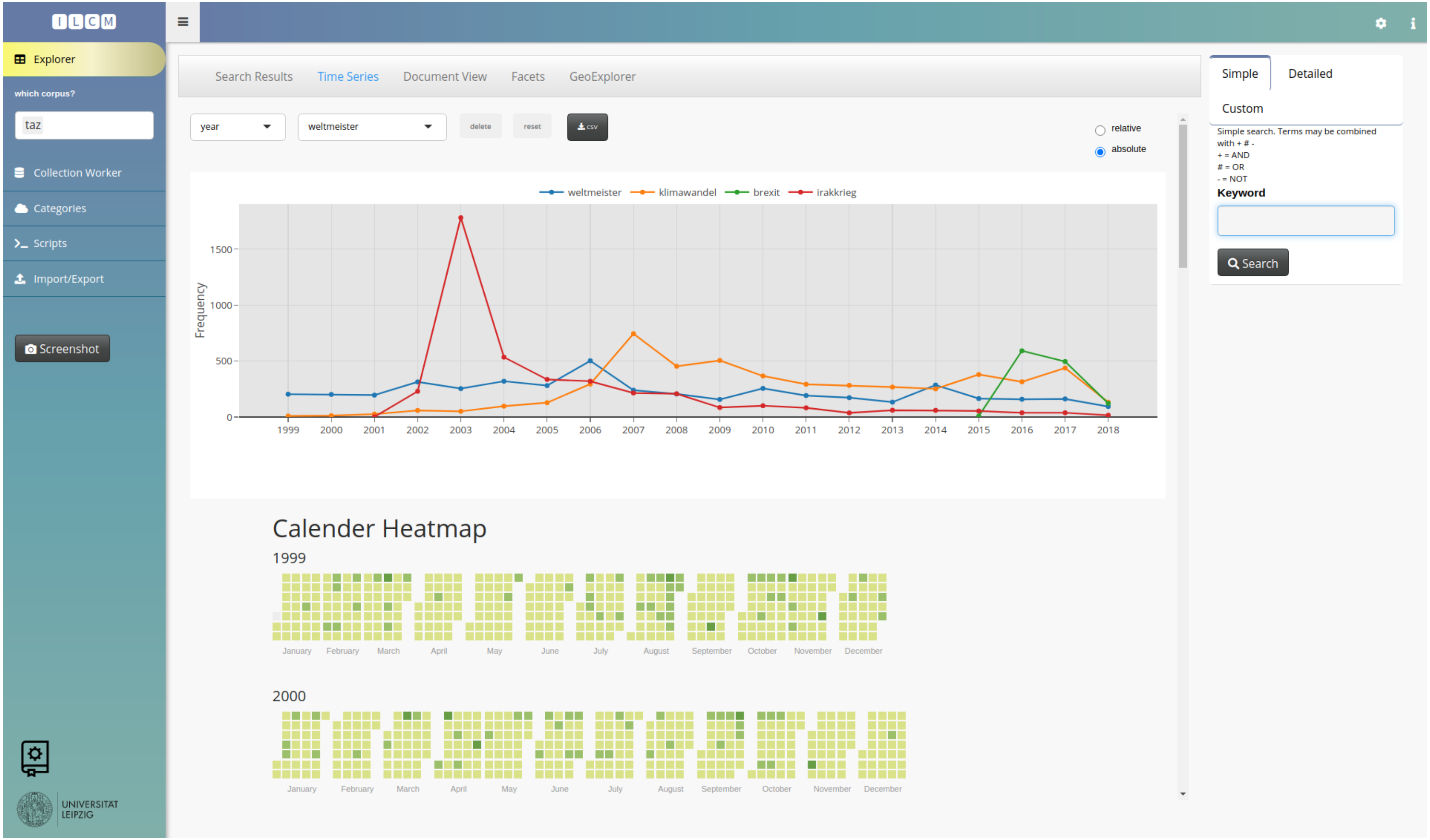
Seite 3

Pags . 3/11 MITIGATION CONTRIBUTION The reference point 22 % of the absolute GHG from the Business as Usual (BAU) in 2025 . 35 % of the absolute GHG from the BAU in 2030 . Time frame and/or The time frame to implement the iNDC is from 2020 - 2030 , mid - term Implementation review in 2025 . period Type of St. Kitts and Nevis type of commitment is in terms of absolutes numbers commitment or GHG reduction from the BAU . contribution Scope and coverage All the economic sectors are covered and GHG coverage : CO2 targeted into St. Kitts and Nevis' national contributions , but with special attention to Geographical coverage : the Energy and Transport sectors , since they National are the highest contributors to the GHG Percentage of National national matrix . The high percentage of coverage : 100 % consumption is based on fossil fuels . Planning processes The iNDC preparation has been built upon from previous climate and non - climate activities and includes a consultative process , involving key actors to obtain useful information at the sectoral level . The implementation process is still under discussion to ensure that there is highest participation and ownership from the key actors . A very comprehensive plan for the implementation phase must take place to

https://ilcm.informatik.uni-leipzig.de/ [Niekler et. al.]

NLP Problems

Knowledge and Information Management



which corpus? taz

year ▾ weltmeister ▾

○ relative
● absolute

Simple search. Terms may be combined with + # -
+ = AND
= OR
- = NOT

Keyword

https://ilcm.informatik.uni-leipzig.de/ [Niekler et. al.]

NLP Problems

State of Affairs: Mostly Solved

- Spam detection.

  Let’s go to Agra vs. Buy V1Agra

- Part-of-speech (POS) tagging.

  Colorless/Adjective green/Adjective ideas/Noun sleep/Verb furiously/Adverb.

- Named entity recognition (NER).

  Einstein:Person met with UN:Organization officials in Princeton:Location.

NLP Problems

State of Affairs: Making Good Progress

- Sentiment detection.

  Best pizza in town. vs. The waiter ignored us for 20 minutes.

