Detecting Web Functions

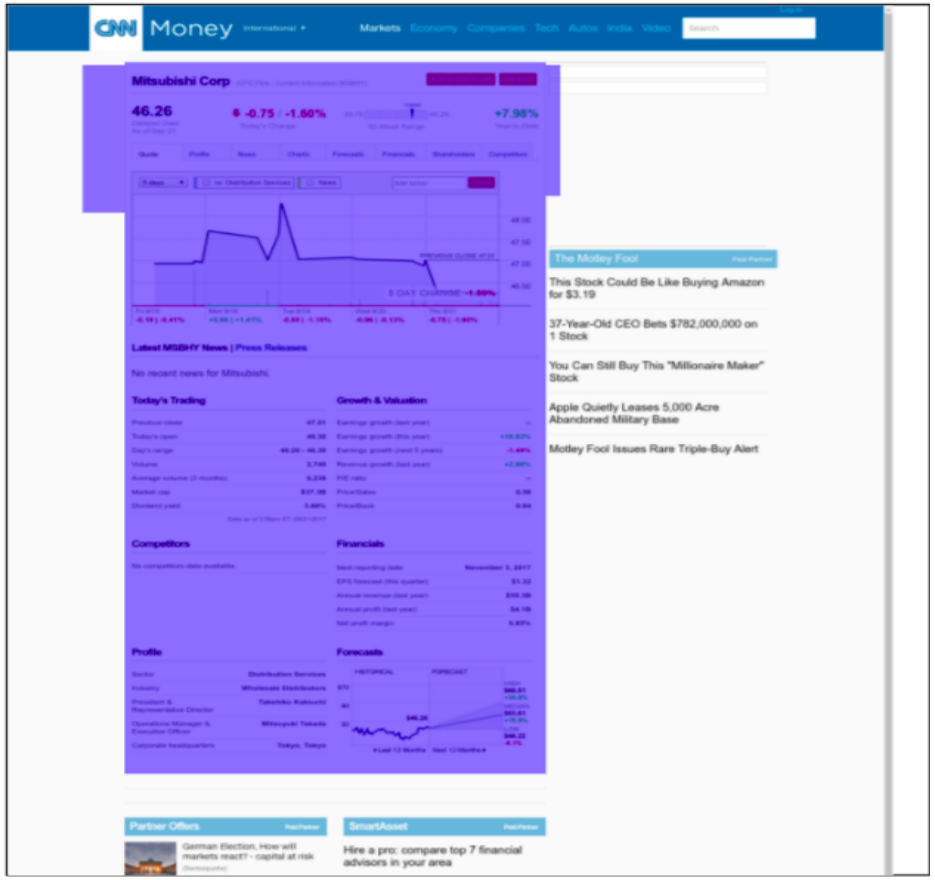
Bauhaus University Weimar
Ademola Eric Adewumi
Master's Thesis
12.03.2021

Consistent Identification of Sections

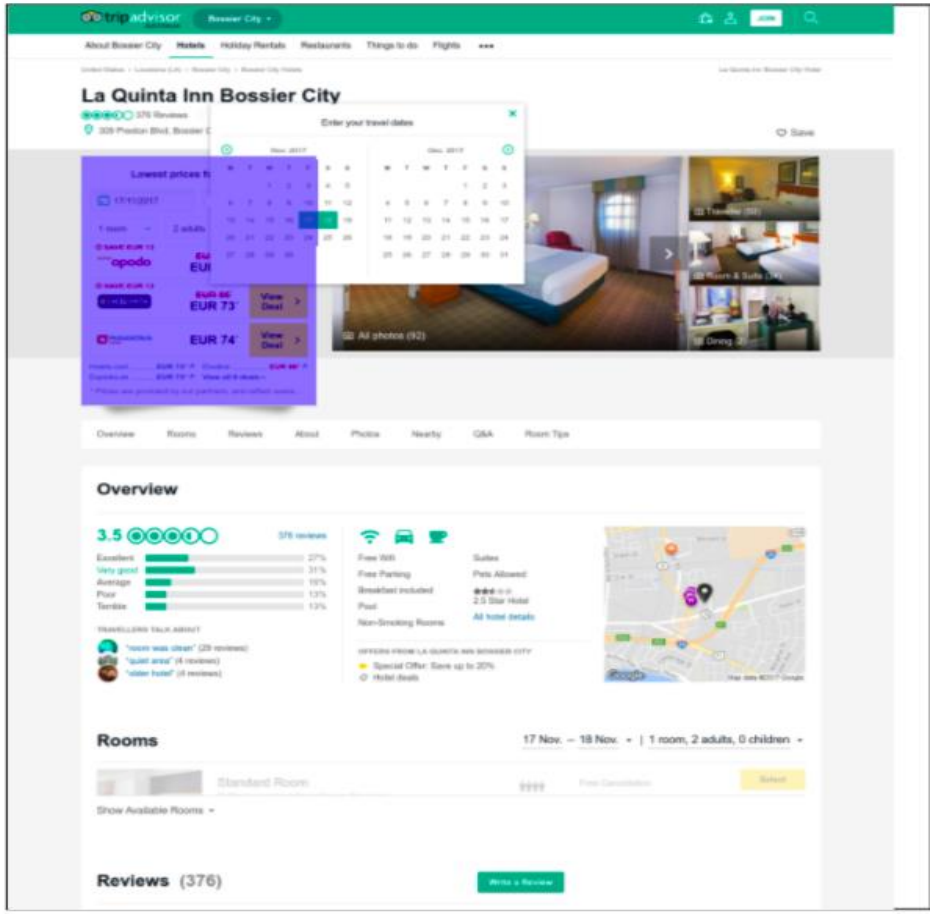
Hypertext Markup displayed in the World Wide Web is a Webpage.

Document Object Model elements make Sections that show Patterns of Consistent Identification of Visual Features for web users

Task 3



Task 1



Genre Identification of Webpages

- Chat Forums
- Links
- Forms
- Shopping Cart
- Videos
- Notifications

This web page wants	mainly	also	not
to allow the visitor to discuss with others	○	◉	○
to suggest pages to the visitor (ignore ads!)	○	◉	○
to get information from the visitor	○	◉	○
to sell to or buy from the visitor (ignore ads!)	◉	○	○
to entertain the visitor	○	◉	○
to inform the visitor	○	◉	○

Multi Annotator Competence

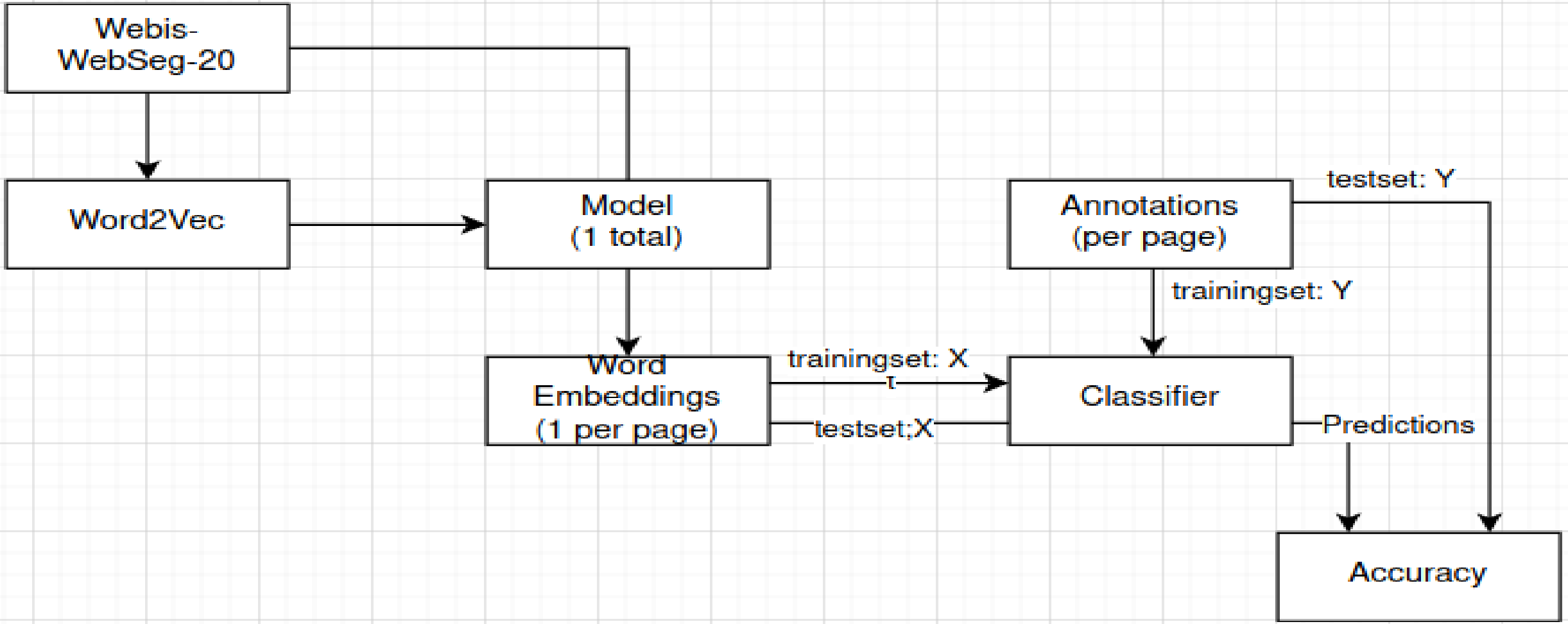
- 439 Workers
- > 9000 pages annotated
- Weight of performing annotators is +positive
- Weight of non-performing annotators –negative

Workers

	Name ↓	Assignments	Approved ↑	Rejected
☐	AZXM77IPUDQOS	7	7	0
☐	AZOO712WKGS0I	1	1	0
☐	AYZNYPT2TVDWF	1	1	0
☐	AYSTMCRE2AE7T	2	2	0
☐	AYJ2Z50W4IN8V	2	2	0
☐	AYHIH9NTPYFLY	179	179	0
☐	AYGOIYMWWWGF2	1	1	0
☐	AYF300N0NPCMU	1	1	0
☐	AY5ZTLIRK9IOS	1	1	0

Worker	Sec.
A18..E9F (0.66)	178
A3S..3LZ (0.68)	217
A34..EAH (0.70)	139
AUK..VEQ (0.25)	330
AI8..44H (0.52)	561

Random Forest Classifier



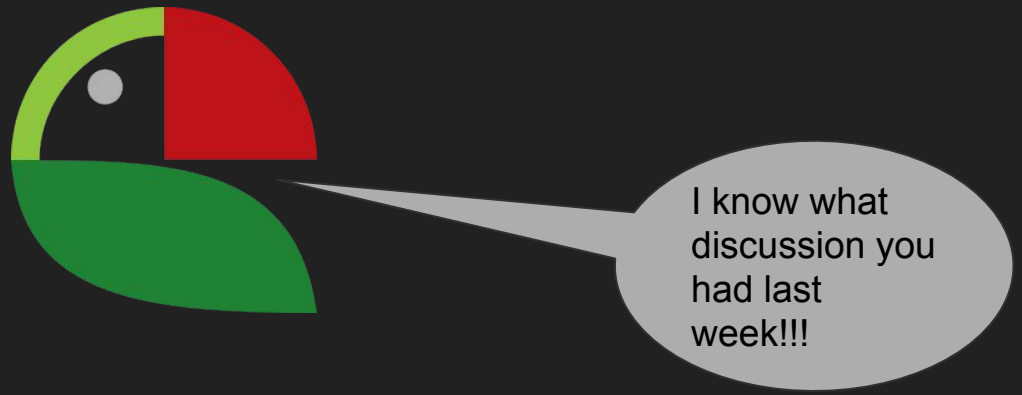
Words, Vectors, Embeddings

```
['bildbeschreibung', 'einblenden']
['verbindlich', 'im', 'ton', 'klar', 'im', 'ziel', 'stephan', 'weil']
['bild', 'l', 'decke', 'matthias']
['spd', 'spitzenkandidat', 'weil', 'muss', 'seinen', 'dienstsitz', 'nur', 'ein', 'paar', 'stra', 'enz',
[]
['ass', 'die', 'spd', 'es', 'mit', 'den', 'gr', 'nen', 'mit', 'denen', 'sie', 'niedersachsen', 'regiere
['stephan', 'weil']
['sitzungen', 'und', 'verhandlungen', 'mit', 'den', 'gr', 'nen', 'im', 'stadtrat', 'von', 'hannover', '
['seinen', 'dienstsitz', 'muss', 'der', 'bisherige', 'oberb', 'rgermeister', 'von', 'hannover', 'vom',
['zwei', 'erfahrungswelten']
['der', 'k', 'nftige', 'ministerpr', 'sident', 'st', 'tzt', 'sich', 'auf', 'zwei', 'erfahrungswelten',
['hannover', 'ist', 'aus', 'tief', 'verankerter', 'tradition', 'sozialdemokratisch', 'ist', 'es', 'wahr
['ayg', 'l', 'zkan']
['aufstellt', 'frau', 'zkan', 'k', 'nnte', 'insbesondere', 'frauen', 'und', 'zuwanderer', 'als', 'w', '
['freundlicher', 'umgang', 'mit', 'mcallister']
['lange', 'hatte', 'der', 'jahre', 'alte', 'weil', 'den', 'ruf', 'eines', 'spr', 'den', 'gespr', 'chspa
['ohne', 'emotionen', 'ist', 'weil', 'der', 'viel', 'liest', 'und', 'seine', 'ruhe', 'bei', 'langl', 'u
['mehr', 'zum', 'thema']
```

```
[-0.39892951  0.94143304  2.85872626  2.95971623  4.65714902  5.88993579
  6.65095437  7.1860415   7.25214559  9.32081875  9.65896907 11.13717788
 12.83100027 13.28471884 14.22987399 15.78768975 16.97390151 16.92146759
 17.28807372 19.29048359 20.19074239 20.41795033 21.75895694 22.57040146
 22.90343666 25.1868642  25.23326331 26.27327681 27.92765573 29.00115197
 29.4320538  30.35311526]
```

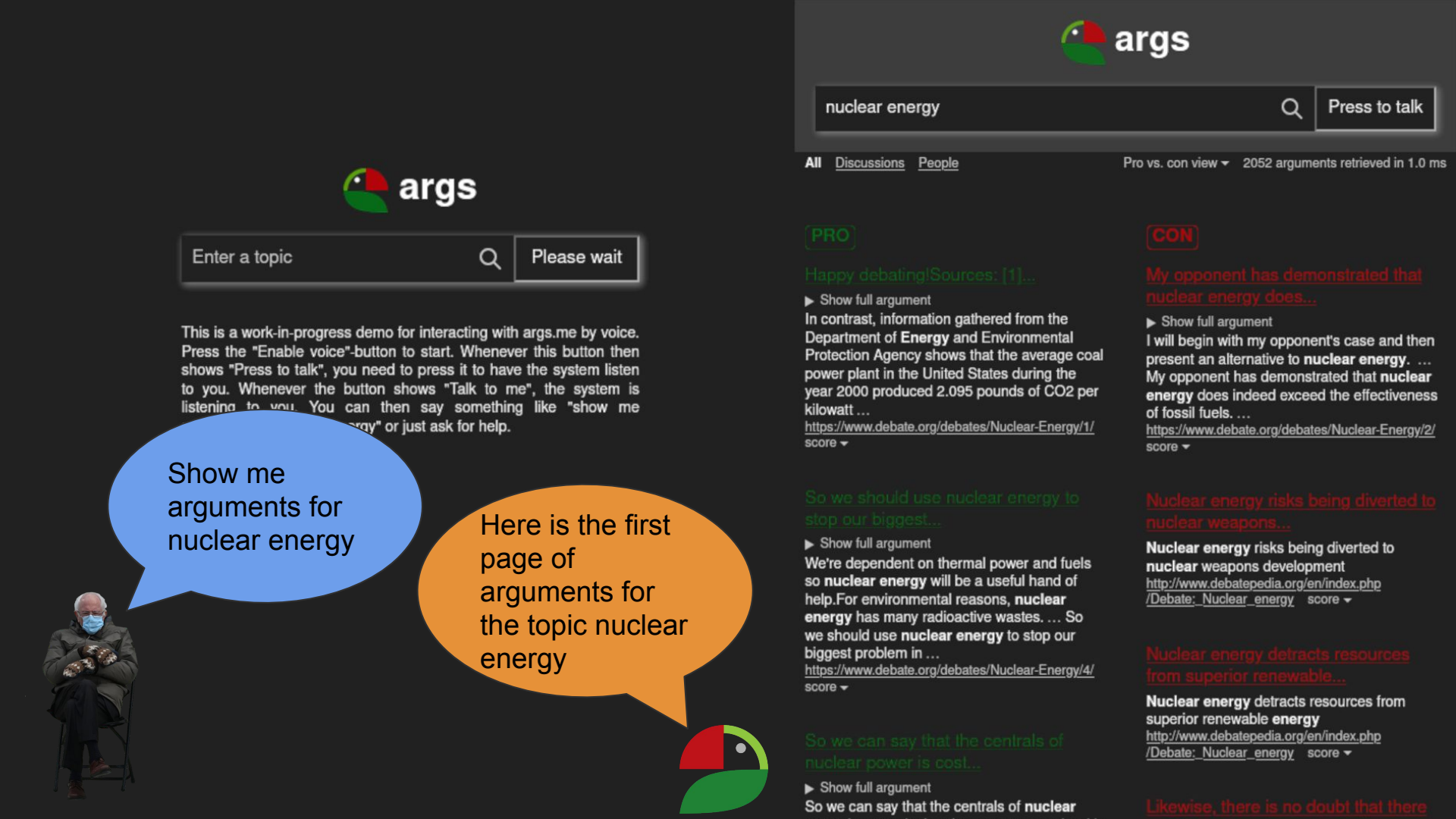
```
Supply:10
Company:109
in:4609
association:10
with:1787
BillDesk:5
brings:11
to:7700
you:2598
simple,:9
convenient:12
secure:22
way:153
pay:89
your:2003
```

THE END



Conversational Argument Search

A Creepy Search Engine?


args
Enter a topic
Please wait
This is a work-in-progress demo for interacting with args.me by voice. Press the "Enable voice"-button to start. Whenever this button then shows "Press to talk", you need to press it to have the system listen to you. Whenever the button shows "Talk to me", the system is listening to you. You can then say something like "show me
or just ask for help.
Show me arguments for nuclear energy
Here is the first page of arguments for the topic nuclear energy
args
nuclear energy
Press to talk
All
Discussions
People
Pro vs. con view
2052 arguments retrieved in 1.0 ms
PRO
Happy debating!Sources: [1]...
Show full argument
In contrast, information gathered from the Department of Energy and Environmental Protection Agency shows that the average coal power plant in the United States during the year 2000 produced 2.095 pounds of CO2 per kilowatt ...
https://www.debate.org/debates/Nuclear-Energy/1/
score
So we should use nuclear energy to stop our biggest...
Show full argument
We're dependent on thermal power and fuels so nuclear energy will be a useful hand of help.For environmental reasons, nuclear energy has many radioactive wastes. ... So we should use nuclear energy to stop our biggest problem in ...
https://www.debate.org/debates/Nuclear-Energy/4/
score
So we can say that the centrals of nuclear power is cost...
Show full argument
So we can say that the centrals of nuclear
CON
My opponent has demonstrated that nuclear energy does...
Show full argument
I will begin with my opponent's case and then present an alternative to nuclear energy. ... My opponent has demonstrated that nuclear energy does indeed exceed the effectiveness of fossil fuels. ...
https://www.debate.org/debates/Nuclear-Energy/2/
score
Nuclear energy risks being diverted to nuclear weapons...
Nuclear energy risks being diverted to nuclear weapons development
http://www.debatepedia.org/en/index.php/Debate:_Nuclear_energy
score
Nuclear energy detracts resources from superior renewable...
Nuclear energy detracts resources from superior renewable energy
http://www.debatepedia.org/en/index.php/Debate:_Nuclear_energy
score
Likewise, there is no doubt that there

What we are working on!

- Annotating Arguments with Wikipedia Links - WikiLinks
- Improve skill development and deployment experience

WIP:

- Research studies on how individuals and groups interact in arguments
- (meta) data collection and usage with RDF & SPARQL

A Convo with an Argument Search Engine

User:

Hey. I was wondering are cars with internal combustion engines better than electric cars?

A Convo with an Argument Search Engine

https://9gag.com/gag/ajqeR4R

ARGS:

Sorry, I didn’t get that.

Did you mean

electric cats?

A Convo with an Argument Search Engine

User:

No. I want to inform myself about the disadvantages of electric cars.

https://tecfa.unige.ch/guides/svg/ex/html5/svg-import/like_a_sir_original.svg

A Convo with an Argument Search Engine

ARGS:

Okay. I get that you are interested in electricity. What about nuclear energy. I could tell you about the disadvantages of nuclear power plants.

A Convo with an Argument Search Engine

User:

No. Just list arguments against electric vehicles.

A Convo with an Argument Search Engine

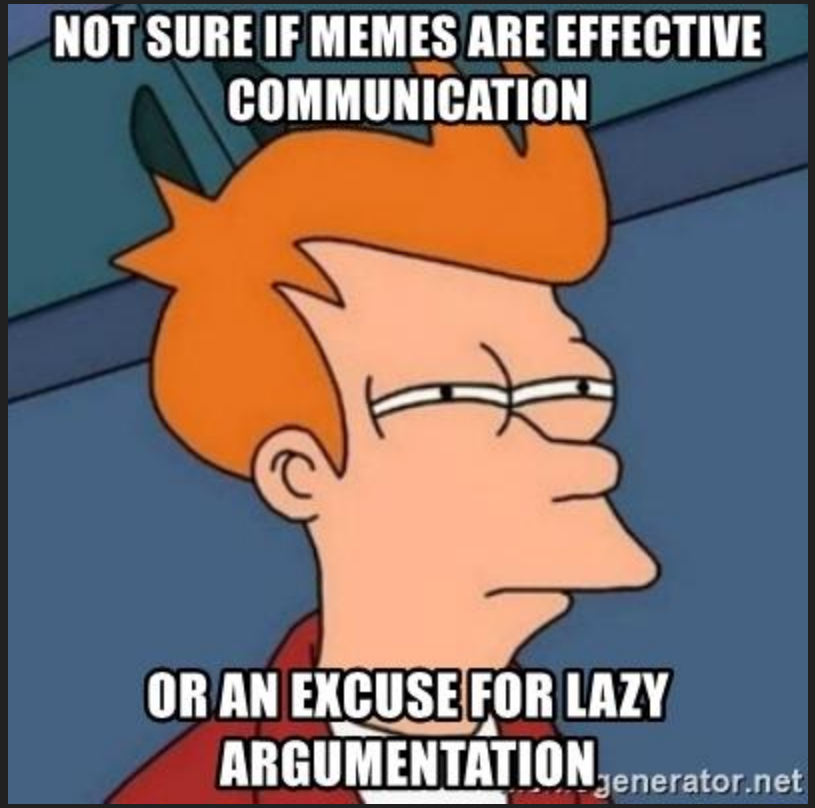
ARGS:

Okay. Based on your previous interest, here are the first three arguments against electric cats.

A Convo with an Argument Search Engine

Apache Jena

- RDF and SparQL supported
- RDF is a standard model for data interchange on the Web.
  - Evolution supported
  -


NOT SURE IF MEMES ARE EFFECTIVE COMMUNICATION
OR AN EXCUSE FOR LAZY ARGUMENTATION
generator.net

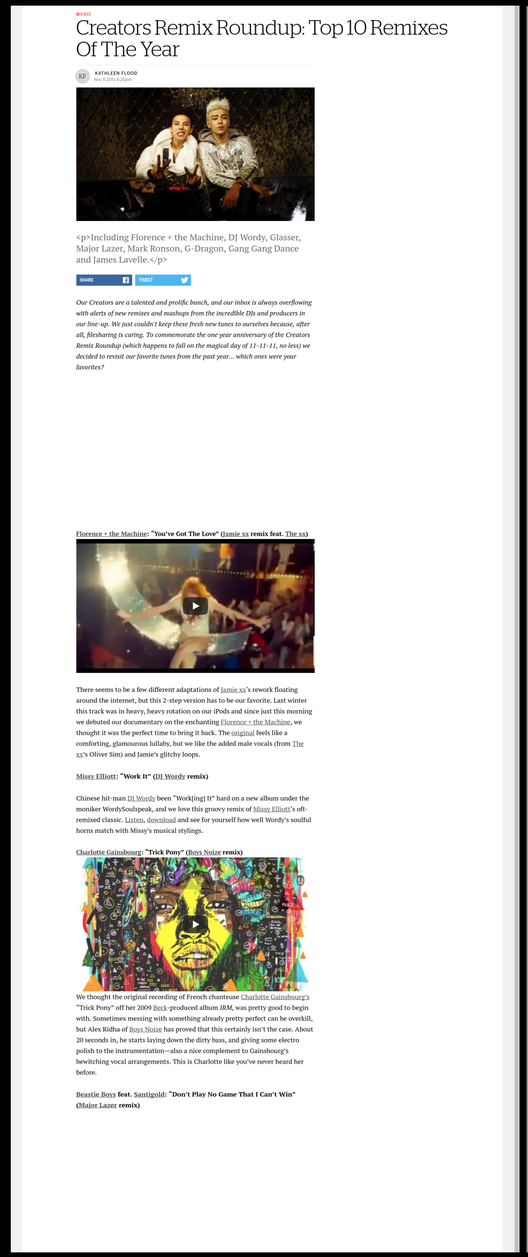
DETERMINING THE QUALITY OF ARCHIVED WEB PAGES FROM SCREENSHOT DIFFERENCES

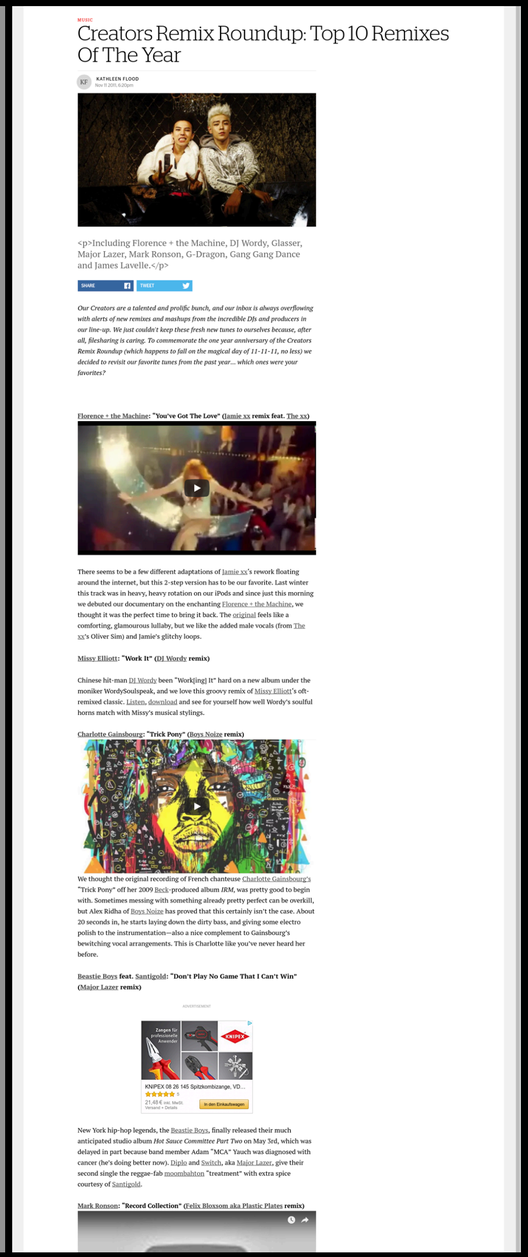
Master’s Thesis by Theresa Elstner

MOTIVATION

QUALITY ASSESSMENT

- Relative pixel error: **32 %**
- **pixel error does not adequately represent quality**





archived page

Pixel Error:

same differing

original page

BASICALLY...

REDUCE THE NOISE IN PIXEL ERROR

BASICALLY...

REDUCE THE NOISE IN PIXEL ERROR

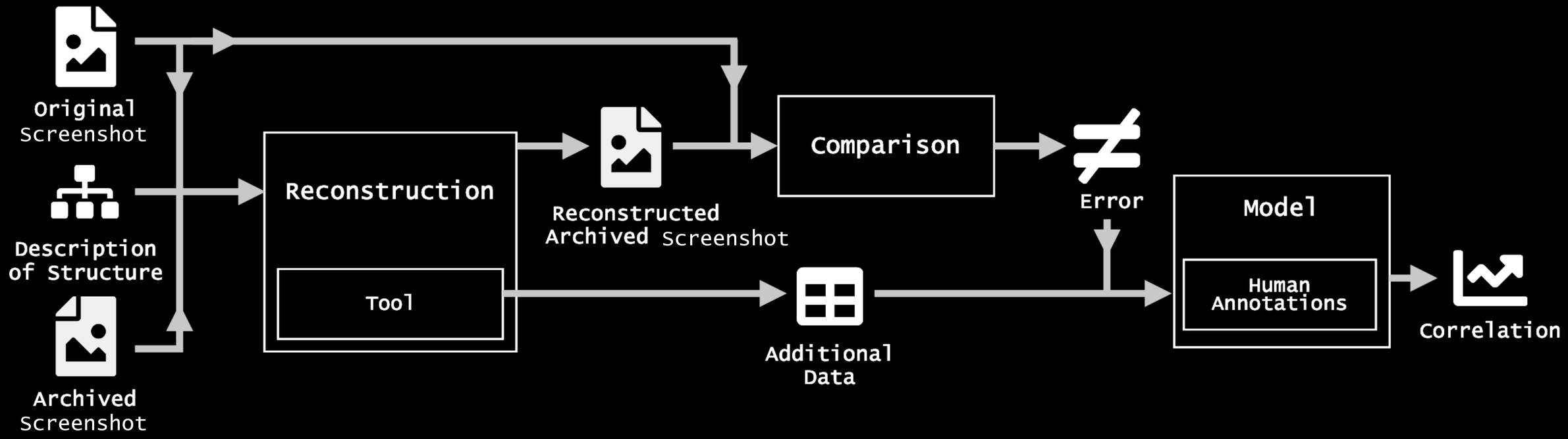
BASICALLY...

REDUCE THE NOISE IN PIXEL ERROR

OBTAIN A BETTER INDICATOR FOR QUALITY

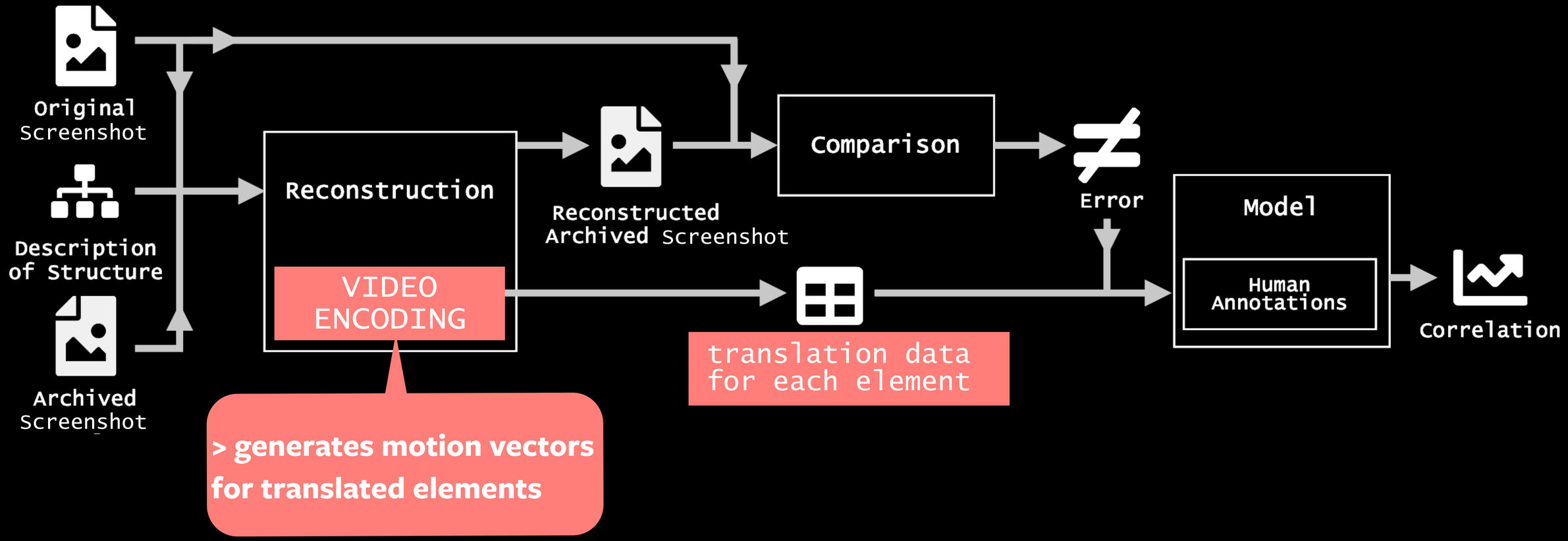
GENERAL FRAMEWORK

HOW TO DEVELOP AUTOMATIC QUALITY ASSESSMENT FOR WEB ARCHIVES



IMPLEMENTATION OF THE FRAMEWORK

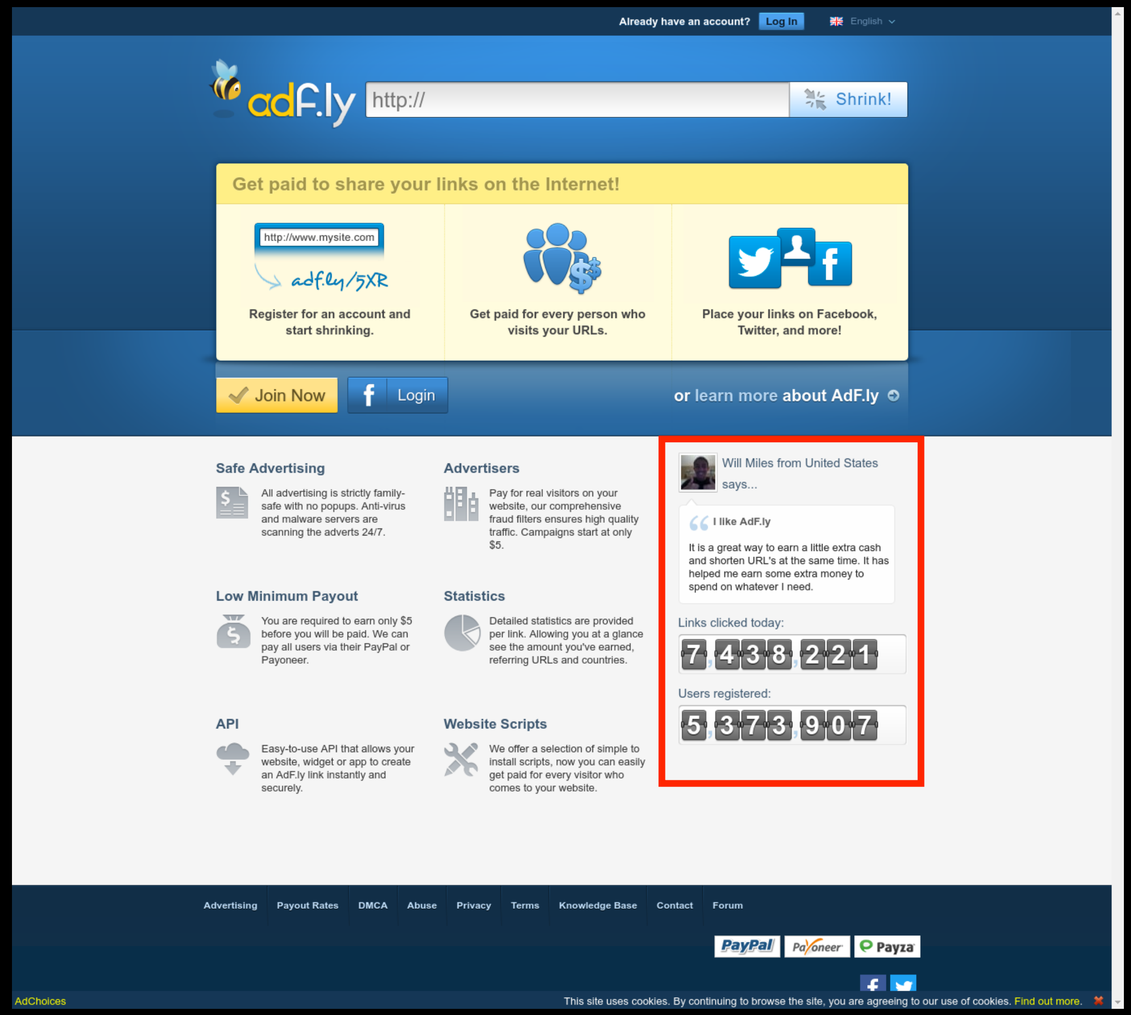
REDUCING PIXEL ERROR THROUGH ELIMINATING TRANSLATION ERROR



IMPLEMENTATION OF THE FRAMEWORK

RECONSTRUCTING ARCHIVED SCREENSHOTS

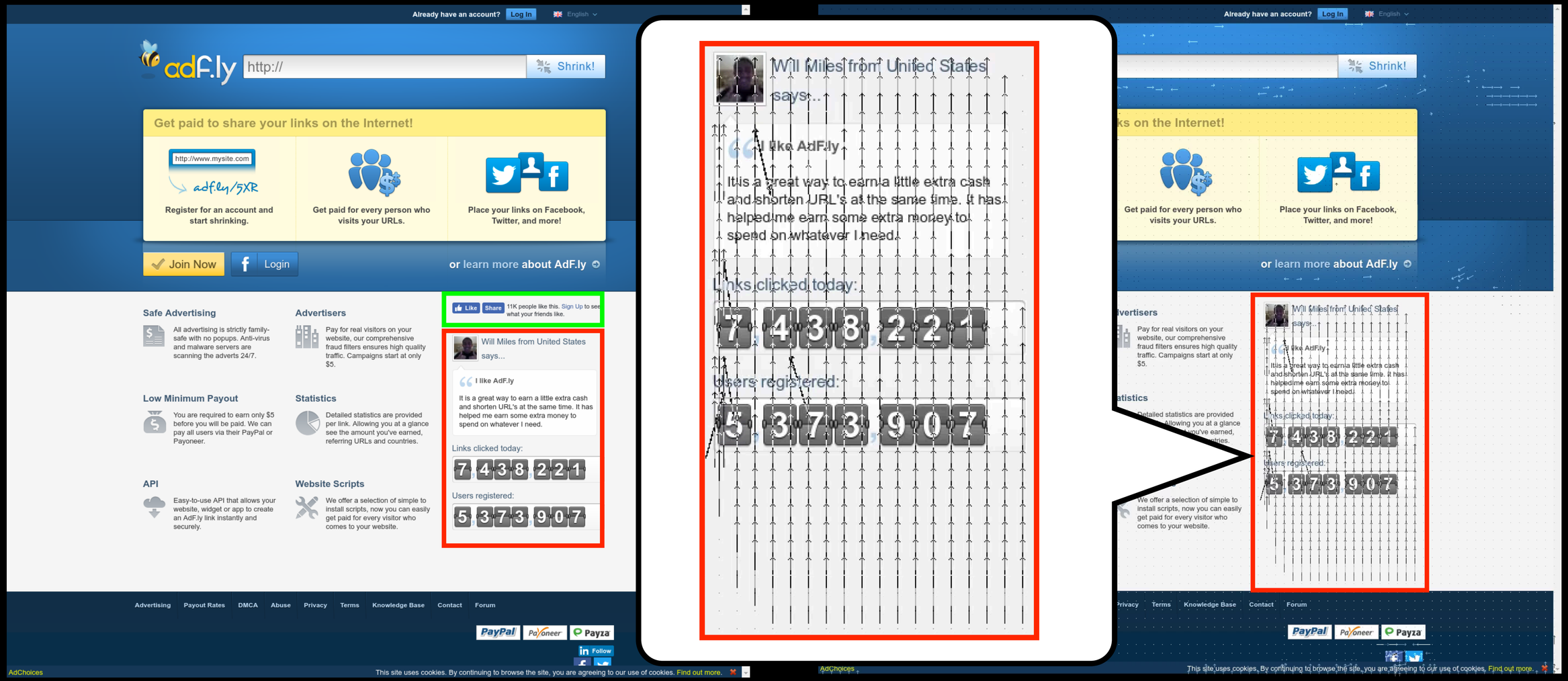
First Frame: Original Screenshot



Second Frame: Reconstructed Screenshot

IMPLEMENTATION OF THE FRAMEWORK

RECONSTRUCTING ARCHIVED SCREENSHOTS



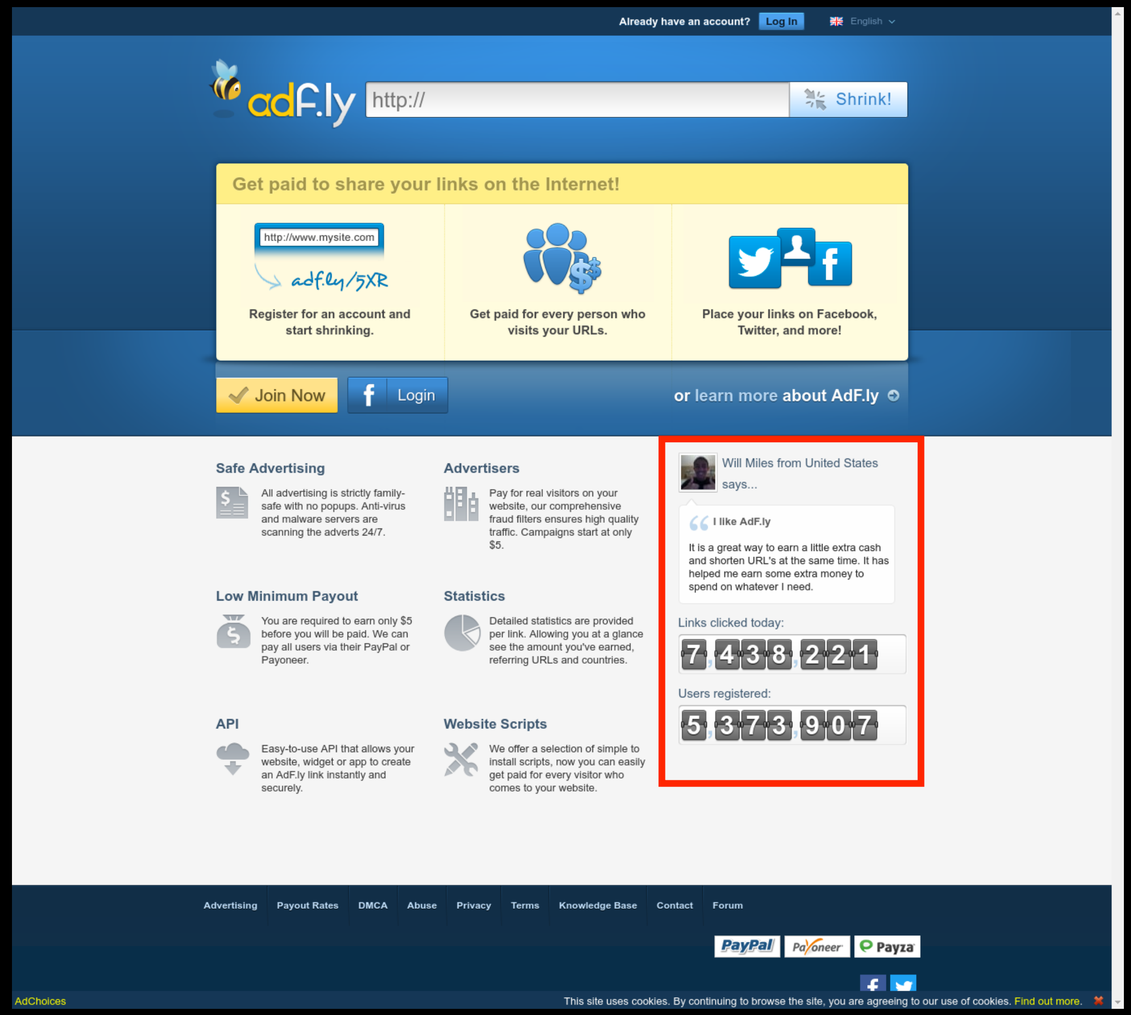
First Frame: Original Screenshot

Second Frame: Reconstructed Screenshot

IMPLEMENTATION OF THE FRAMEWORK

RECONSTRUCTING ARCHIVED SCREENSHOTS

First Frame: Original Screenshot



Second Frame: Reconstructed Screenshot

IMPLEMENTATION OF THE FRAMEWORK

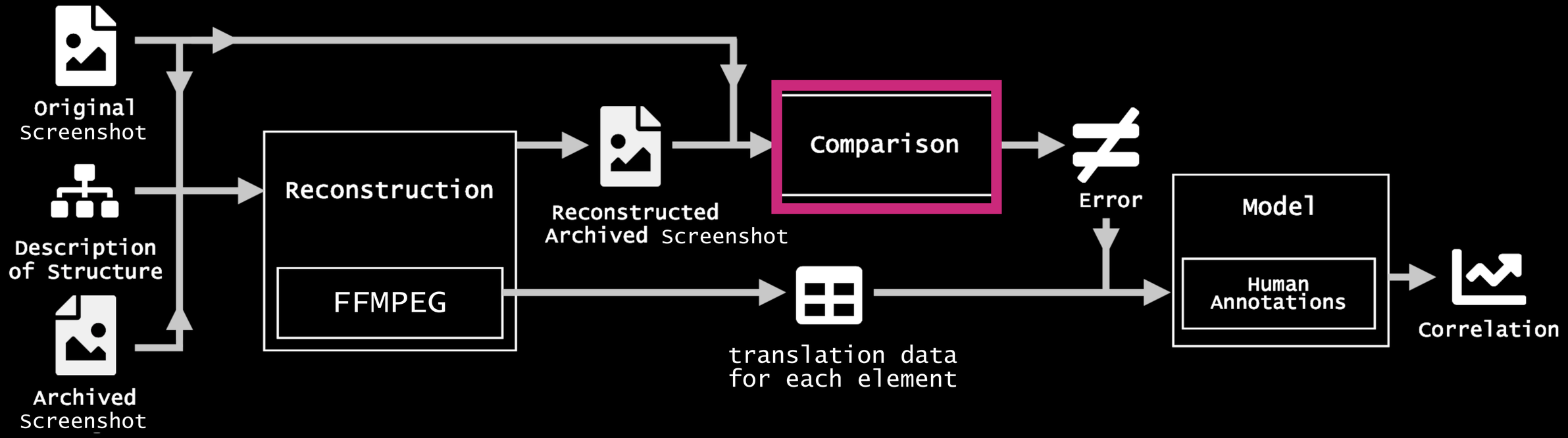
RECONSTRUCTING ARCHIVED SCREENSHOTS

First Frame: Original Screenshot

Second Frame: Reconstructed Screenshot

IMPLEMENTATION OF THE FRAMEWORK

COMPARISON



IMPLEMENTATION OF THE FRAMEWORK

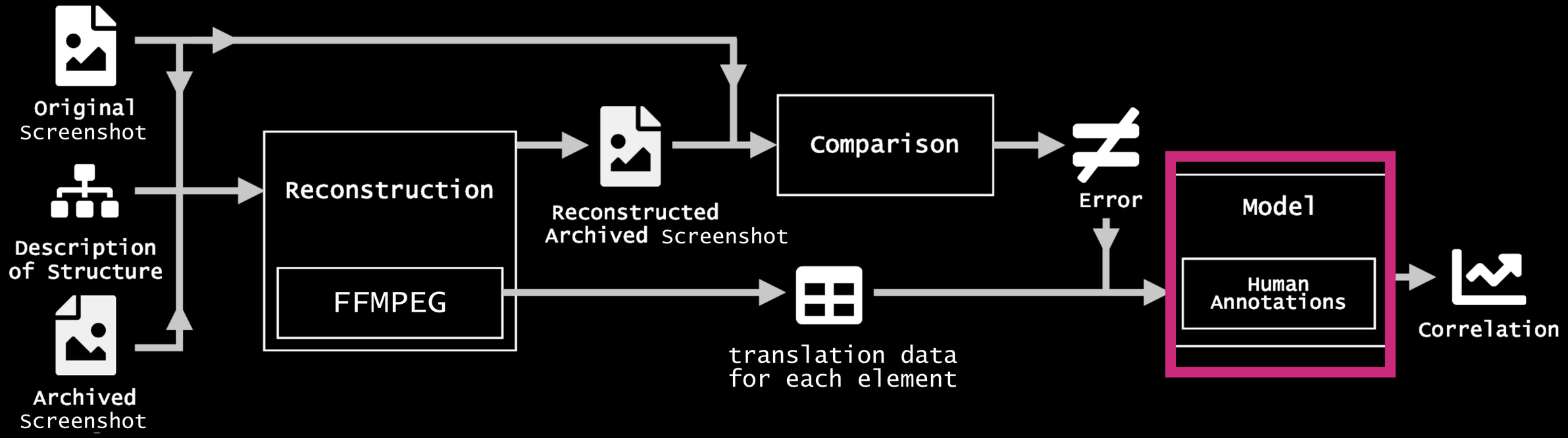
COMPARISON

	AVG RELATIVE PIXEL ERROR
ARCHIVED SCREENSHOTS VS ORIGINAL	23.11 %
RECONSTRUCTED ARCHIVED SCREENSHOTS	21.07 %

REDUCTION OF PIXEL ERROR: 2.04 %

IMPLEMENTATION OF THE FRAMEWORK

MODEL



LINEAR REGRESSION RESULTS

	ACCURACY
ARCHIVED SCREENSHOTS VS ORIGINAL	74.8 %
RECONSTRUCTED ARCHIVED SCREENSHOTS VS ORIGINAL	76.6 %

IMPROVEMENT THROUGH RECONSTRUCTION: 1.8 %

Transferring Relevance Judgments Between Different Web Crawls

The Good,

the Bad,

and the Ugly

Maik Fröbe Big Data Analytics

12.03.2021

webis.de

Relevance Label Transfer

Use Case

- Evaluation of retrieval models for web search
- Requirements for evaluation:
  - Web Crawl
  - Relevance labels
- Problem:
  - The WWW evolves

Relevance Label Transfer

Use Case

- Evaluation of retrieval models for web search
- Requirements for evaluation:
  - Web Crawl
  - Relevance labels
- Problem:
  - The WWW evolves
- Idea:

Use near-duplicate detection to **transfer** relevance judgments to **newer crawls**

Relevance Label Transfer

Relevance Transfer from ClueWeb09 to ClueWeb12

- Assumption: Near duplicates have the same relevance

- ClueWeb09
- Relevant for query “designer dog breeds”

- ClueWeb12

Relevance Label Transfer

Relevance Transfer from ClueWeb09 to ClueWeb12

- Starting point: ClueWeb09
  - 73,883 relevance judgments
  - Judgment effort: **4-8 months**

Relevance Label Transfer

Relevance Transfer from ClueWeb09 to ClueWeb12

- Starting point: ClueWeb09
  - 73,883 relevance judgments
  - Judgment effort: **4-8 months**

- Pilot Study:
  - $\geq$ **51%** of judged URLs available in 2012

Relevance Label Transfer

Relevance Transfer from ClueWeb09 to ClueWeb12

- Starting point: ClueWeb09
  - 73,883 relevance judgments
  - Judgment effort: **4-8 months**

- Pilot Study:
  - $\geq$ **51%** of judged URLs available in 2012

Relevance Label Transfer

Relevance Transfer from ClueWeb09 to ClueWeb12

- Starting point: ClueWeb09
  - 73,883 relevance judgments
  - Judgment effort: **4-8 months**

- Pilot Study:
  - $\geq$ **51%** of judged URLs available in 2012

- Judged ClueWeb09 URLs in ClueWeb12:
  - **24%** crawled in ClueWeb12
  - **8%** are near-duplicates in ClueWeb12

Relevance Label Transfer

Relevance Transfer from ClueWeb09 to ClueWeb12

- Starting point: ClueWeb09
  - 73,883 relevance judgments
  - Judgment effort: **4-8 months**

- Pilot Study:
  - $\geq$ **51%** of judged URLs available in 2012

- Judged ClueWeb09 URLs in ClueWeb12:
  - **24%** crawled in ClueWeb12
  - **8%** are near-duplicates in ClueWeb12

Relevance Label Transfer

Relevance Transfer from ClueWeb09 to ClueWeb12

- Starting point: ClueWeb09
  - 73,883 relevance judgments
  - Judgment effort: **4-8 months**

- Pilot Study:
  - $\geq$ **51%** of judged URLs available in 2012

- Judged ClueWeb09 URLs in ClueWeb12:
  - **24%** crawled in ClueWeb12
  - **8%** are near-duplicates in ClueWeb12

- Full-corpus duplicate detection:
  - **10%** near-duplicates in ClueWeb12
    Judgment effort: **1-3 weeks**

Relevance Label Transfer

Relevance Transfer from ClueWeb09 to ClueWeb12

- Starting point: ClueWeb09
  - 73,883 relevance judgments
  - Judgment effort: **4-8 months**

- Pilot Study:
  - $\geq$ **51%** of judged URLs available in 2012

- Judged ClueWeb09 URLs in ClueWeb12:
  - **24%** crawled in ClueWeb12
  - **8%** are near-duplicates in ClueWeb12

- Full-corpus duplicate detection:
  - **10%** near-duplicates in ClueWeb12
    Judgment effort: **1-3 weeks**

Relevance Label Transfer

Relevance Transfer from ClueWeb09 to ClueWeb12

- Starting point: ClueWeb09
  - 73,883 relevance judgments
  - Judgment effort: **4-8 months**

- Pilot Study:
  - $\geq$ **51%** of judged URLs available in 2012

- Judged ClueWeb09 URLs in ClueWeb12:
  - **24%** crawled in ClueWeb12
  - **8%** are near-duplicates in ClueWeb12

- Full-corpus duplicate detection:
  - **10%** near-duplicates in ClueWeb12
    Judgment effort: **1-3 weeks**

Thank You!