- Coreference resolution.

  ?My trophy did not fit into ?the suitcase because it is too big.

- Word sense disambiguation (WSD)

  I need new batteries for my mouse.

NLP Problems

State of Affairs: Making Good Progress (continued)

- Machine translation.

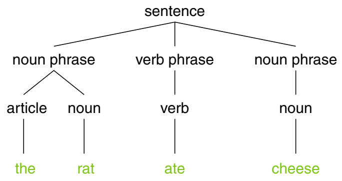
  Is getting better and better. → Wird immer besser.

- Information extraction.

  Come to our first lecture, April 15. → Calendar update: Lecture (April 15)

- Parsing.

  The rat ate cheese. →



NLP Problems

State of Affairs: Still Challenging

- ❑ Question answering (QA).
  Is ibuprofen effective in reducing fever for patients with acute febrile illness?

- ❑ Paraphrasing.
  XYZ acquired ABC yesterday vs. ABC has been taken over by XYZ

- ❑ Summarization.
  Dow Jones is up + house prices rose $\rightarrow$ Economy is good

- ❑ Dialogue.
  User: Best pizza around?
  Echo/Siri/Now: Antonio's. Want a table tonight?

Remarks:

- On referring to the field (roughly):

  1. Natural Language Processing/Language Engineering. Devising methods for processing specific language phenomena (e.g. resolving pronouns); operationalizing formal models of language (e.g. computational formal grammars)
  2. Language Technology/Text Technology/Speech Technology. Applications of NLP (various sub-areas: MT, Dialogue Systems, etc.)
  3. Computational Linguistics. Linguistics/Language science research using computational means

  Unfortunately, these terms are often used interchangeably.

- For an overview of history of NLP see, for example, Karen Sparck Jones (1994) Natural Language Processing: A Historical Review

- Food for thought. 2019 IBM Project Debater held its first public live debate with Harish Natarajan who holds the world record for most debate competitions won; the event can be viewed here. Watch (parts of) the debate and then go back to the schema of Watson's architecture.
  - What kind of functionalities/functional components do you think are required for such a system?
  - Can you decompose the debating task into components, some of which require NLP?