UNIVERSITÄT
LEIPZIG

Phoenix - Scientific Text Reuse

Lukas Gienapp

March 11, 2021

Leipzig University

PHOENIX

- BMBF-funded research project between Bauhaus-Universität Weimar, the DZHW Berlin, and Leipzig University
- Aim of the project: reflect on the practice of scientific authorship and scientific writing; How are scientific texts produced today?
- Project in Leipzig: study **Scientific Text Reuse** – citing, paraphrasing, summarizing, copying – how widespread are these in current scientific writing?

Scientific Text Reuse

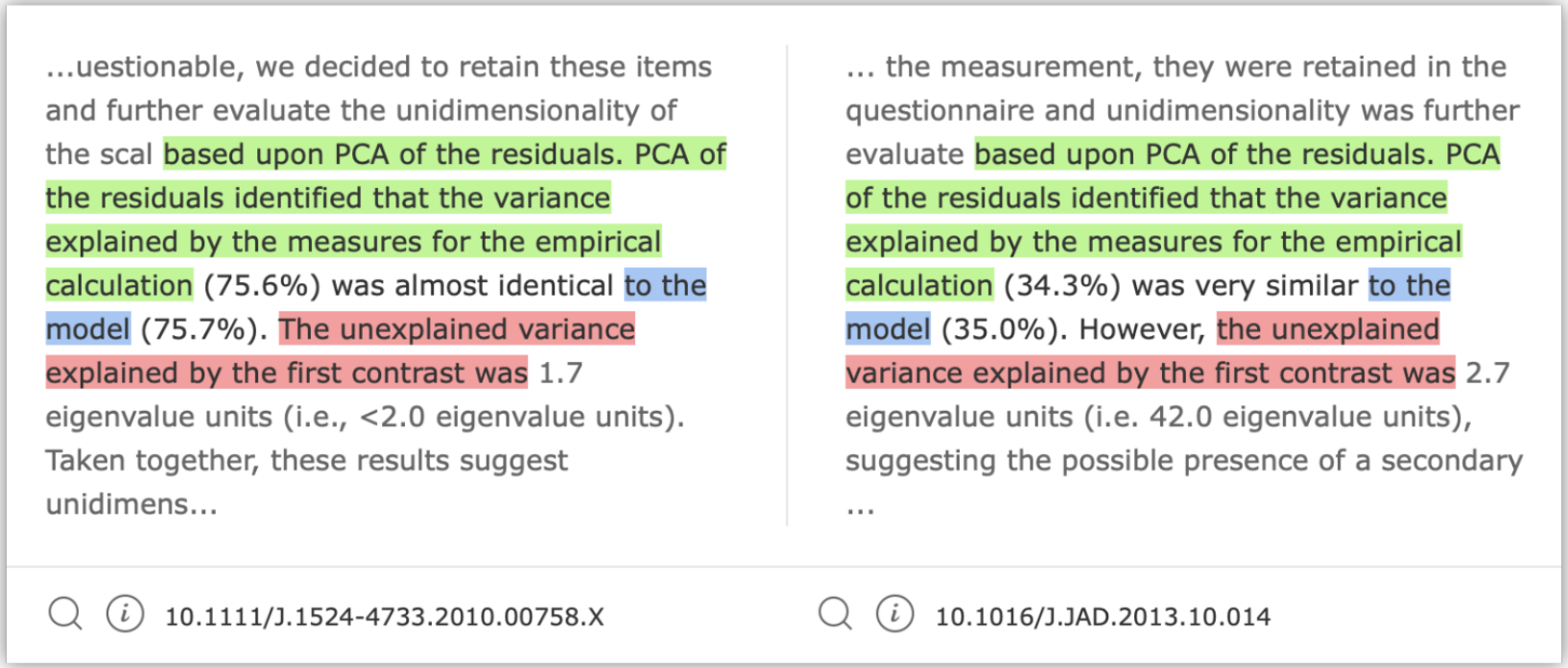


Figure 1: Example of reuse case (standardized experiment description text)

Text Alignment

- **Candidate Retrieval**: given a document collection, identify all pairings that are highly probable to contain text reuse
  - recall-optimized approach
  - MinHash on small document chunks to identify word overlap
  - current granularity: 15 overlapping words in a 50 word chunk
- **Alignment**: given two texts, compute the position of all sub-spans of text that are highly similar in both
  - precision-optimized approach
  - *Seeding*: compare chunked documents with a distance function to identify similar sub-passages
  - *Extending*: combine sub-passages close to each other
  - current setup: 8-grams with hash identity

Experiment

Documents	$\approx$	6×10^6
Possible Comparisons	$\approx$	36×10^{12}
Conducted Comparisons	$\approx$	9×10^9
Found Reuse Cases[1]	$\approx$	80×10^6

[1] Estimated, alignment process is ongoing

Roadmap

- Ongoing:
  - finish alignment process
  - aggregate metadata + infer scientific discipline
- Planned for the (near) future:
  - deduplication/clustering
  - classify text reuse cases into a taxonomy
  - web demo for accessible data exploration
- End goal:
  - analysis of text reuse wrt. scientific discipline & reuse type

Image Captions as Paraphrases

Webis Flash Talks 2021

Marcel Gohsen

What is a Paraphrase?

What is a paraphrase?

What is a Paraphrase?

What is a paraphrase?

How do you define a paraphrase?

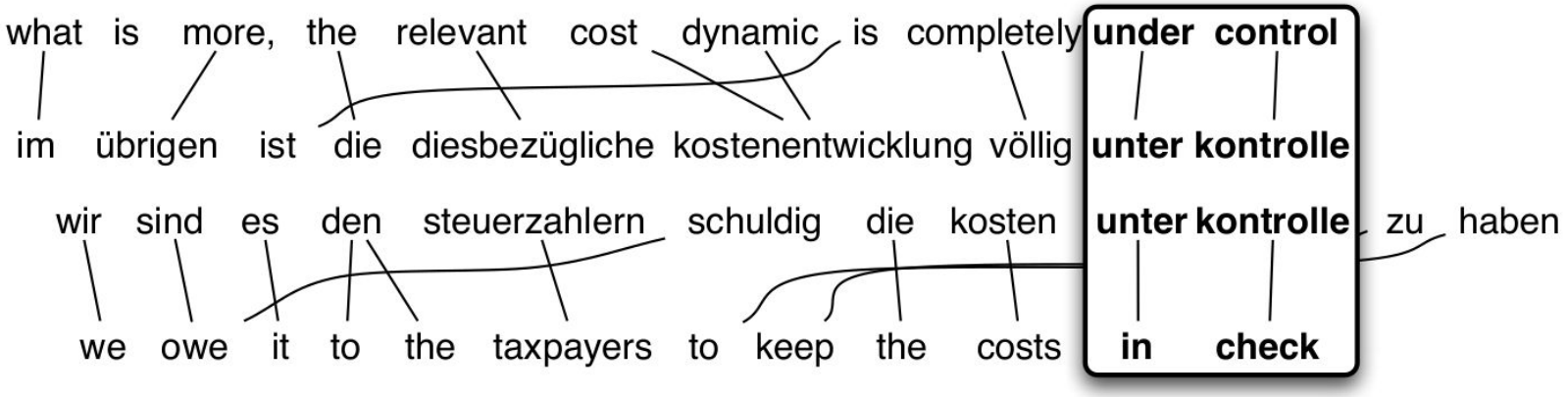
What is a Paraphrase?

What is a paraphrase?

How do you define a paraphrase?

Motivation

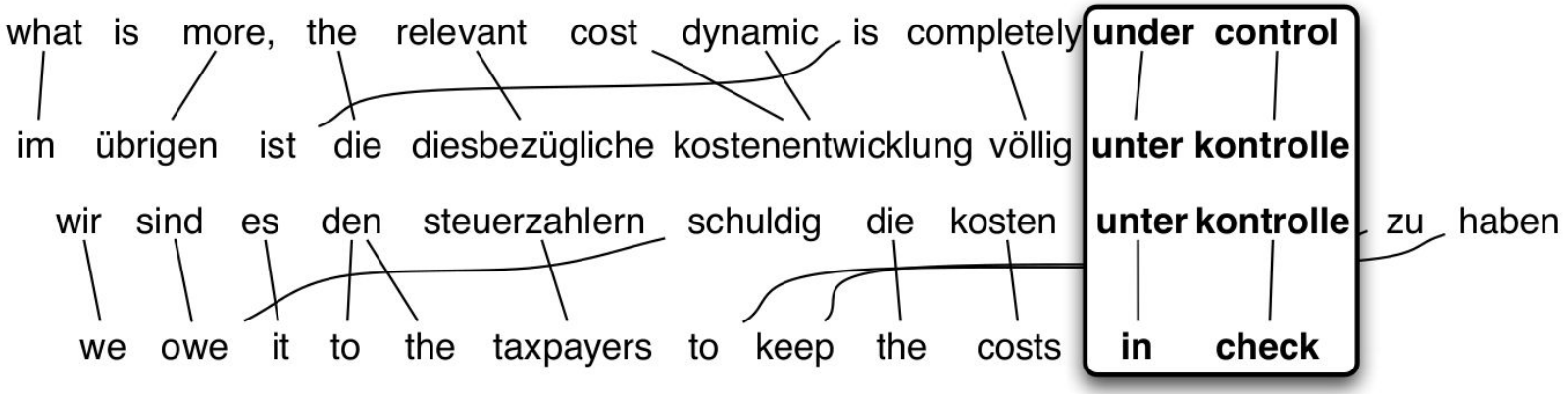
Most automatic paraphrase acquisition methods are based on this approach:



Bannard et al., 2005, ACL 2005

Motivation

Most automatic paraphrase acquisition methods are based on this approach:

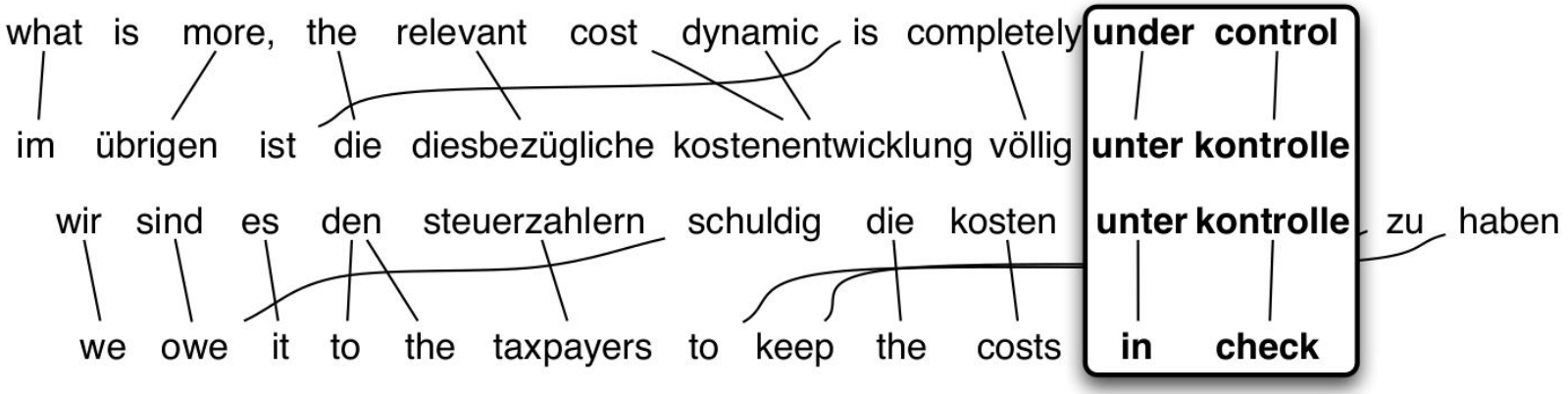


Bannard et al., 2005, ACL 2005

- Only phrasal paraphrases

Motivation

Most automatic paraphrase acquisition methods are based on this approach:

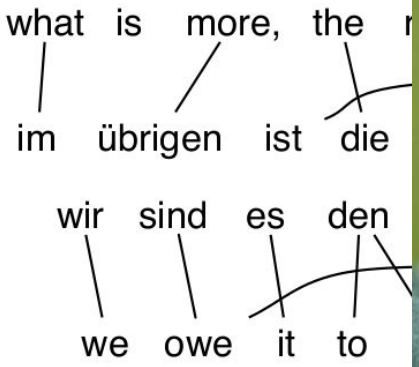


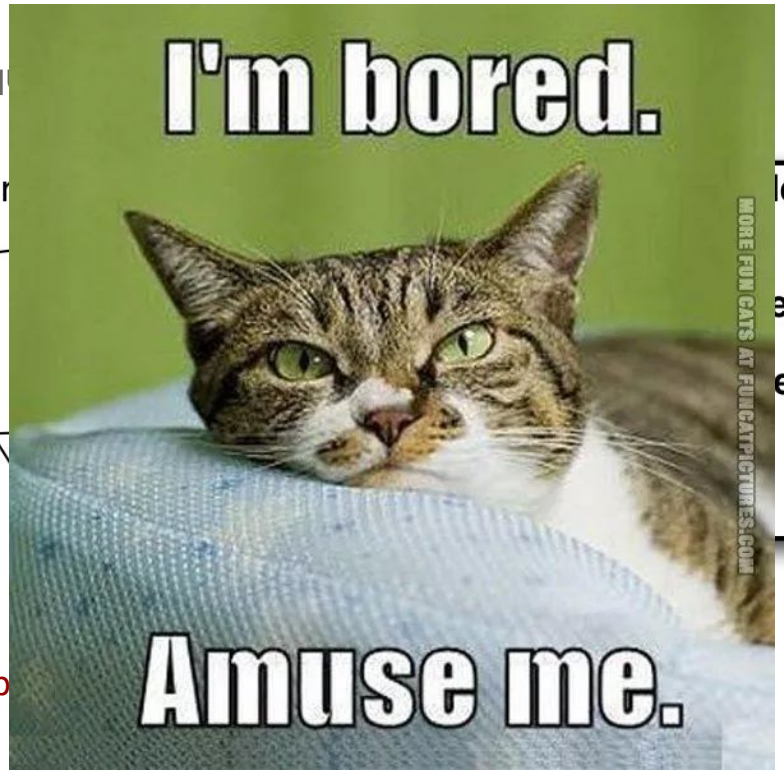
Bannard et al., 2005, ACL 2005

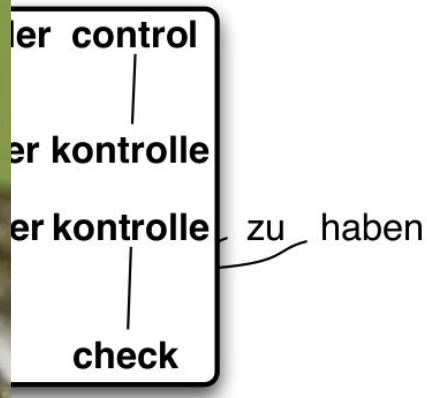
- Only phrasal paraphrases
- Require large parallel corpora

Motivation

Most automatic paraphrase acq







Bannard et al., 2005, ACL 2005

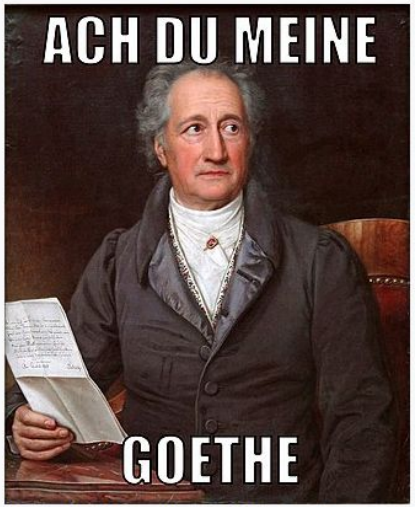
- Only phrasal paraphrases
- Require large parallel corp

Novel Paraphrase Approach

Novel Paraphrase Approach

https://en.wikipedia.org/wiki/Johann_Wolfgang_von_Goethe

Johann Wolfgang von Goethe

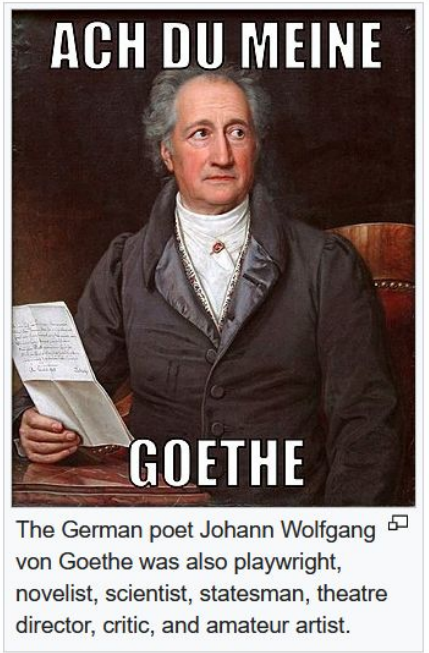


Johann Wolfgang von Goethe was a German poet, playwright, novelist, scientist, statesman, theatre director, critic, and amateur artist.

Novel Paraphrase Approach

https://en.wikipedia.org/wiki/Johann_Wolfgang_von_Goethe

https://en.wikipedia.org/wiki/Goethe%E2%80%93Schiller_Monument



Novel Paraphrase Approach

https://en.wikipedia.org/wiki/Johann_Wolfgang_von_Goethe

Paraphrase

https://en.wikipedia.org/wiki/Goethe%E2%80%93Schiller_Monument

Data

Wikimedia dumps

- High ratio of image captions
- High textual quality

- Limited number of pages
  (~ 20 million English articles)
- No “physical” images

“Small” gold standard corpus

Data

Wikimedia dumps

- High ratio of image captions
- High textual quality

- Limited number of pages
  (~ 20 million English articles)
- No “physical” images

→ “Small” gold standard corpus

INTERNET ARCHIVE

Web Archive

- Large number of pages
- Many images (hundreds of millions)

- Smaller ratio of images to captions
- Varying textual quality
- Noisy caption extraction strategies

→ Large silver corpus

Thank you!  I’m grateful!

On Simulating Human Weirdness

Sebastian Günther

The SINIR Project

- **S**imulating **IN**teractive **I**nformation **R**etrieval

Problem:

- Evaluating changes to a search engine/digital library is:
  - Cost intensive
  - Slow
  - You need actual users 😐

Solution:

Idea

- Develop a simulation framework to evaluate visual and retrieval system changes

- Aspects to cover:

Query Formulation

Eye Tracking!

Clicking Behaviour

Tyypos!!1!

Cost/Gain

Reading Time?

Stopping Behavior

What about missclicks!?

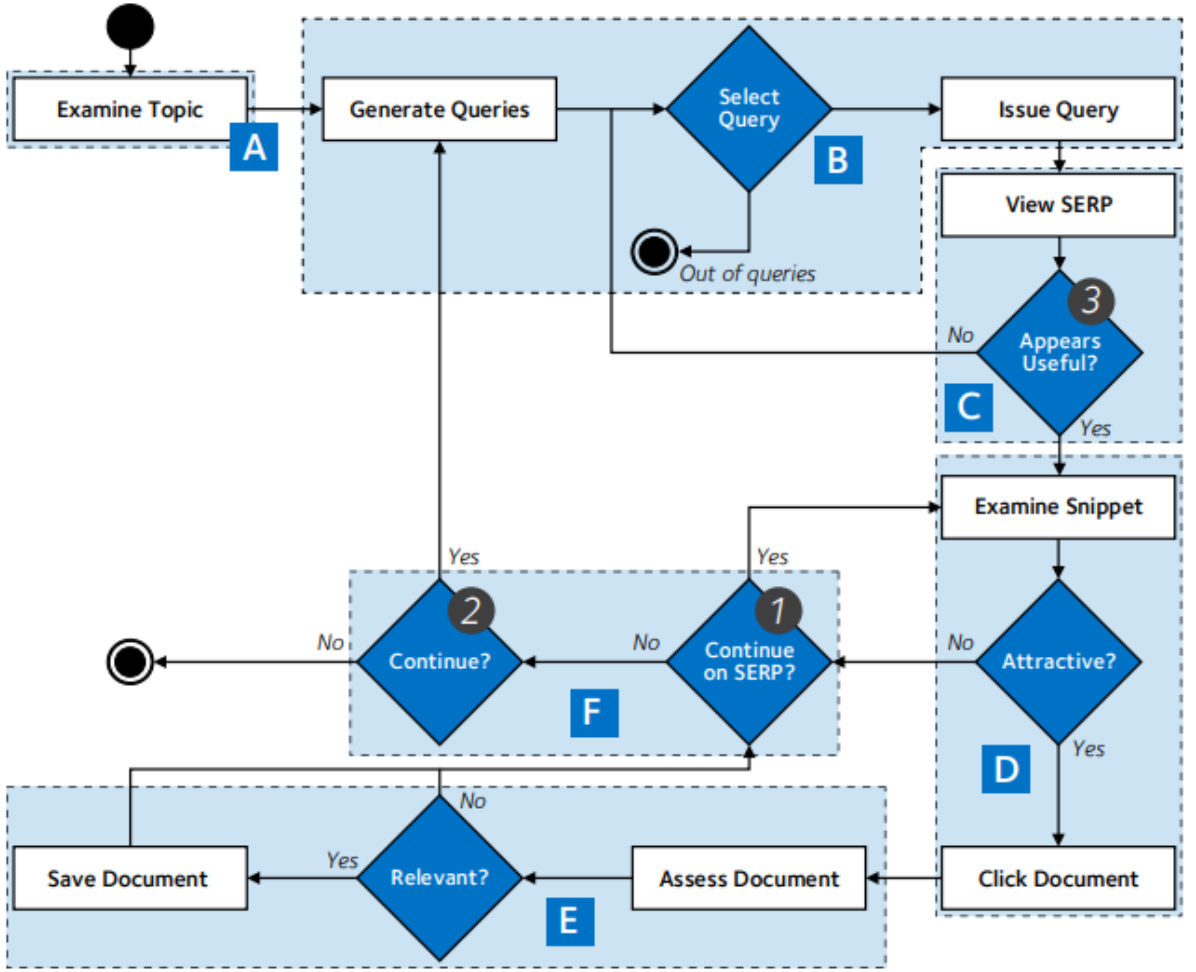
Query Re-Formulation

Knowledge Model

...

The Complex Searcher Model

- Information need
- Queries
- List of results
- Examing Snippets
- Examing Documents
- Stopping:
  - New query
  - Satisfied
  - …give up?



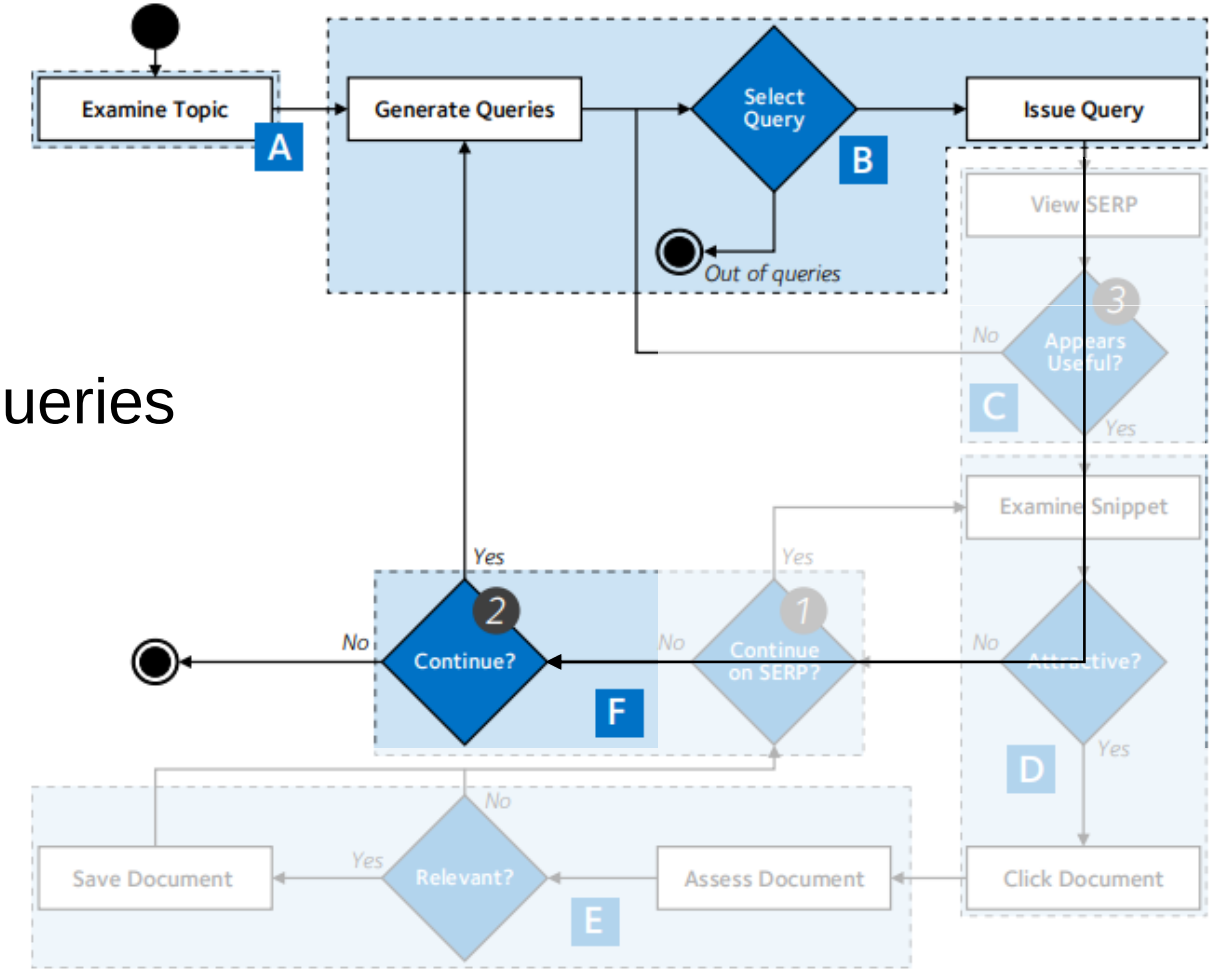
source: D. Maxwell [1]

So far: It‘s all about Queries!

Task:

- Build an artifical session
  - Queries only
  - Containing multiple related queries
  - On a certain topic

A 3rd party should <u>not</u> be able to tell if the session is real or generated!



source: D. Maxwell [1]

Can we just… utilize Google?

- They get queries all day
- … and they use queries to build their suggestions
- … and they provide their suggestions

…profit?

- A good start!
- But: neglects the human aspects
- Let‘s try it anyways…

Google Query Logs Experiments (1)

TREC Robust 05 Topic #303

- `Hubble Telescope Achievements`
- `hubble space telescope achievements` *// okay, fine!*
- `hubble space telescope achievement` *// …uhm yeah?*
- `hubble space telescope achievements` *// what?*
- `hubble space telescope achievement` *// this is going nowhere*

Google Query Logs Experiments (2)

TREC Robust 05 Topic #389

- `illegal technology transfer`
- `forced technology transfer`
- `forced technology transfer wto`
- `trips agreement technology transfer`
- `transfer of technology`
- `transfer of technology pdf`
- `transfer of technology ppt`
- `transfer of technology in agriculture ppt`
- `transfer of technology in agriculture pdf`
- `transfer of technology in agriculture ppt` *// seriously?*

Pitfalls

- There might be loops
- We may stray away from a topic (instead of getting more specific)
- Sometimes there are even 0 (zero!) suggestions
- Heavily personalized (language, location, history, etc.)

Solution:

- utilize multiple sources + intelligent selection

On the Horizon…

- A small framework to support the development of query log generation approaches
  + Evaluation of query logs

- Thesis: Learning user behaviour from an action log
  - We have all the actions of all the users for one year

Top 3 learnings

- Good datasets are rare
- Googles wisdom is finite
  - (or at least they make it appear so)
- Humans are weird
  - (read *any* query logs!)

Thank you!

References

[1] Maxwell, David Martin. *Modelling search and stopping in interactive information retrieval*. Diss. University of Glasgow, 2019.

Title image: by Oladimeji Ajegbile (https://www.pexels.com/de-de/foto/mann-der-mit-einem-laptop-arbeitet-2696299/)

UNIVERSITÄT
LEIPZIG

How to get FAME in 7 steps

Leipzig, 12.03.2021

Ahmad Dawar Hakimi

UNIVERSITÄT
LEIPZIG

How to get "A Framework for Argumentation and Evaluation" in 7 steps

Leipzig, 12.03.2021

Ahmad Dawar Hakimi

FAME Project

Abstract: Two different perspectives on argumentation have been pursued in computer science research:

1. Argument mining from natural language texts at large scale
2. Formal argument evaluation

So far largely independent and unrelated ⇨ project goal: link 1 + 2

- Project hypothesis: controlled natural language (CNL) can serve as an intermediate representation of argumentative text
- We start with Attempto Controlled English (ACE) for representation of arguments (Fuchs et al. 2008)
- (Semi-)automatic retrieval of argumentative units on selected issues from empirical texts ⇨ manual encoding/mapping into ACE representation

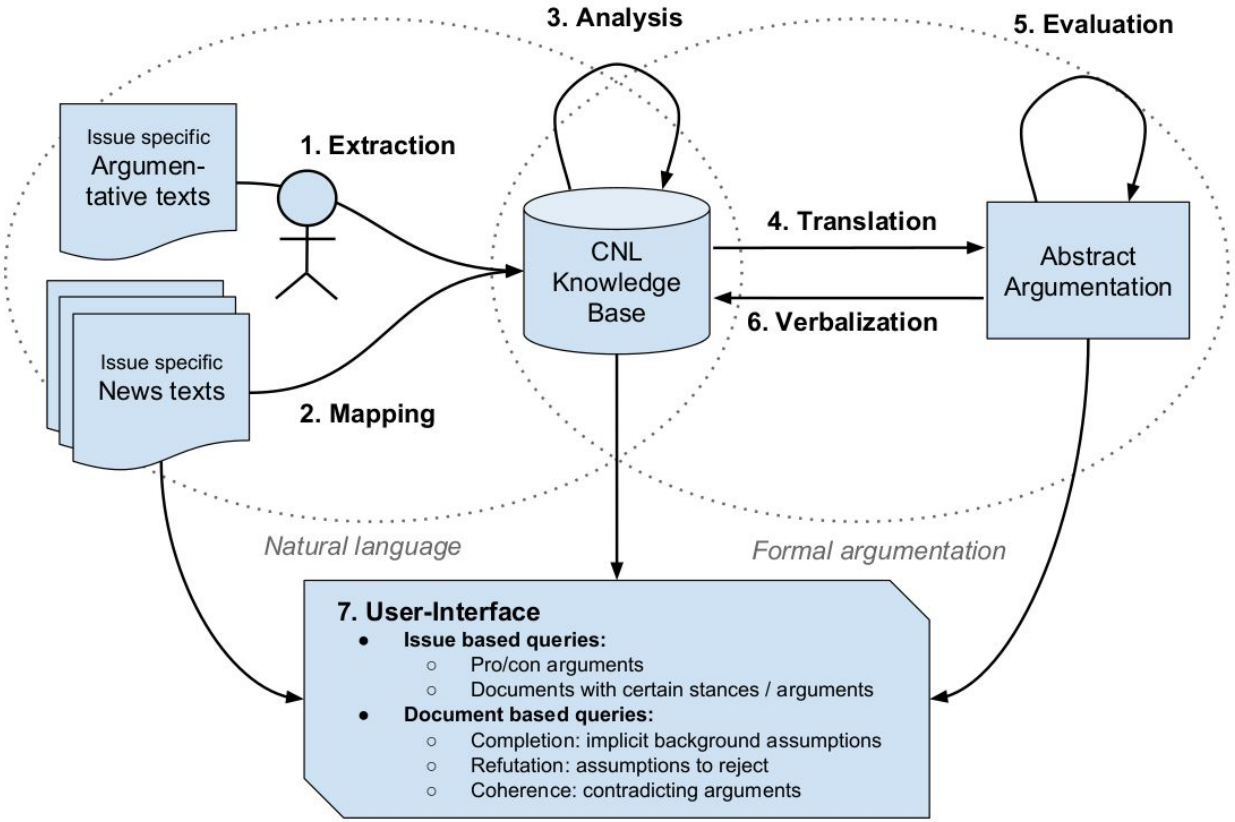


Figure 1: FAME architecture

UNIVERSITÄT
LEIPZIG

Thank You!

Ahmad Dawar Hakimi
Automatische Sprachverarbeitung
Augustusplatz 10, 04109 Leipzig

hakimi@informatik.uni-leipzig.de
http://asv.informatik.uni-leipzig.de/de/staff/Ahmad_Hakimi

Webis Flash Talks 2021

What does my Language Model know?

Felix Helfer
ASV Leipzig

Answer-aware Question Generation

- As a form of information retrieval from unstructured documents (for chatbot engines).

- Adapted GPT-2 (and variants) give somewhat solid first results.

S: *The selected language will be stored in a cookie on your computer and will be automatically selected at your next visit.*

A: *The selected language*

→ ***What will be stored in a cookie on your computer?***

Language Models: What do they know? Do they know things? Let's find out!

S: *He encourages a distinction between lawful protest demonstration, nonviolent civil disobedience, and violent civil disobedience.*

A: *lawful protest demonstration, nonviolent civil disobedience, and violent civil disobedience*

→ What did Whitehead want?

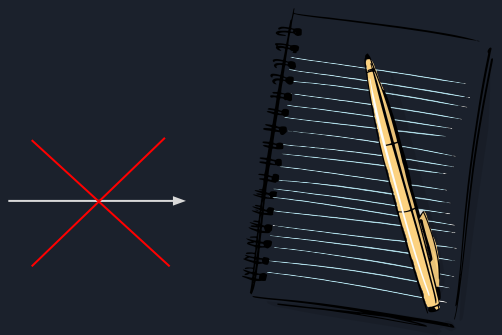
Pierre Legrand

Arthur North Whitehead

But...

Wrong Legrand(e)!

Pierre Legrand



Nonviolent Civil Disobedience and Police Enforcement Policy (1968)

J. L. LeGrande

Language Models and Information

- Inherent “semantic knowledge” is unreliable at best when the LM is used for a focused task.

- Need context, attention, preprocessing, etc. to “steer” a generative model in the right direction.

… but what about a more “freestyle” task?

Generating Punchlines

- Similar model, input is now a “setup” sentence.
- Trained on a reddit joke dataset, filtered for one-liners with a hyphen that got at least one upvote.

S: *How many physicists does it take to screw in a light bulb?*

→ *None, they just oscillate the room.*

S: *A computer scientist and a programmer walk into a bar*

→ *they are greeted with an awkward silence*

Topical Punchlines?

How about some topical jokes, GPT-2?

S: *What's the difference between lawful protest demonstration and violent civil disobedience?*

→ *No one will beat you on the head*

S: *Whitehead walks into a bar*

→ *then a table at the far corner and a chair at the top.*

Topical Punchlines?

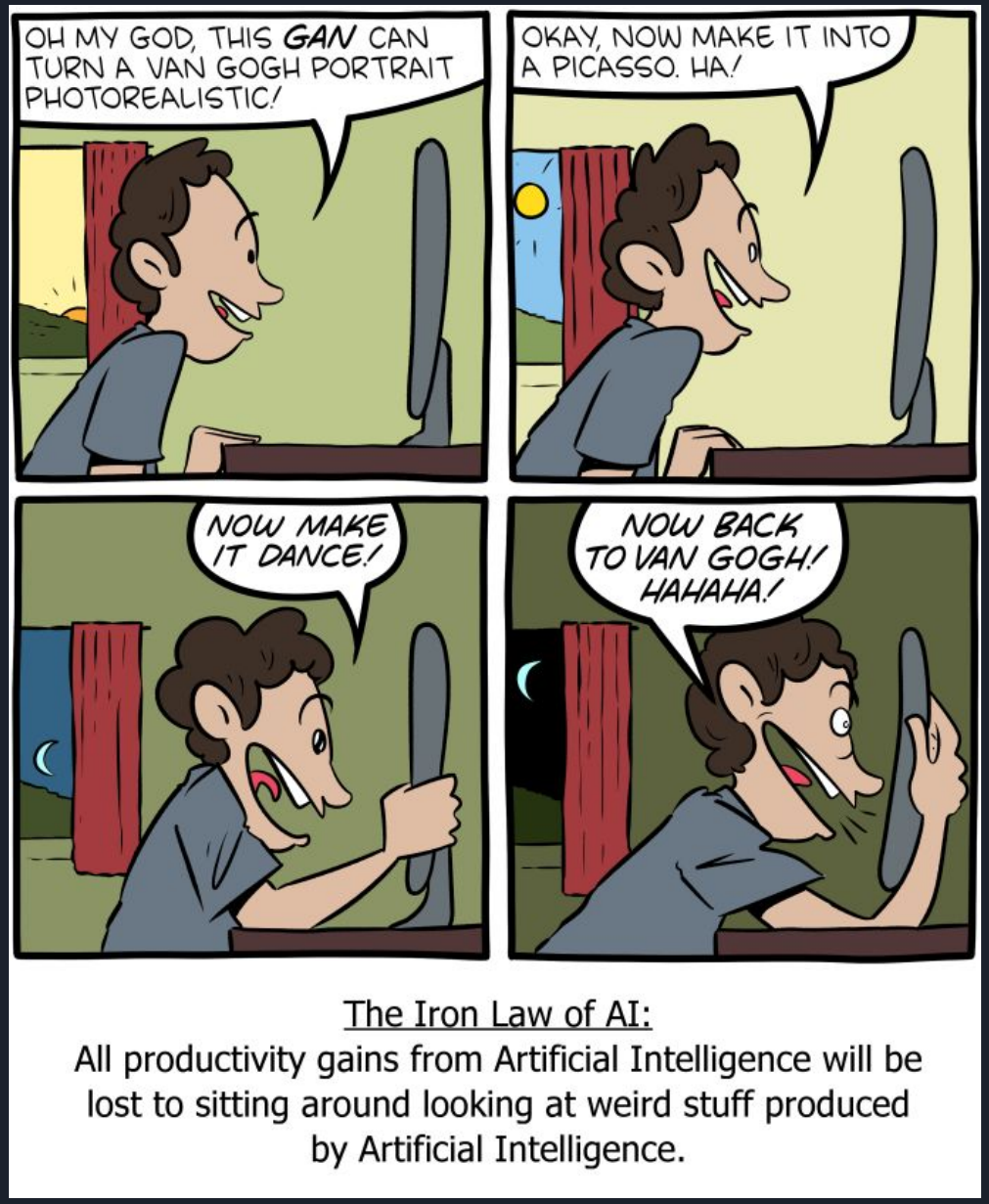
Mabe anything tying our two authors together?

S: *What do Legrand and Whitehead have in common?*

→ *They're both dead.*

… Fair enough.

Bonus comic



Source:
https://www.smbc-comics.com/comic/gan

Thanks for listening!

Image sources

https://en.wikipedia.org/wiki/Pierre_Legrand#/media/File:Legrand,_Pierre.jpg

https://en.wikipedia.org/wiki/Alfred_North_Whitehead#/media/File:Alfred_North_Whitehead_in_1936.jpg

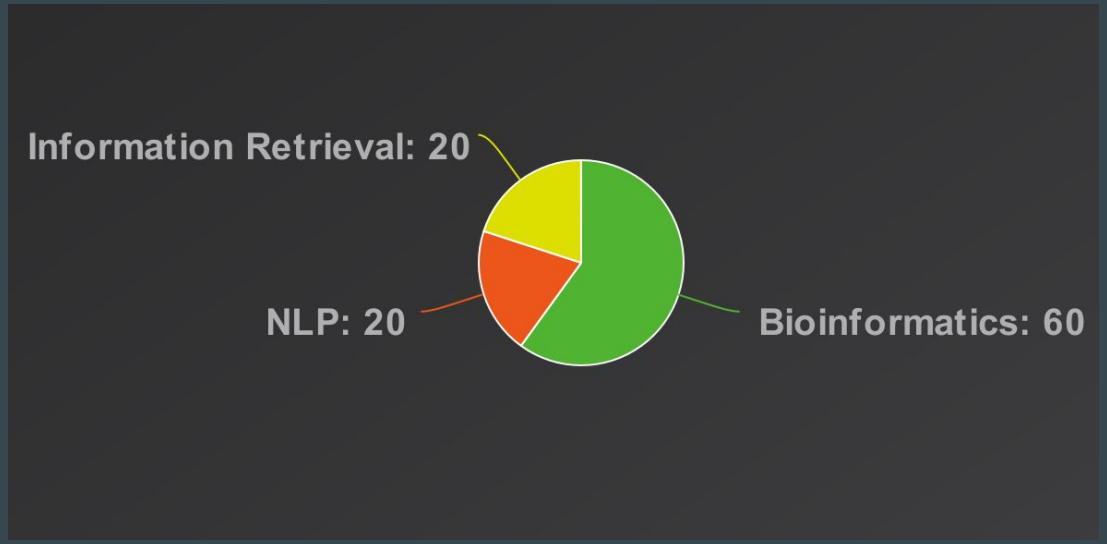
https://www.seekpng.com/ipng/u2q8a9u2u2t4u2t4_notebook-png-transparent-photo-pen-and-paper-cartoon/

https://www.smbc-comics.com/comic/gan

Understanding Comparative Questions

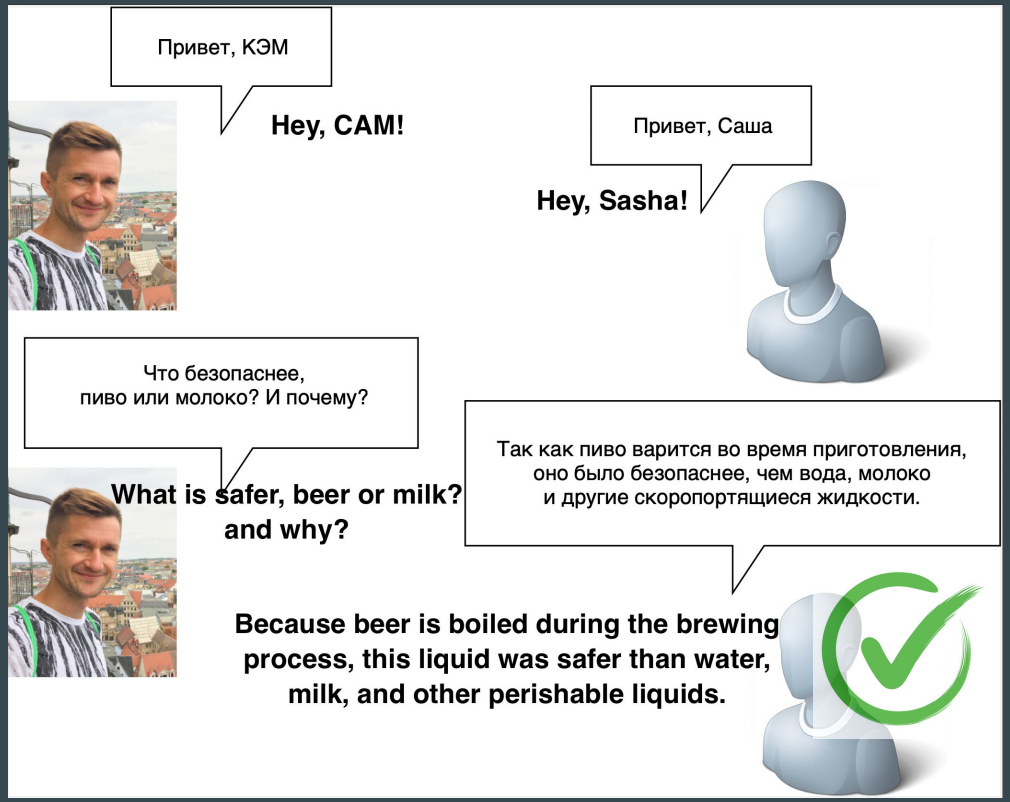
Jonas Hirsch, Alexander Bondarenko (Sascha) and Matthias Hagen

Who I am



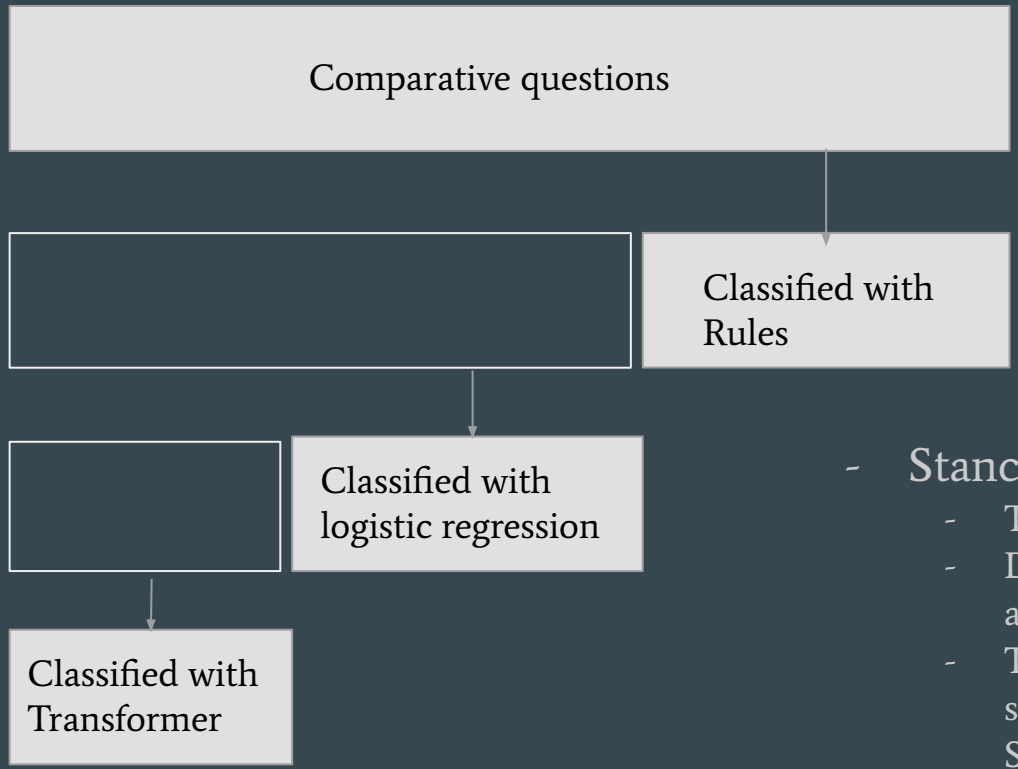
- 5th semester Bioinformatics Bachelor student
- Since October 2020 student assistant in the Big Data Analytics team at Halle
- Working with Sascha on comparative questions

What we do



- Bondarenko et al, 2020, Comparative Web Search Questions
- Found 65% of all comparative questions with precision of 1.0 on Russian question queries (1.5 billion)
- NOW: achieve better results on English question queries + stance detection (= beer is safer than milk)

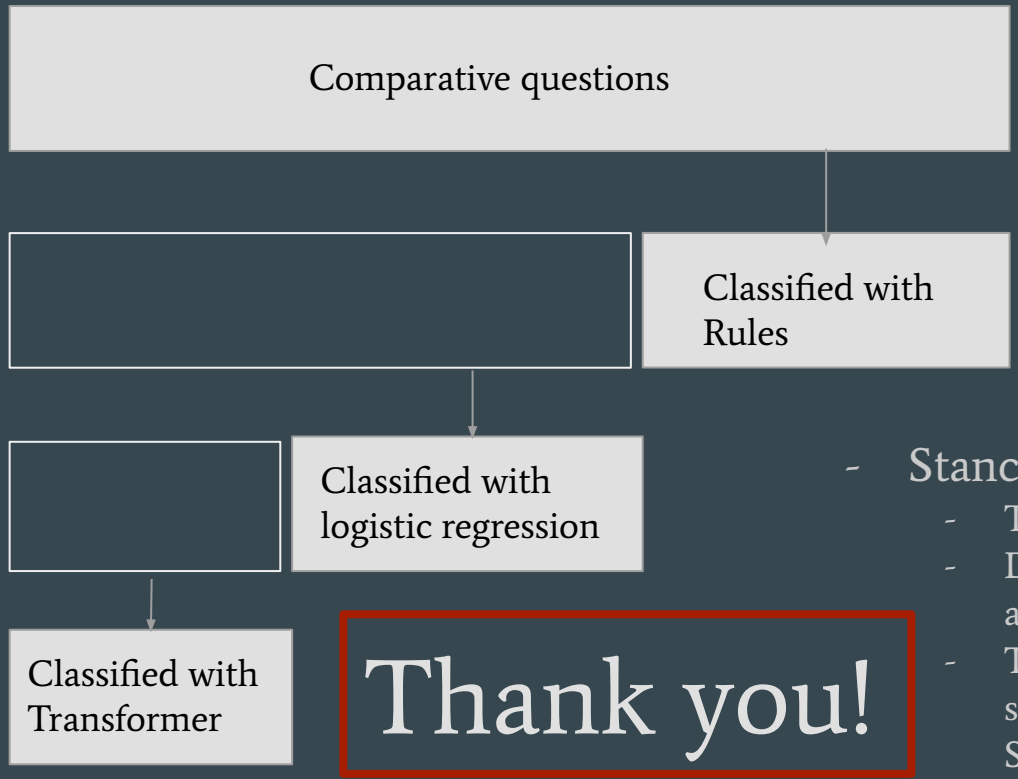
How we do it



- Ensemble of:
  - Handmade rules
  - Logistic regression
  - Transformer (BERT, RoBERTa, ALBERT)

- Stance classification:
  - Target identification
  - Dataset with questions and answers + stance annotation
  - Transformer, Stance flow (local and global stance following Wachsmuth et al, 2015, Sentiment Flow)

How we do it



- Ensemble of:
  - Handmade rules
  - Logistic regression
  - Transformer (BERT, RoBERTa, ALBERT)

- Stance classification:
  - Target identification
  - Dataset with questions and answers + stance annotation
  - Transformer, Stance flow (local and global stance following Wachsmuth et al, 2015, Sentiment Flow)

Thank you!

Applying the iLCM to stop climate change

CHRISTIAN KAHMANN

12.03.2021

What is the iLCM?

IMMIGRANT LAW CENTER — of MINNESOTA —

Supporting families and strengthening communities for more than 20 years

ILCM INTERNATIONAL LADIES' CLUB OF MACAU 澳門國際婦女會

ILCM 2016

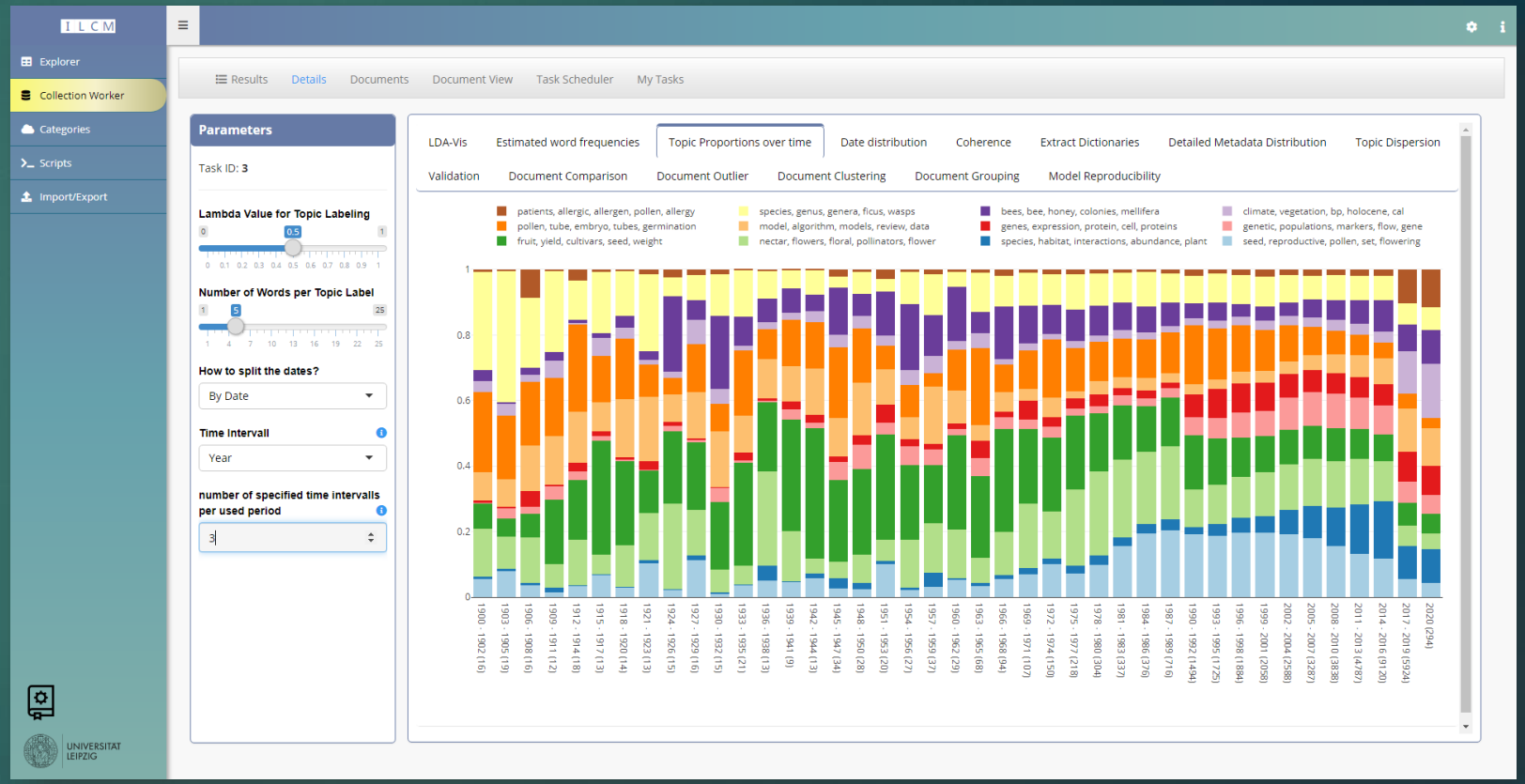
Indian Conference on

Life Cycle Management (ILCM) 2016

17-18 October 2016 | New Delhi, India

Interactive Leipzig Corpus Miner

- Project funded by DFG
- Text Mining Infrastructure
- specifically designed for use in the Social Sciences / Digital Humanities
- Enables even unexperienced users to apply text mining and machine learning algorithms to their data and problems
- Focus on adaptability and extensibility
- https://github.com/ChristianKahmann/ilcm_Shiny
- https://hub.docker.com/r/ckahmann/ilcm



How does the iLCM help to stop climate change?

- Second Project: Transnorms
- Project at political science department of FU Berlin
- Goal: Analysis of translation processes of international norms at different levels of locality
- Selected norms:
  - Climate protection
  - Child labor
  - Prohibition of torture.

How does the iLCM help to stop climate change?

- Second Project: Transnorms
- Project at political science department of FU Berlin
- Goal: Analysis of translation processes of international norms at different levels of locality
- Selected norms:
  - Climate protection
  - Child labor
  - Prohibition of torture.

How does the iLCM help to stop climate change?

- Second Project: Transnorms
- Project at political science department of FU Berlin
- Goal: Analysis of translation processes of international norms at different levels of locality
- Selected norms:
  - Climate protection
  - Child labor
  - Prohibition of torture.

PARIS2015
UN CLIMATE CHANGE CONFERENCE
COP21·CMP11

How does the iLCM help to stop climate change?

- Second Project: Transnorms
- Project at political science department of FU Berlin
- Goal: Analysis of translation processes of international norms at different levels of locality
- Selected norms:
  - Climate protection
  - Child labor
  - Prohibition of torture.

PARIS2015
UN CLIMATE CHANGE CONFERENCE
COP21·CMP11

Donald J. Trump
@realDonaldTrump
Following

The concept of global warming was created by and for the Chinese in order to make U.S. manufacturing non-competitive.

RETWEETS 24,831 LIKES 14,654

2:15 PM - 6 Nov 2012

Donald J. Trump
@realDonaldTrump
Following

Snowing in Texas and Louisiana, record setting freezing temperatures throughout the country and beyond. Global warming is an expensive hoax!

RETWEETS 428 LIKES 358

1:27 AM - 29 Jan 2014

How does the iLCM help to stop climate change?

- Data:
  - National Determined Contributions (NDC) of 183 countries
  - NDC: "*NDCs embody efforts by each country to reduce national emissions and adapt to the impacts of climate change. The Paris Agreement (Article 4, paragraph 2) requires each Party to prepare, communicate and maintain successive nationally determined contributions (NDCs) that it intends to achieve.*"
- Analysis:
  - Combination of quantitative and qualitative methods
  - First: Topic Modelling
  - Qualitative Interpretation and Labeling of Topics
  - Combination of Metadata (e.g. % of renewable energy) and found topics

How does the iLCM help to stop climate change?

- Results:
  - The richer the country, the shorter its NDC
  - Poor countries mention in particular the need for financial support to implement projects
  - Content overlaps often surprising
    - E.g. Serbia is classified in topic with label "Water-related and climate change impacts and dangers", in which apart from that only island states are to be found.
- Next Steps:
  - Strong insight at the global level of locality
  - → what is the situation at national level?
    - Data: laws, action plans, strategies, development plans
- How has the interpretation of the global norm changed at different levels of locality?
- What are the competing targets?

Christian Kahmann

Universität Leipzig

Raum P8-18

Email: kahmann@informatik.uni-leipzig.de

Thank You!

TEXT QUALITY IN SEARCH – SUPPORTED WRITING

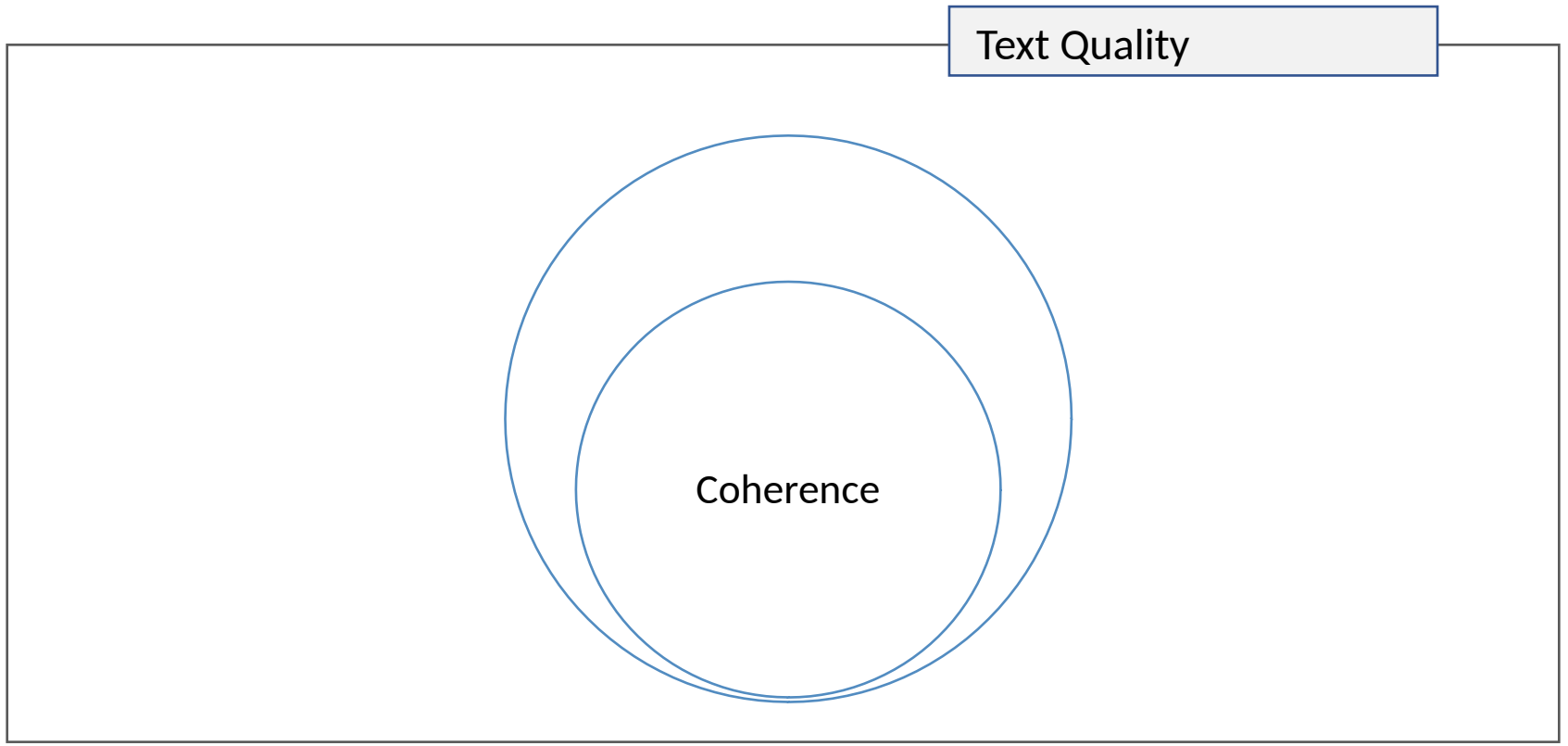
Supervised by: Michael Völske and Magdalena Wolska

Presenter: Bibek Khadayat

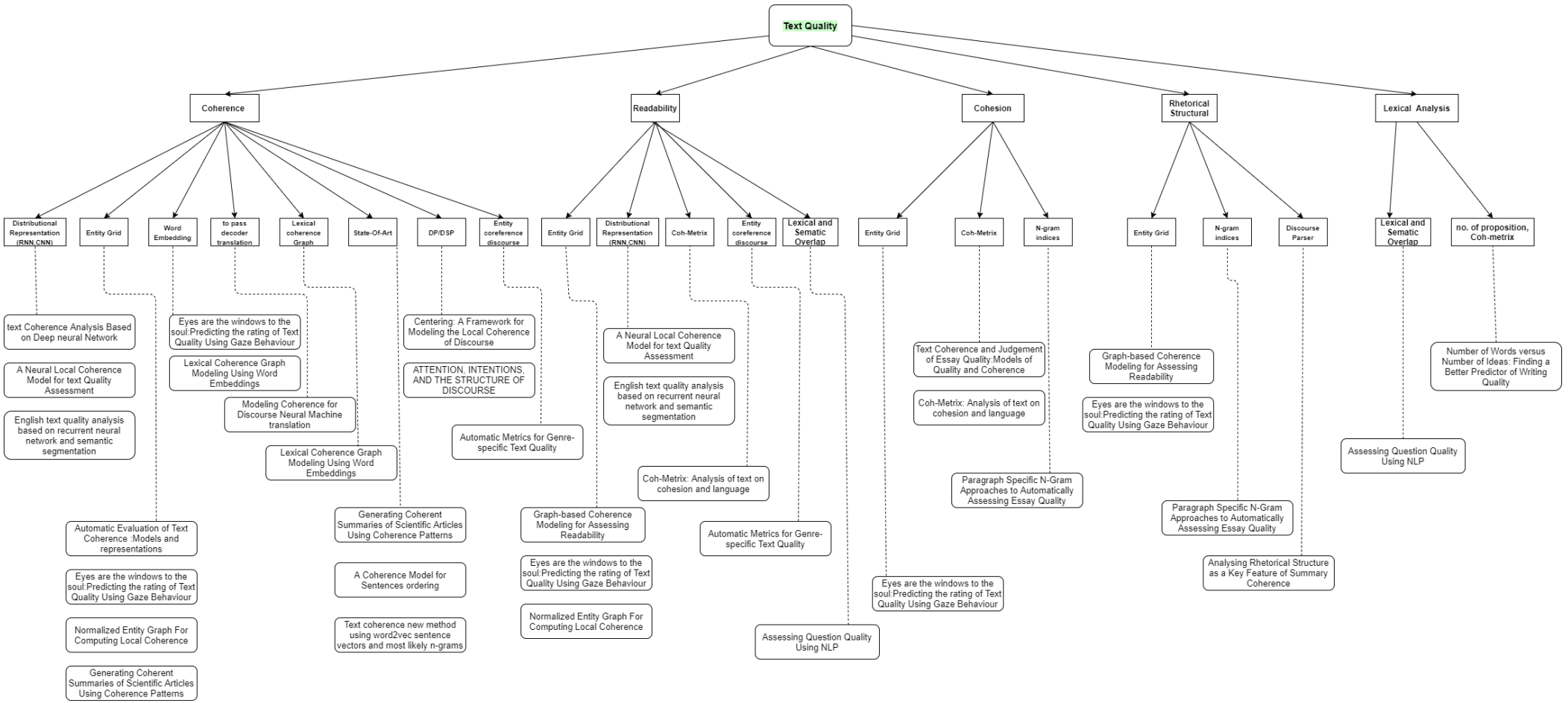
12/03/2020

Overview

- Meaning full sentence at least with a verb and full stop is a good text.
- Well organized, cohesive, coherence, readable etc. has high text quality.
- Coherence → Latin verb co-herence → to stick together.
- Coherence is Link or connection based on idea , concept or theme with smooth text flow.
- Cohesion is on basis of grammar ,word, phrase or clause .
- Readability is ease to understand the text, it depend on word selection , text format and organization


Text Quality
Coherence

Literature Overview



Dataset

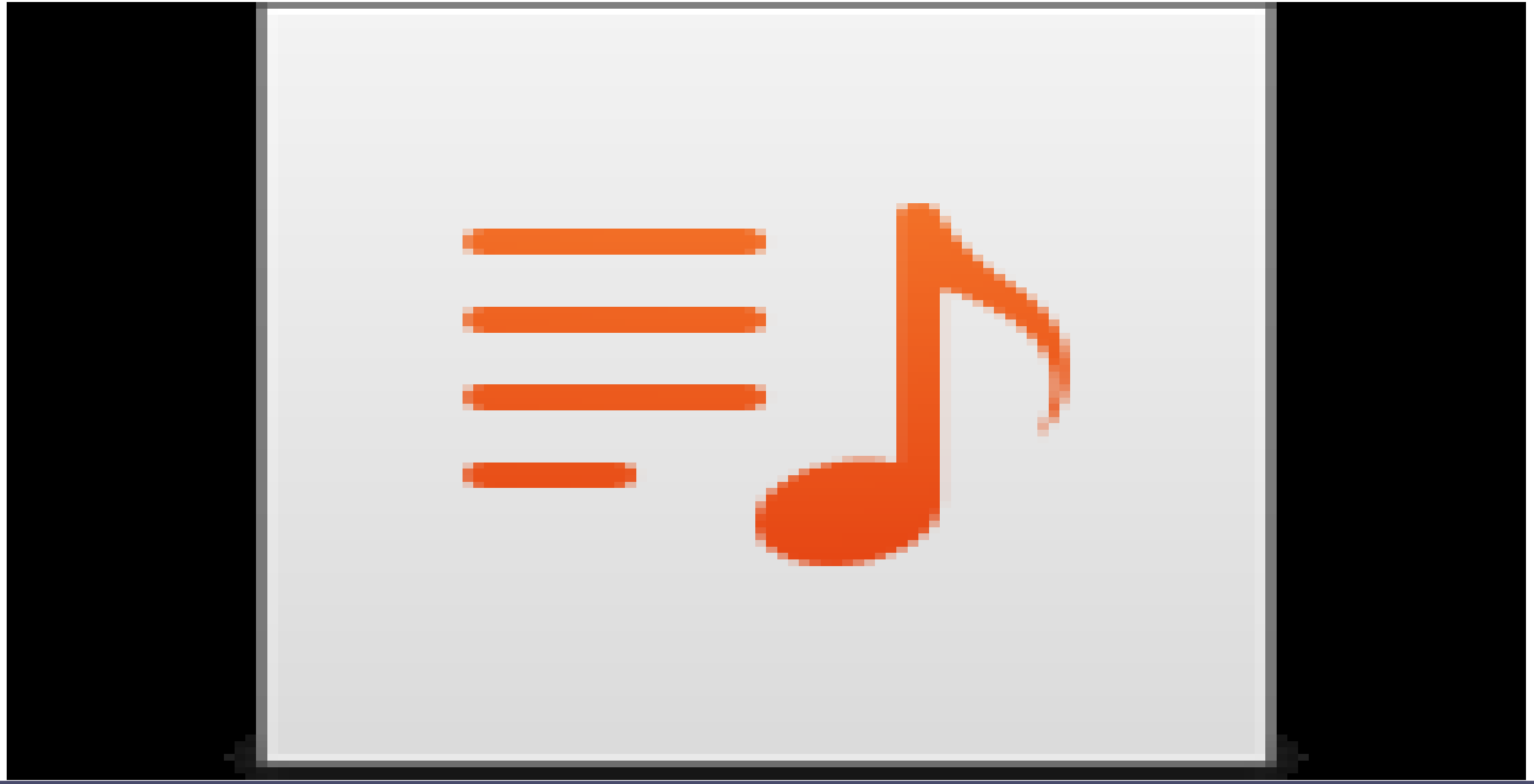
https://webis.de/data/webis-trc-12.html

Webis Text reuse Corpus 2012 (Webis-TRC-12)

Is a manually written document by a hired writers at the crowdsourcing platform oDesk.

Each document in this corpus is about 150 topics used at TREC web tracks 2009-2011.

This corpus contain 150 topics with several revision documents in each topic



Goal

Find text quality of the essay included in the corpus specially Coherence feature.

Examine the sentence is coherent to the text or not.

Looking at the coherence of the essay’s sources , And checking whether that is predictive of a sources being included.

Thank you!

A Stack of Fruits

Or: How-to Alexa

Johannes Kiesel, Bauhaus-Universität Weimar
Webis Flash Talks'21, March 12th 2021

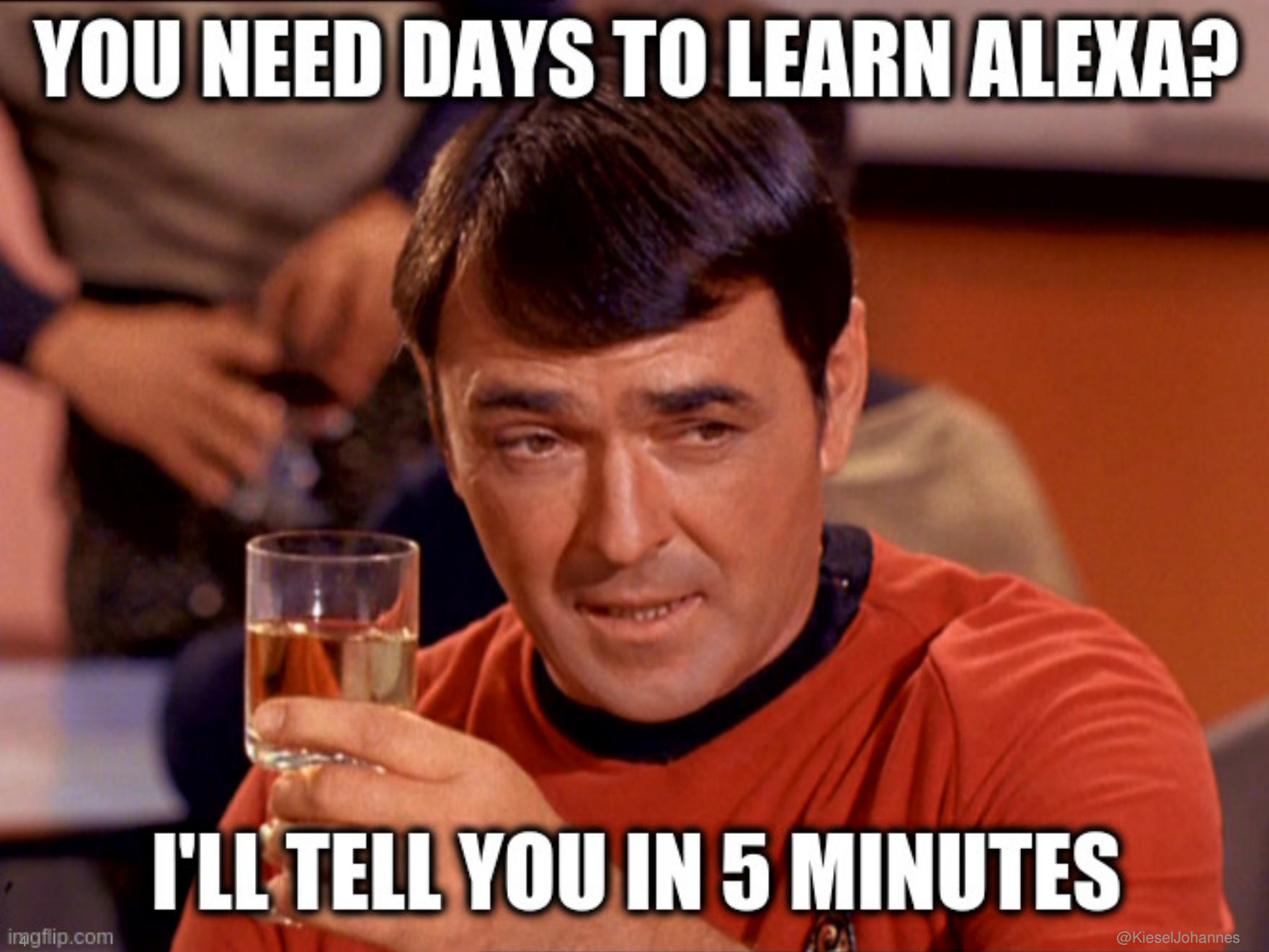
A Stack of Fruits

Or: How-to Alexa

Ademola Adewumi, Alban Bruder, Arefeh Bahrami, David Fries, Fabienne Hubricht, Fan Fan, Fera Al Sabaa, Hauke Sandhaus, Henrik Leisdon, Janek Bevendorff, Joel Arukwe, Johannes Kiesel, Kai Lorenz, Kevin Lang, Larisa Sorokina, Lars Meyer, Lucky Chandrautama, Marcel Gohsen, Maximilian Kullmann, Mohammed Udaipurwala, Roxanne El Baff, Sandy Nader, Sebastian Laverde, Sebastian Reichmann, Xiaoni Cai, Yamen Ajjour

Webis Flash Talks'21, March 12th 2021

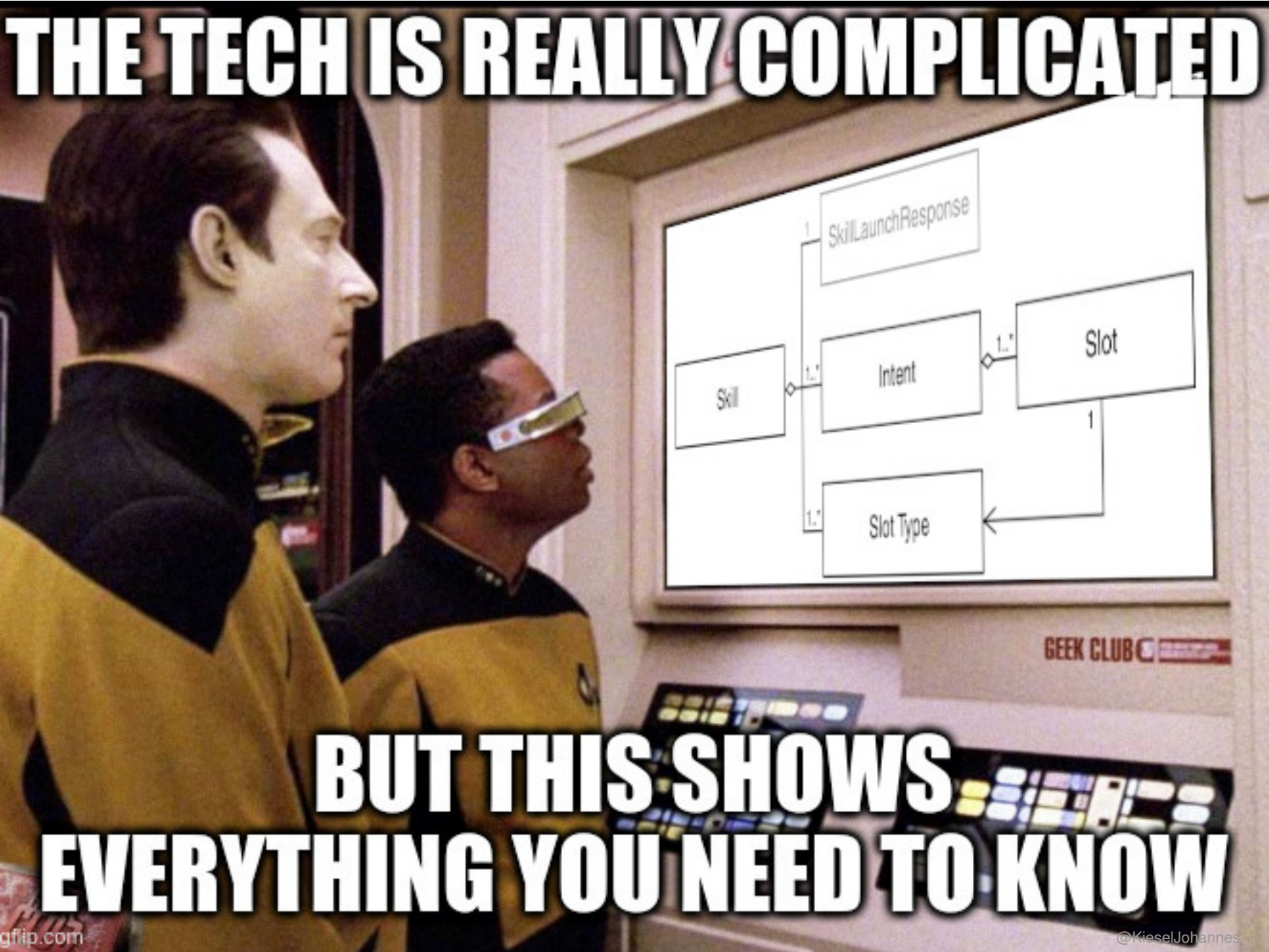

2286 vs. 1986
I QUIT SMOKING


YOU NEED DAYS TO LEARN ALEXA?
I'LL TELL YOU IN 5 MINUTES
imgflip.com
@KieselJohannes

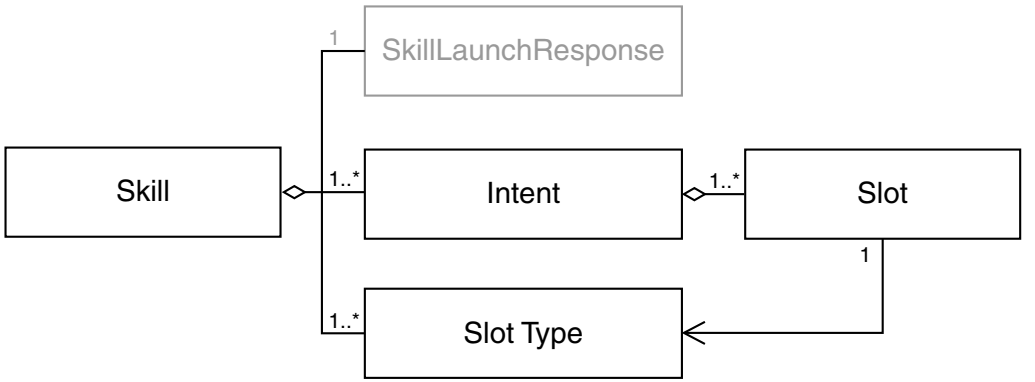
code-research / conversational-search / **alexa-example-fruit-stack**
Example skill for Alexa: a stack of fruits.

A stack of fruits

- Push() *“Add a cherry”*
- Pop() →  *“Remove the last one”*
- Clear() →  *“Eat it all”*


THE TECH IS REALLY COMPLICATED
SkillLaunchResponse
1
Skill
1..*
Intent
1..*
Slot
1
1..*
Slot Type
GEEK CLUB
BUT THIS SHOWS
EVERYTHING YOU NEED TO KNOW
gflip.com
@KieselJohannes

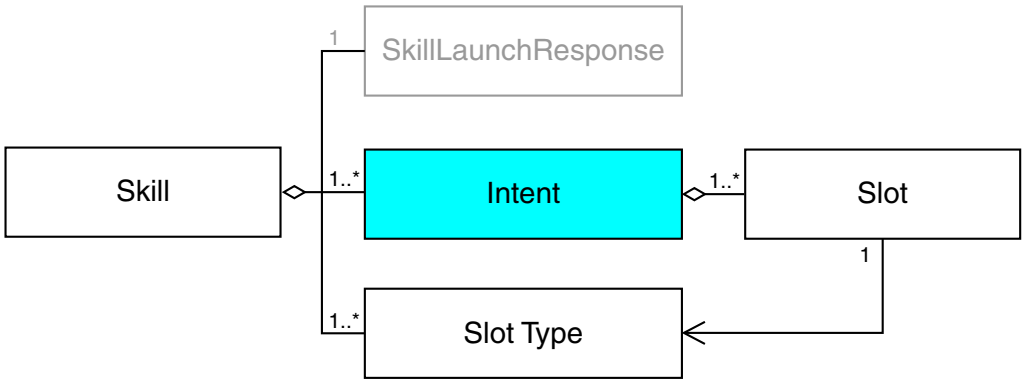
code-lib / aitools / **aitools4-commons-alexa**



Speech

Logic

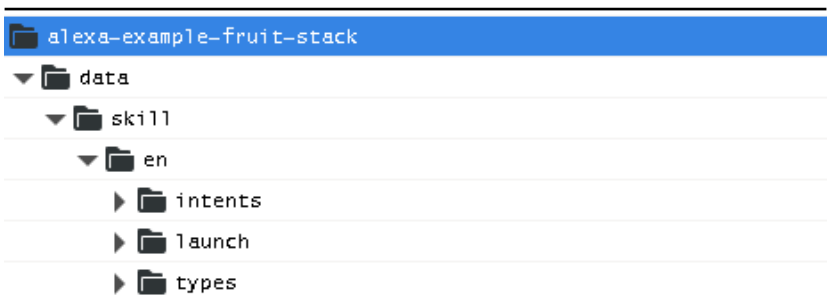
code-lib / aitools / **aitools4-commons-alexa**



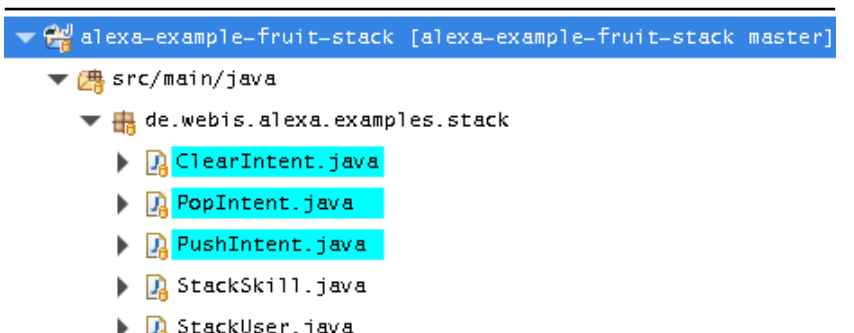
code-research / conversational-search / **alexa-example-fruit-stack**

Example skill for Alexa: a stack of fruits.

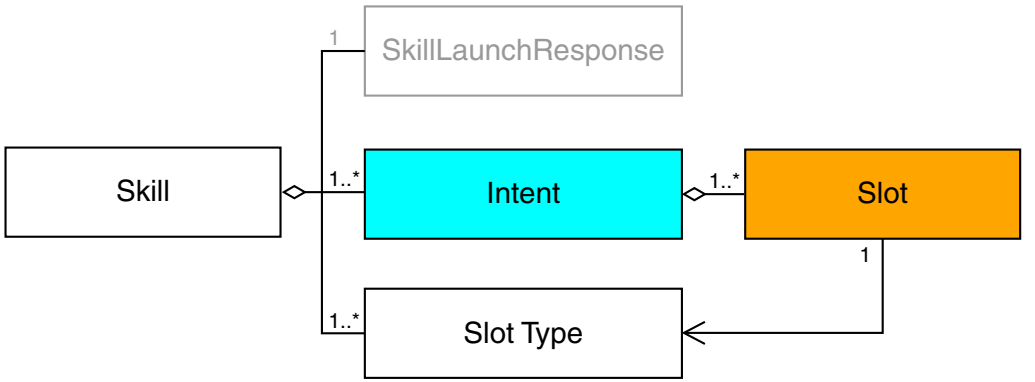
Speech



Logic



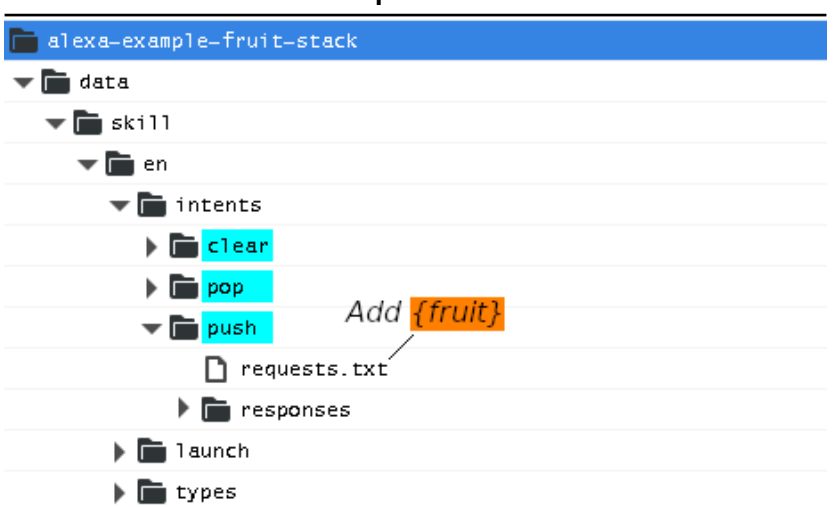
code-lib / aitools / **aitools4-commons-alexa**



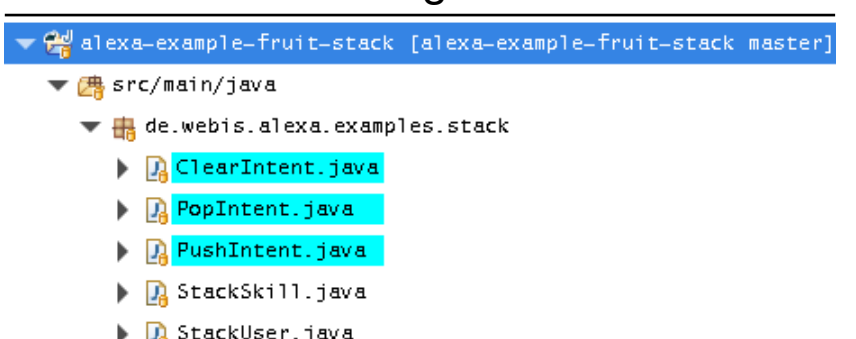
code-research / conversational-search / **alexa-example-fruit-stack**

Example skill for Alexa: a stack of fruits.

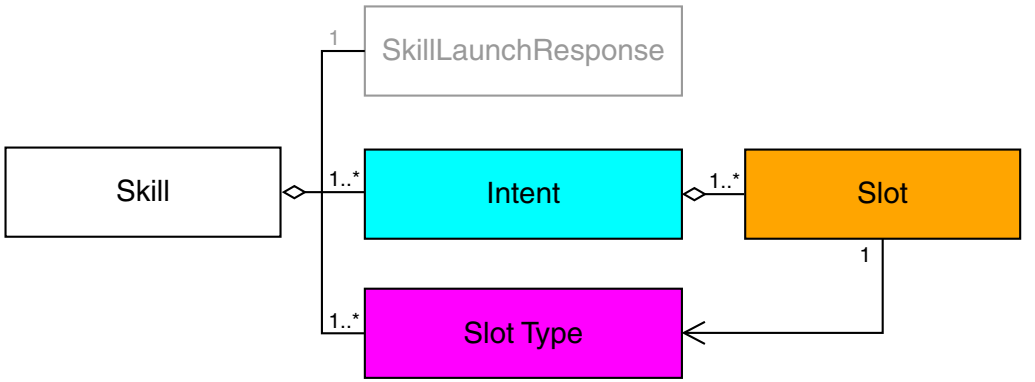
Speech



Logic



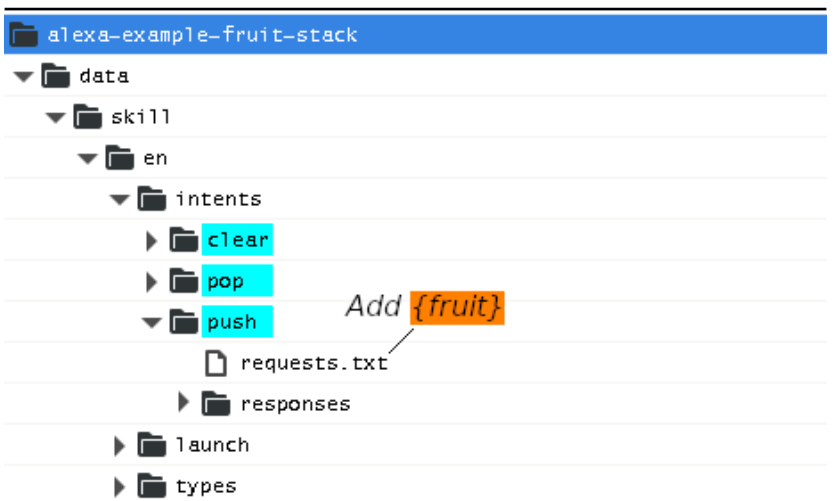
code-lib / aitools / **aitools4-commons-alexa**



code-research / conversational-search / **alexa-example-fruit-stack**

Example skill for Alexa: a stack of fruits.

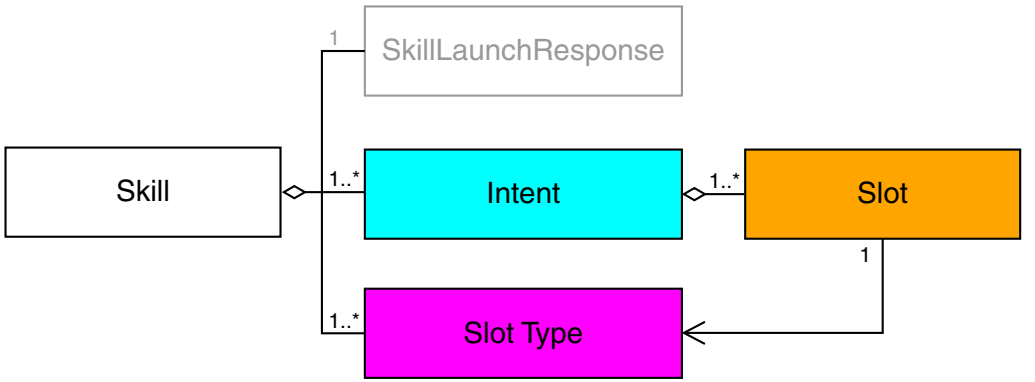
Speech



Logic

```
public class PushIntent extends GenericIntent {
  public PushIntent(...)  throws ... {
    super(new Configuration("push", ...)
        .addSlot(Slot.builder("fruit", "fruit")
            .isRequired(true))
        ...
      );
  }
}
```

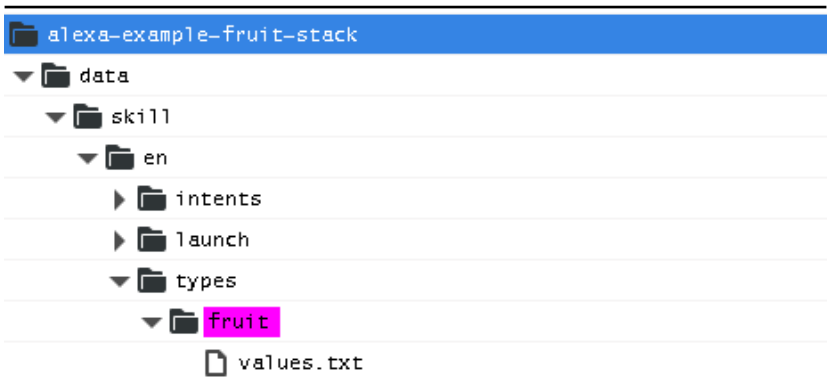
code-lib / aitools / **aitools4-commons-alexa**



code-research / conversational-search / **alexa-example-fruit-stack**

Example skill for Alexa: a stack of fruits.

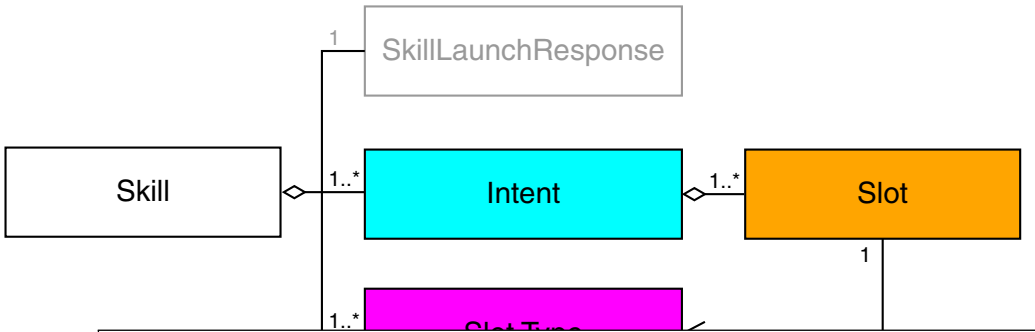
Speech



Logic

```
public class PushIntent extends GenericIntent {
  public PushIntent(...)  throws ... {
    super(new Configuration("push", ...)
        .addSlot(Slot.builder("fruit", "fruit")
            .isRequired(true))
        ...
      );
  }
}
```

code-lib / aitools / **aitools4-commons-alexa**



Spee

```
alexa-example-fruit-stack
  data
    skill
      en
        intents
        launch
        types
          fruit
            values.txt
```

Not logged in Talk Contributions Create account Log in

Page Talk Read Change Change source View history Search Wikipedia

Simple English WIKIPEDIA

List of fruits

From Simple English Wikipedia, the free encyclopedia

Fruits on this list are defined as the word is used in everyday speech. It does not include vegetables, whatever their origin.

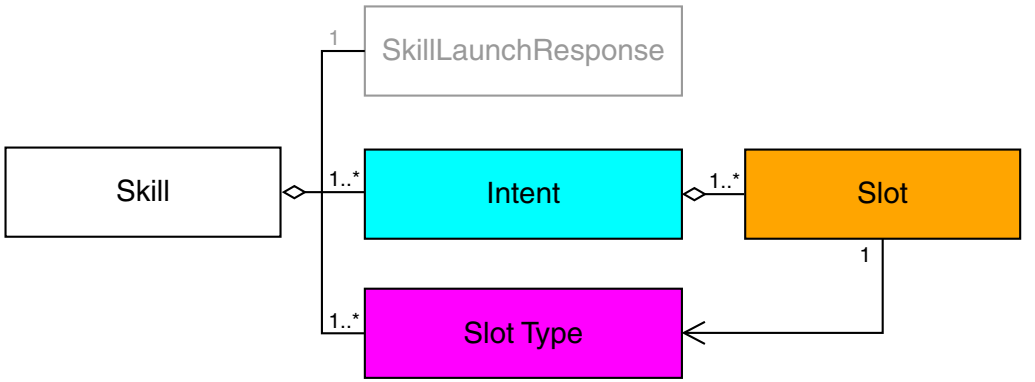
- Abiu
- Açaí
- Acerola
- Ackee
- African cucumber
- Apple
- Apricot
- Avocado
- Banana
- Bilberry
- Blackberry
- Blackcurrant
- Black sapote
- Blueberry
- Boysenberry

Main page
Simple start
Simple talk
New changes
Show any page
Help
Contact us
Give to Wikipedia
About Wikipedia

Tools

What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Cite this page
Wikidata item
Sandbox

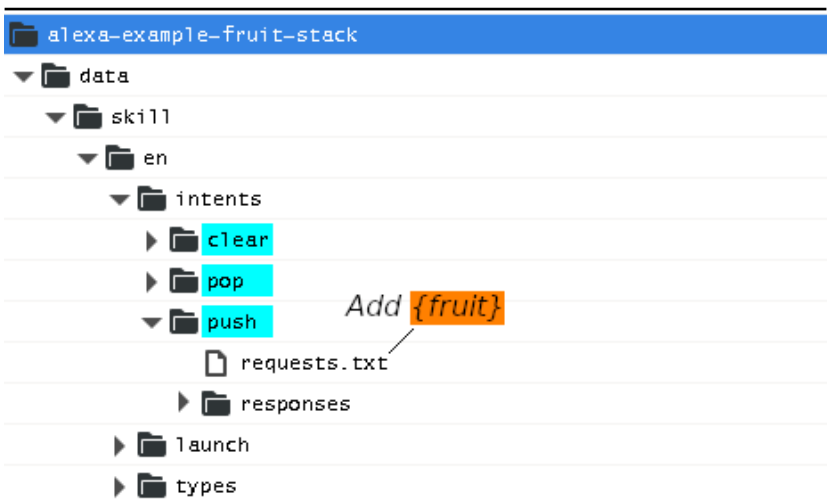
code-lib / aitools / **aitools4-commons-alexa**



code-research / conversational-search / **alexa-example-fruit-stack**

Example skill for Alexa: a stack of fruits.

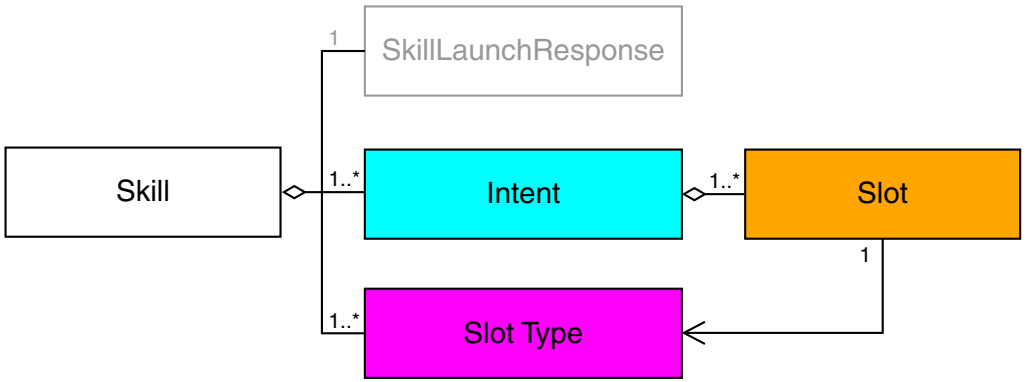
Speech



Logic

```
public class PushIntent extends GenericIntent {
  public PushIntent(...)  throws ... {
    super(new Configuration("push", ...)
        .addSlot(Slot.builder("fruit", "fruit")
            .isRequired(true))
        ...
      );
  }
}
```

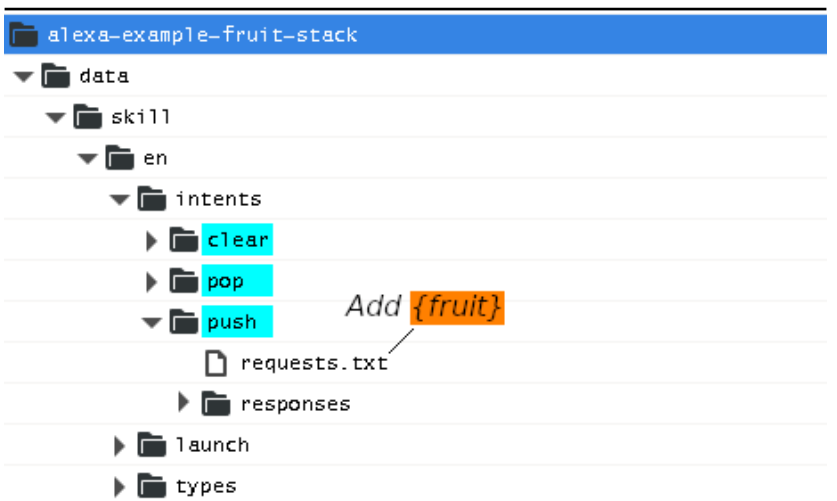
code-lib / aitools / **aitools4-commons-alexa**



code-research / conversational-search / **alexa-example-fruit-stack**

Example skill for Alexa: a stack of fruits.

Speech

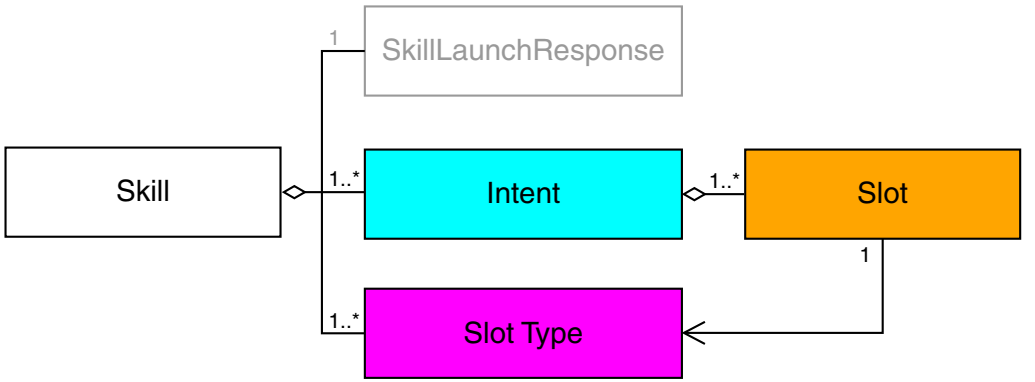


Logic

```
public class PushIntent extends GenericIntent {
  protected Response.Builder onRequest(...
        Map<String, SlotValue> slots, User user) {
    user.pushFruit(slots.get("fruit"));
```

```
  }
}
```

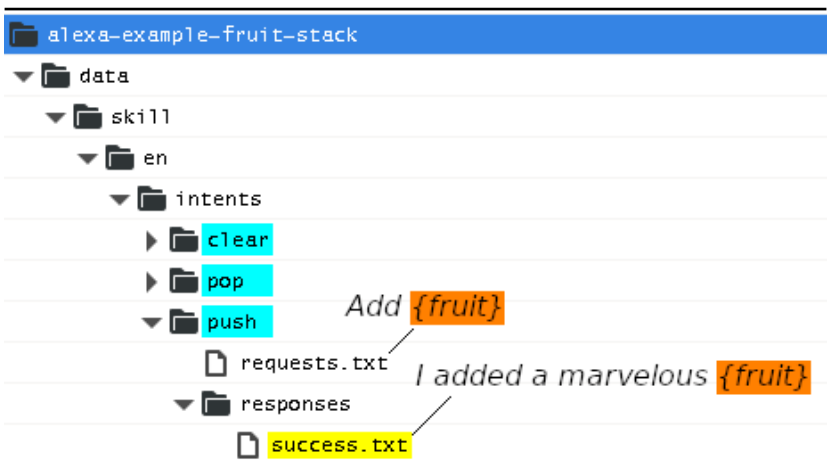
code-lib / aitools / **aitools4-commons-alexa**



code-research / conversational-search / **alexa-example-fruit-stack**

Example skill for Alexa: a stack of fruits.

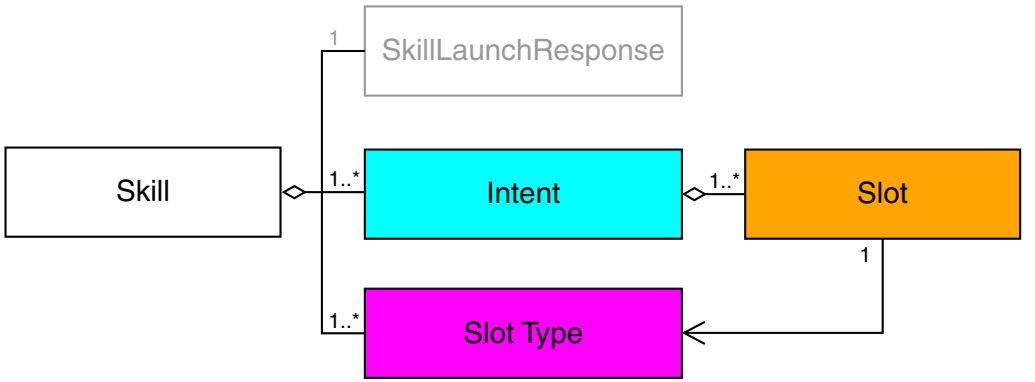
Speech



Logic

```
public class PushIntent extends GenericIntent {
  protected Response.Builder onRequest(...
        Map<String, SlotValue> slots, User user) {
    user.pushFruit(slots.get("fruit"));




  }
}
```

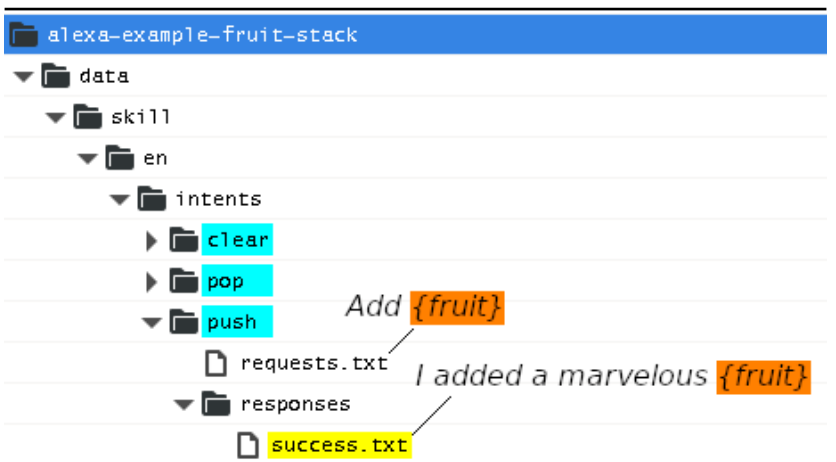
code-lib / aitools / **aitools4-commons-alexa**



code-research / conversational-search / **alexa-example-fruit-stack**

Example skill for Alexa: a stack of fruits.

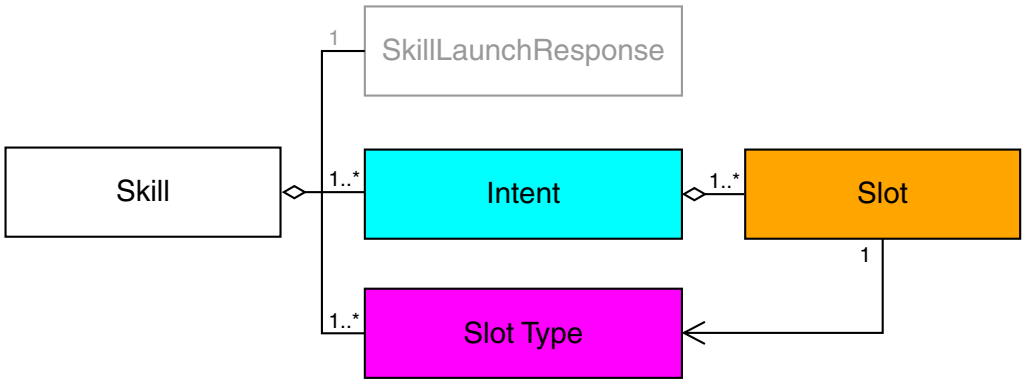
Speech



Logic

```
public class PushIntent extends GenericIntent {
  protected Response.Builder onRequest(...
        Map<String, SlotValue> slots, User user) {
    user.pushFruit(slots.get("fruit"));


    return this.respond("success", slots, ...);



  }
}
```

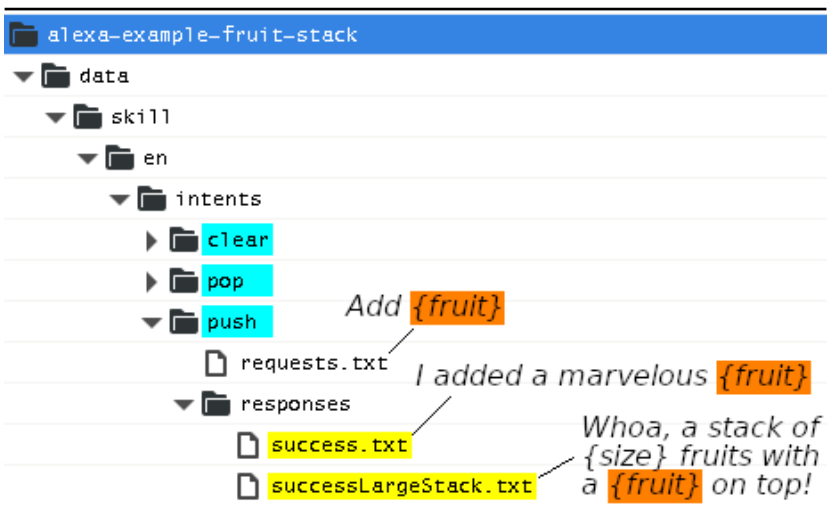
code-lib / aitools / **aitools4-commons-alexa**



code-research / conversational-search / **alexa-example-fruit-stack**

Example skill for Alexa: a stack of fruits.

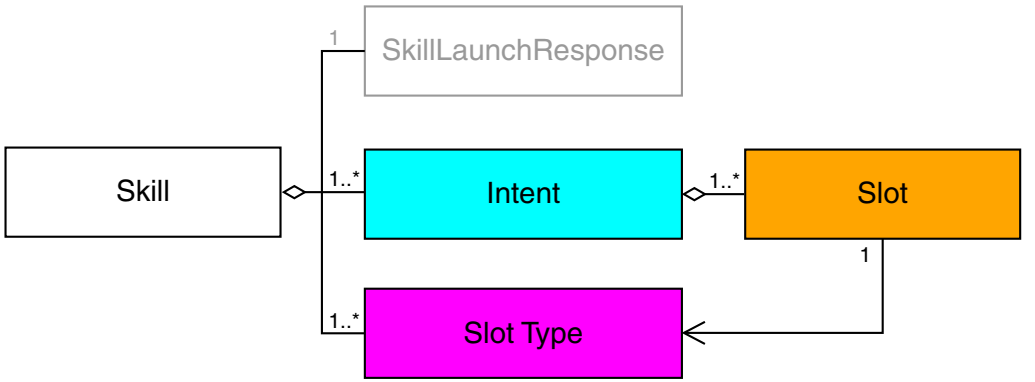
Speech



Logic

```
public class PushIntent extends GenericIntent {
  protected Response.Builder onRequest(...
        Map<String, SlotValue> slots, User user) {
    user.pushFruit(slots.get("fruit"));

    return this.respond("success", slots, ...);

  }
}
```

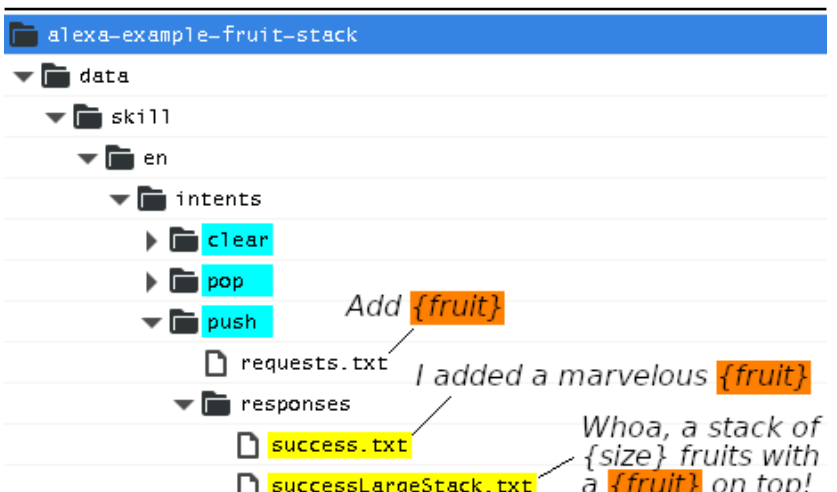
code-lib / aitools / **aitools4-commons-alexa**



code-research / conversational-search / **alexa-example-fruit-stack**

Example skill for Alexa: a stack of fruits.

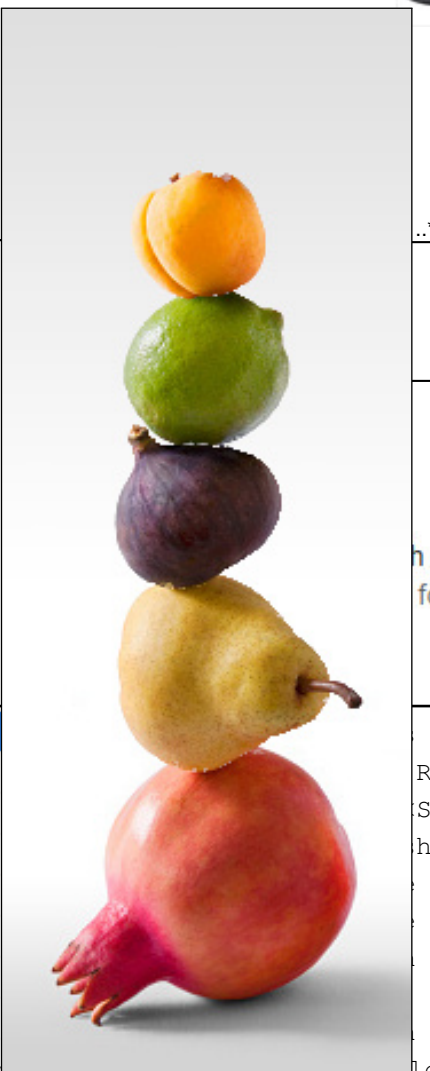
Speech



Logic

```
public class PushIntent extends GenericIntent {
  protected Response.Builder onRequest(...
        Map<String, SlotValue> slots, User user) {
    user.pushFruit(slots.get("fruit"));
    int size = user.getStack().size();
    if (size < 3) {
      return this.respond("success", slots, ...);
    } else {
      return this.respond("successLargeStack",
          slots , Map.of( "size" , size), ...);
    }
  }
}
```

code-lib / aitools / **aitools4-commons-alexa**



Skill

Slot

1

h / conversational-search / **alexa-example-fruit-stack**

for Alexa: a stack of fruits.

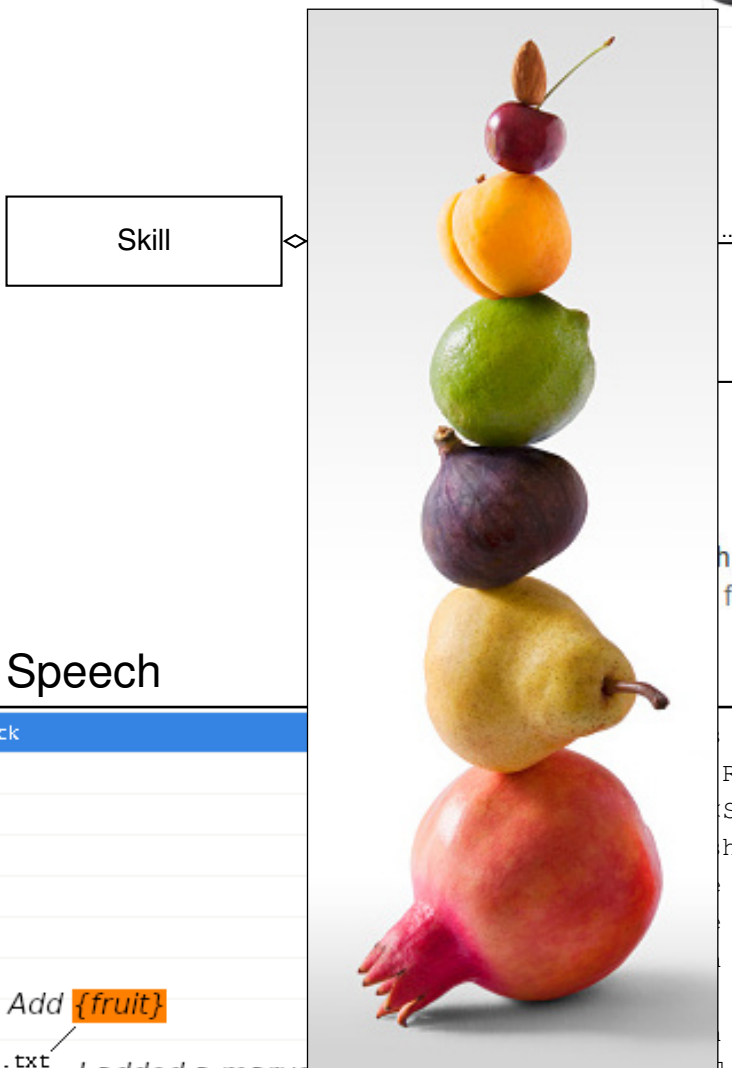
Speech

- alexa-example-fruit-stack
  - data
    - skill
      - en
        - intents
          - clear
          - pop
          - push
            - requests.txt — *Add {fruit}*
          - responses
            - success.txt — *I added a marvelous {fruit}*
            - successLargeStack.txt — *Whoa, a stack of {size} fruits with a {fruit} on top!*

Logic

```
PushIntent extends GenericIntent {
Response.Builder onRequest(...
String, SlotValue> slots, User user) {
hFruit(slots.get("fruit"));
 = user.getStack().size();
 < 3) {
 this.respond("success", slots, ...);

 this.respond("successLargeStack",
lots , Map.of("size", size), ...);
}
}
}
```

code-lib / aitools / **aitools4-commons-alexa**



Skill

Slot

1

h / conversational-search / **alexa-example-fruit-stack**

for Alexa: a stack of fruits.

Speech

```
alexa-example-fruit-stack
data
  skill
    en
      intents
        clear
        pop
        push
          requests.txt
          responses
            success.txt
            successLargeStack.txt
```

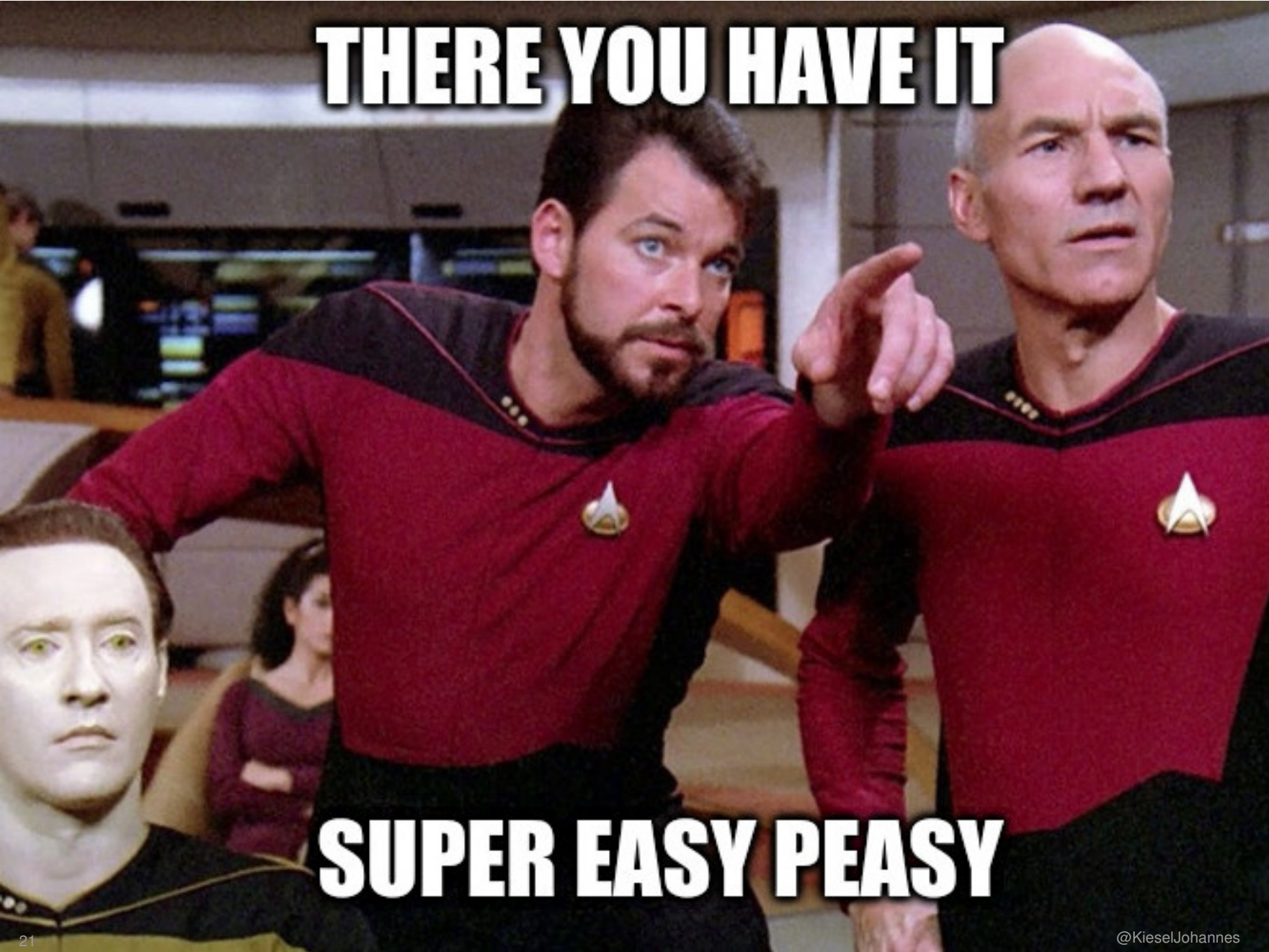
Add {fruit}

I added a marvelous {fruit}

Whoa, a stack of {size} fruits with a {fruit} on top!

Logic

```
PushIntent extends GenericIntent {
Response.Builder onRequest(...
String, SlotValue> slots, User user) {
hFruit(slots.get("fruit"));
 = user.getStack().size();
 < 3) {
 this.respond("success", slots, ...);

 this.respond("successLargeStack",
lots , Map.of("size", size), ...);
}
}
}
```


THERE YOU HAVE IT
SUPER EASY PEASY

Documentation

The core of each skill's *logic* is the Skill, which you define through extending the `Skill.Builder` class. [example]

Each skill's *speech* is located in `data/skill/locale/` (for the respective locale). `locale` may either be a complete locale (e.g., `en-us`) or just the language part (e.g., `en`), in which case the corresponding localized skill serves all countries for that language. [example]

See how to add a skill launch response and how to add an intent to start building your skill. See how to run the skill server to get it running.

Contents

project-template-alexa

Project ID: 2331

Star 0 Fork 0

45 Commits **1** Branch **0** Tags **358 KB** Files **358 KB** Storage

Template for Alexa projects (extends project-template-java)

Read more

master project-template-alexa / History Find file Web IDE Clone

Alexa Project Setup

This project explains the basic workflow for setting up Alexa projects and dependency management, and serves as a simple example of implementing this workflow. Those unfamiliar with Alexa are advised to read the documentation, look up unfamiliar terminology, and ask your supervisor and fellow students.

Practice Tasks

These tasks use the fruit-stack example skill. First download the repository.

1. Environment setup. Make sure you have the ASK-CLI installed and that you can compile with `./gradlew shadowjar` and run with `java -jar build/libs/project-template-alexa-0.1.0-all.jar`: You should see `error: No application provided` and usage information for the "Aitools Alexa CLI".
2. Skill setup. Run the skill setup for a new configuration: `data/skill/alexa-example-fruit-stack-<your-name>.conf`. Invocation name for `en_US`: `<your-name> fruits`. Use this configuration in the following.

Command line interface

Built-in and the same for each skill that uses  code-lib / aitools / **aitools4-commons-alexa**

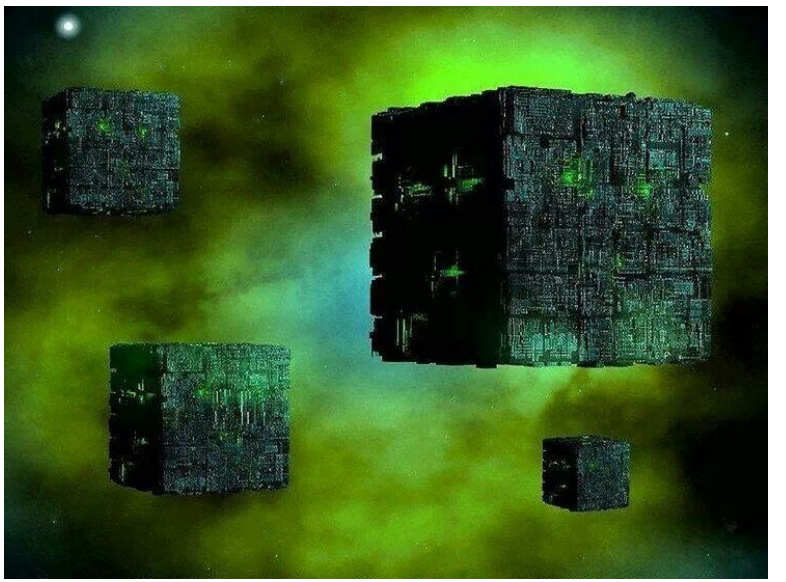
```
$ java -jar build/libs/alexa-example-fruit-stack-0.1.0-all.jar

usage:  <application> [<application specific options>]

Aitools Alexa CLI
Applications:
  chat       Starts an interactive chat with the skill.
  configure  Creates or updates the skill configuration on the Amazon
             server and the -skill file.
  delete     Deletes the configuration file and the skill in the Amazon
             Developer Console for each locale.
  deploy     Guides through the Kubernetes deployment.
  serve      Runs a server that serves as endpoint for Alexa.
  update     Compiles the interaction model and uploads it to the Amazon
             Developer Console for each locale.
```

Kubernetes deployment

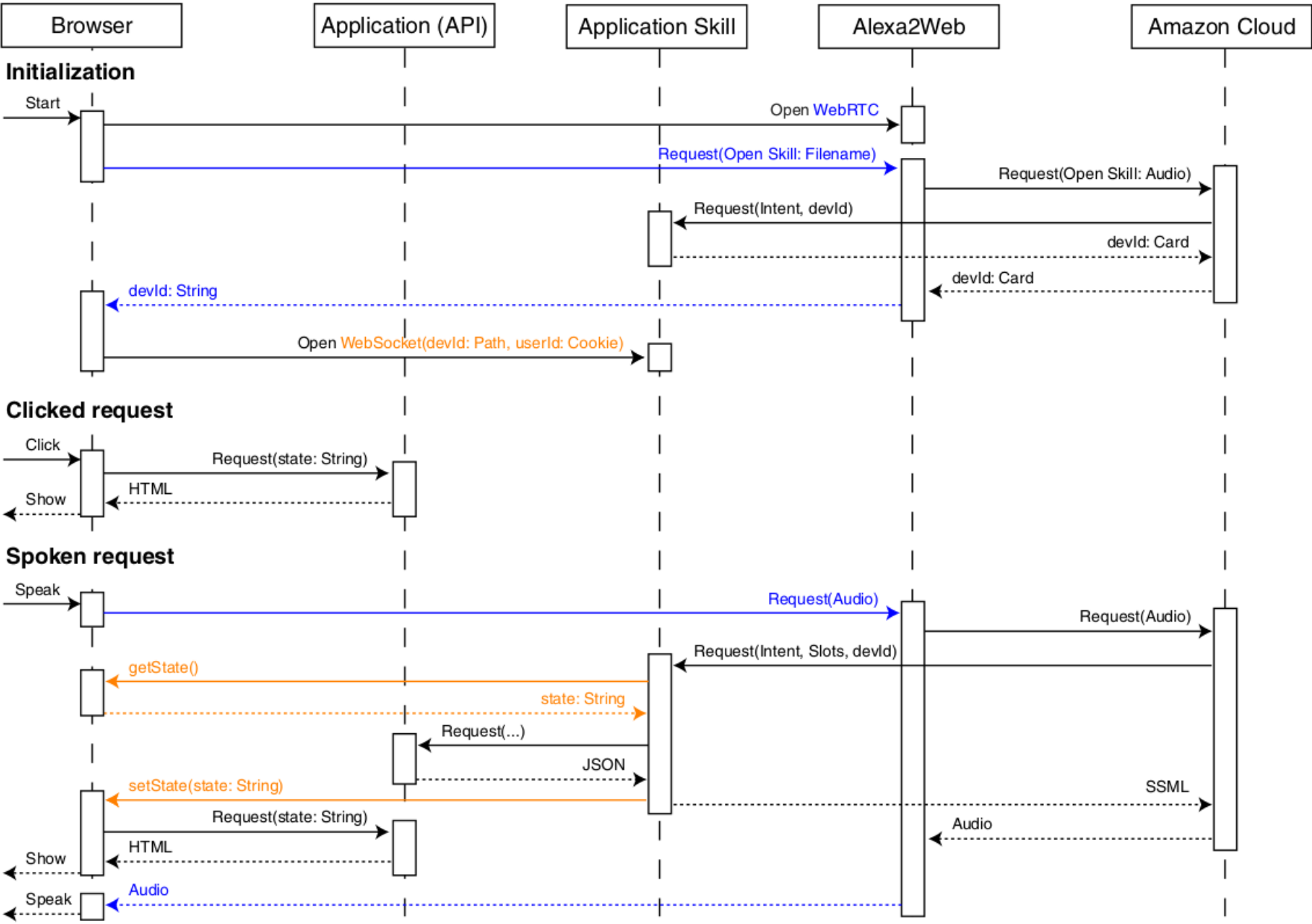
Guided deployment per `deploy`

Alexa skill server container

→



Our cloud


Browser
Application (API)
Application Skill
Alexa2Web
Amazon Cloud
Initialization
Start
Open WebRTC
Request(Open Skill: Filename)
Request(Open Skill: Audio)
Request(Intent, devId)
devId: Card
devId: Card
devId: String
Open WebSocket(devId: Path, userId: Cookie)
Clicked request
Click
Request(state: String)
HTML
Show
Spoken request
Speak
Request(Audio)
Request(Audio)
Request(Intent, Slots, devId)
getState()
state: String
Request(...)
JSON
setState(state: String)
SSML
Request(state: String)
Audio
HTML
Show
Speak
Audio

Your Fruits

- peach
- lemon
- fig
- pear
- pomegranate

cherry Add!

Click on the fruits to remove them.

Session

Start Stop Talk

What documents are needed to close an open-pit mine?

Lydia Müller
Universität Leipzig/InfAI e.V - Sardine

Open-pit Mine Espenhain

The Data

- Documentation on (closed) open-pit mines from LMBV
- > 40k scanned documents with georeferences
- Some Metadata: Size, original filename, description, sometimes tags
- Quality of metadata is bad
- Quality of the scanned documents varies

Aim and Solution

Aim: Prototypical documentation for all types of georeferences

Solution:

- Classify documents into 11 document classes inspired by the description and tags
- Aggregate all classes for the different types of georeferences and call that prototypical documentation
- Find georeferences with missing documentation

Does it work?

It‘s going in the right direction!

Problems:

- LMBV is not sure about the document types
- Metadata is heterogenous, over-specific or under-specific

Approch:

- Define multi-label document classes based on document content
  - Topics from topic modelling
  - Train a text classifier

Differential Bias

Alonso Palomino

Bias

[1] https://www.allindata.org/bias-in-big-data-implications-for-multi-sector-data-sharing/

Is this biased?

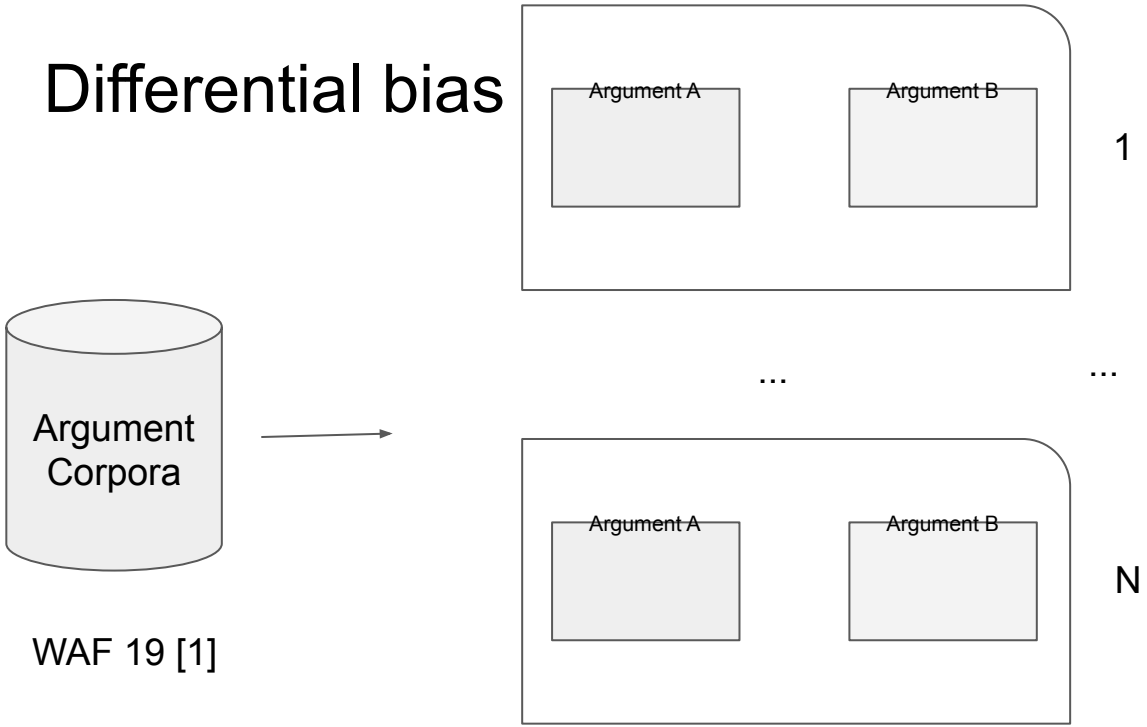
Zoos are not capable of sustaining all endangered species. According to the world conservation union which keeps records of endangered species, there are 5428 threatened animals on a recent 'red list'. yet the IUCN says that even if the world's zoos pooled their resources, they could only expect to sustain about 2000 species in captivity.

Which is more biased?

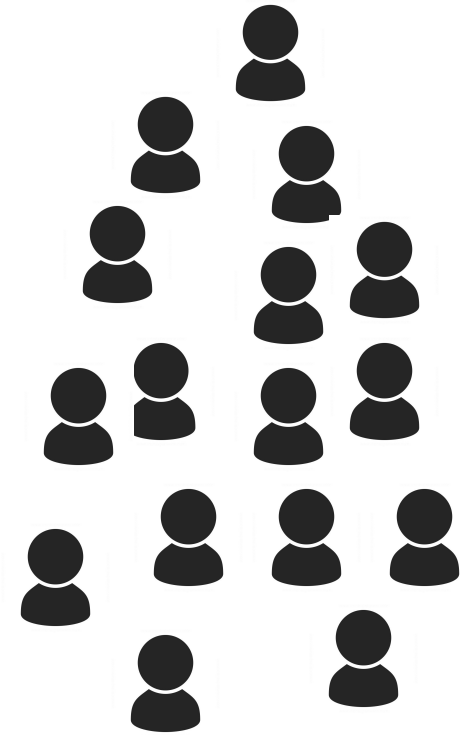
Banning Zoos

Argument A	Argument B
Zoos are not capable of sustaining all endangered species. According to the world conservation union which keeps records of endangered species, there are 5428 threatened animals on a recent 'red list'. yet the IUCN says that even if the world's zoos pooled their resources, they could only expect to sustain about 2000 species in captivity. Most of the animals that you see in zoos aren't endangered. While some argue that zoos are a means to protecting endangered species, the reality is that very few animals in zoos are actually endangered. In other words, this is really not the reason why zoos exist and so should not be put forward as a justification for them.	Zoos can raise awareness of endangered species. Visitors to zoos may raise their awareness of endangered species by being directly exposed to them. If nature was appropriately preserved, we would not need zoos. Michael fox, sierra, november-december 1990 - "Zoos are becoming facsimiles - or perhaps caricatures - of how animals once were in their natural habitat. if the right policies toward nature were pursued, we would need no zoos at all."

Differential bias



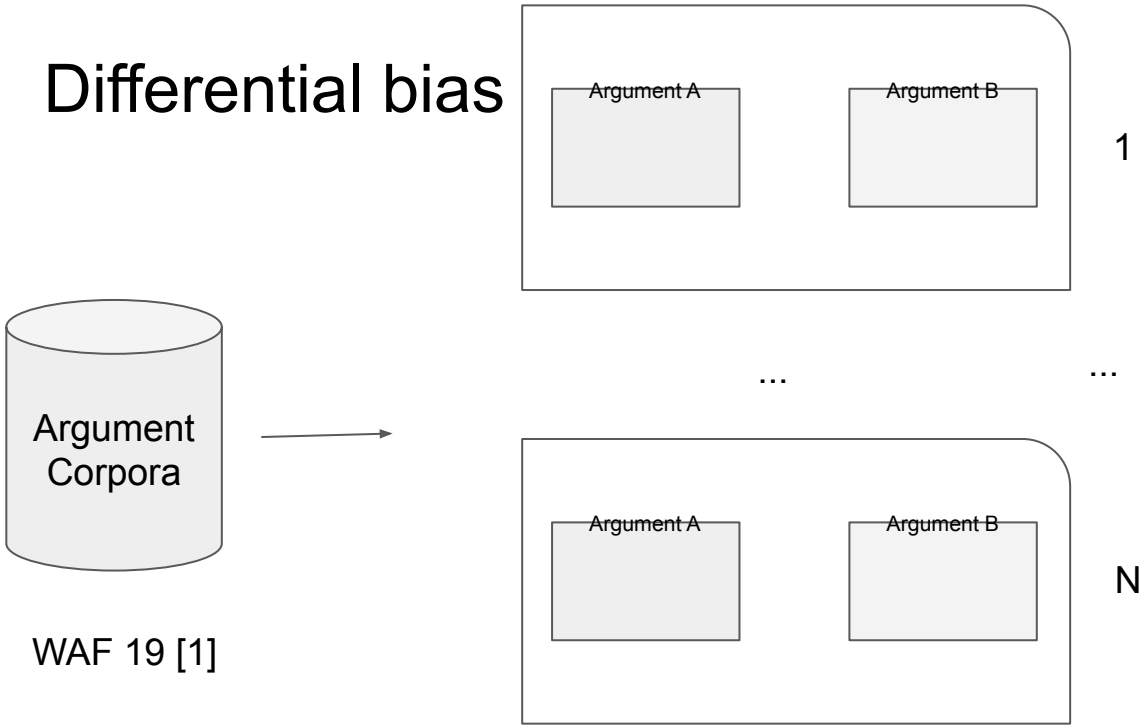
WAF 19 [1]



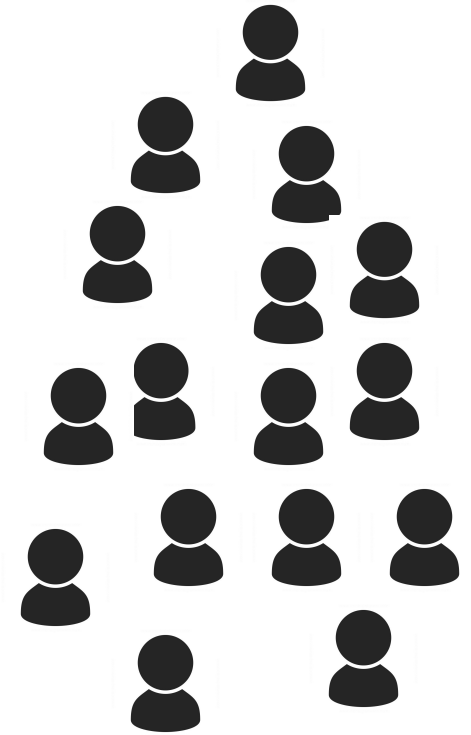
Large scale crowdsourcing labeling

[1] Ajjour, Y., Alshomary, M., Wachsmuth, H., & Stein, B. (2019, November). Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2915-2925).

Differential bias



WAF 19 [1]



Large scale crowdsourcing labeling

[1] Ajjour, Y., Alshomary, M., Wachsmuth, H., & Stein, B. (2019, November). Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2915-2925).

Which one is more biased? (stance)

Banning Zoos

Argument A	Argument B
Zoos are not capable of sustaining all endangered species. According to the world conservation union which keeps records of endangered species, there are 5428 threatened animals on a recent 'red list'. yet the IUCN says that even if the world's zoos pooled their resources, they could only expect to sustain about 2000 species in captivity. Most of the animals that you see in zoos aren't endangered. While some argue that zoos are a means to protecting endangered species, the reality is that very few animals in zoos are actually endangered. In other words, this is really not the reason why zoos exist and so should not be put forward as a justification for them.	Zoos can raise awareness of endangered species. Visitors to zoos may raise their awareness of endangered species by being directly exposed to them. If nature was appropriately preserved, we would not need zoos. Michael fox, sierra, november-december 1990 - "Zoos are becoming facsimiles - or perhaps caricatures - of how animals once were in their natural habitat. if the right policies toward nature were pursued, we would need no zoos at all."

Which one is more biased? (stance)

Banning Zoos

Argument A	Argument B
Zoos are not capable of sustaining all endangered species. According to the world conservation union which keeps records of endangered species, there are 5428 threatened animals on a recent 'red list'. yet the IUCN says that even if the world's zoos pooled their resources, they could only expect to sustain about 2000 species in captivity. **[Pro]** Most of the animals that you see in zoos aren't endangered. While some argue that zoos are a means to protecting endangered species, the reality is that very few animals in zoos are actually endangered. In other words, this is really not the reason why zoos exist and so should not be put forward as a justification for them. **[Pro]**	Zoos can raise awareness of endangered species. Visitors to zoos may raise their awareness of endangered species by being directly exposed to them. **[Pro]** If nature was appropriately preserved, we would not need zoos. Michael fox, sierra, november-december 1990 - "Zoos are becoming facsimiles - or perhaps caricatures - of how animals once were in their natural habitat. if the right policies toward nature were pursued, we would need no zoos at all." **[Con]**

Which one is more biased? (frame)

Banning Zoos

Argument A	Argument B
Zoos are not capable of sustaining all endangered species. According to the world conservation union which keeps records of endangered species, there are 5428 threatened animals on a recent 'red list'. yet the IUCN says that even if the world's zoos pooled their resources, they could only expect to sustain about 2000 species in captivity. Most of the animals that you see in zoos aren't endangered. While some argue that zoos are a means to protecting endangered species, the reality is that very few animals in zoos are actually endangered. In other words, this is really not the reason why zoos exist and so should not be put forward as a justification for them.	Zoos can raise awareness of endangered species. Visitors to zoos may raise their awareness of endangered species by being directly exposed to them. If nature was appropriately preserved, we would not need zoos. Michael fox, sierra, november-december 1990 - "Zoos are becoming facsimiles - or perhaps caricatures - of how animals once were in their natural habitat. if the right policies toward nature were pursued, we would need no zoos at all."

Which one is more biased? (frame)

Banning Zoos

Argument A	Argument B
Zoos are not capable of sustaining all **endangered** species. According to the world conservation union which keeps records of **endangered** species, there are 5428 threatened animals on a recent 'red list'. yet the IUCN says that even if the world's zoos pooled their resources, they could only expect to sustain about 2000 species in captivity. Most of the animals that you see in zoos aren't **endangered**. While some argue that zoos are a means to protecting **endangered** species, the reality is that very few animals in zoos are actually **endangered**. In other words, this is really not the reason why zoos exist and so should not be put forward as a justification for them.	Zoos can raise awareness of **endangered** species. Visitors to zoos may raise their awareness of **endangered** species by being directly exposed to them. If nature was appropriately preserved, we would not need zoos. Michael fox, sierra, november-december 1990 - "Zoos are becoming facsimiles - or perhaps caricatures - of how animals once were in their **natural** habitat. if the right policies toward **nature** were pursued, we would need no zoos at all."

Conclusions

- We proposed a new kind of experiments to measure bias in natural language:
  - **Differential bias:** As bias is not an absolute characteristic but the result of comparing more than a single proposition. To determine whether a premise is biased or not, it is necessary to make a comparative judgment.
  - We operationalize the concept of "differential bias" in the domain of computational argumentation.
  - The previous examples show the scenarios we are using to investigate bias in text at different human detectability levels.

- Goal:
  - With the proposed experiments, we prove that when information systems consider in their design tools that facilitate users to uncover what can be easily overlooked, annotators can differentiate biased narratives easier, helping them form an individual opinion or stance about a topic.

References

1. Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. Modeling Frames in Argumentation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP 2019), pages 2922-2932, November 2019. ACL.
2. Van Laar, J. A. (2007). One-sided arguments. *Synthese*, *154*(2), 307-327.
3. Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013, August). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1650-1659).

Style Transfer

Formal2Informal & Informal2Formal tasks

Wei-Fan Chen

Presenting: Juela Palushi

Paderborn University

March 12, 2021

Overview

01. Introduction

Overview about the Formal to Informal & vice versa tasks.

02. Dataset

Details about the dataset used for fine-tuning, validating and test.

03. Model

Information about the model used during fine-tuning.

04. Metrics

Details on the ROUGE and BLEU metrics.

.

05. Results

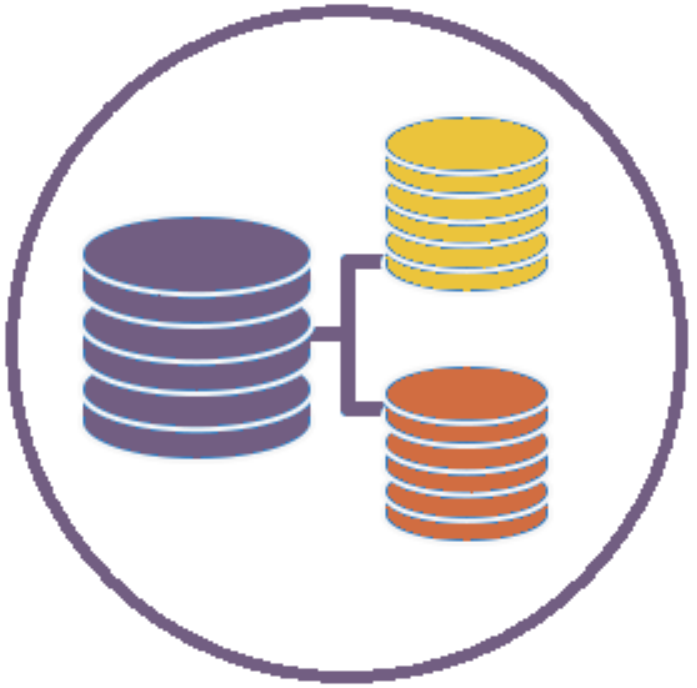
Examples about predicted entries on both fine-tuned models

Introduction

- Two fine-tuned models:
    - Formal2Informal(F2I)
    - Informal2Formal(I2F)

Dataset

- **GYAFC** - Grammarly's Yahoo Answers Formality Corpus
- Data from two domains:
  - Entertainment Music
  - Family Relationships
- Both domains provide entries for:
  - Training: 104K+ entries in total
  - Validation: 5K+ entries in total
  - Test: 2K+ entries in total



Model

- **BART** – from Facebook Research
- Seq2Seq model used for:
  - Natural Language Generation
  - Natural Language Translation
  - Natural Language Comprehension
- Utilized pretrained “*BART base*” model of size ~1GB
- Transformers package from Hugging Face company

Metrics

ROUGE

- ROUGE-1
  - F2I – 0.46 (F-score)
  - I2F – 0.63 (F-score)
- ROUGE-2
  - F2I – 0.26 (F-score)
  - I2F – 0.46 (F-score)

BLEU

- BLEU-1
  - F2I – 0.39
  - I2F – 0.59
- BLEU-2
  - F2I – 0.22
  - I2F – 0.42

Results

Formal to Informal

I enjoy watching my companion attempt to role-play with them.

Prediction: I love watching my friend try to role play with them.

Target: lol i love watchin my lil guy try to act out the things wiht them

Are you posing a rhetorical question?

Prediction: are you asking a rhetorical question?

Target: Sounds like a rhetorical question :)

Informal to Formal

For one thing if it doesn't work out there goes your job.

Prediction: If it does not work out, you lose your job.

Target: For one thing, there goes your job if it does not work out.

I want to be on TV!

Prediction: I would like to be on television.

Target: I would like to be on television.

Thank you

Sampling Bias Due to Near-Duplicates in Learning to Rank

Bachelor's Thesis & SIGIR '20 Paper

Jan Heinrich Reimer
jan.reimer@student.uni-halle.de

Supervisor: Maik Fröbe

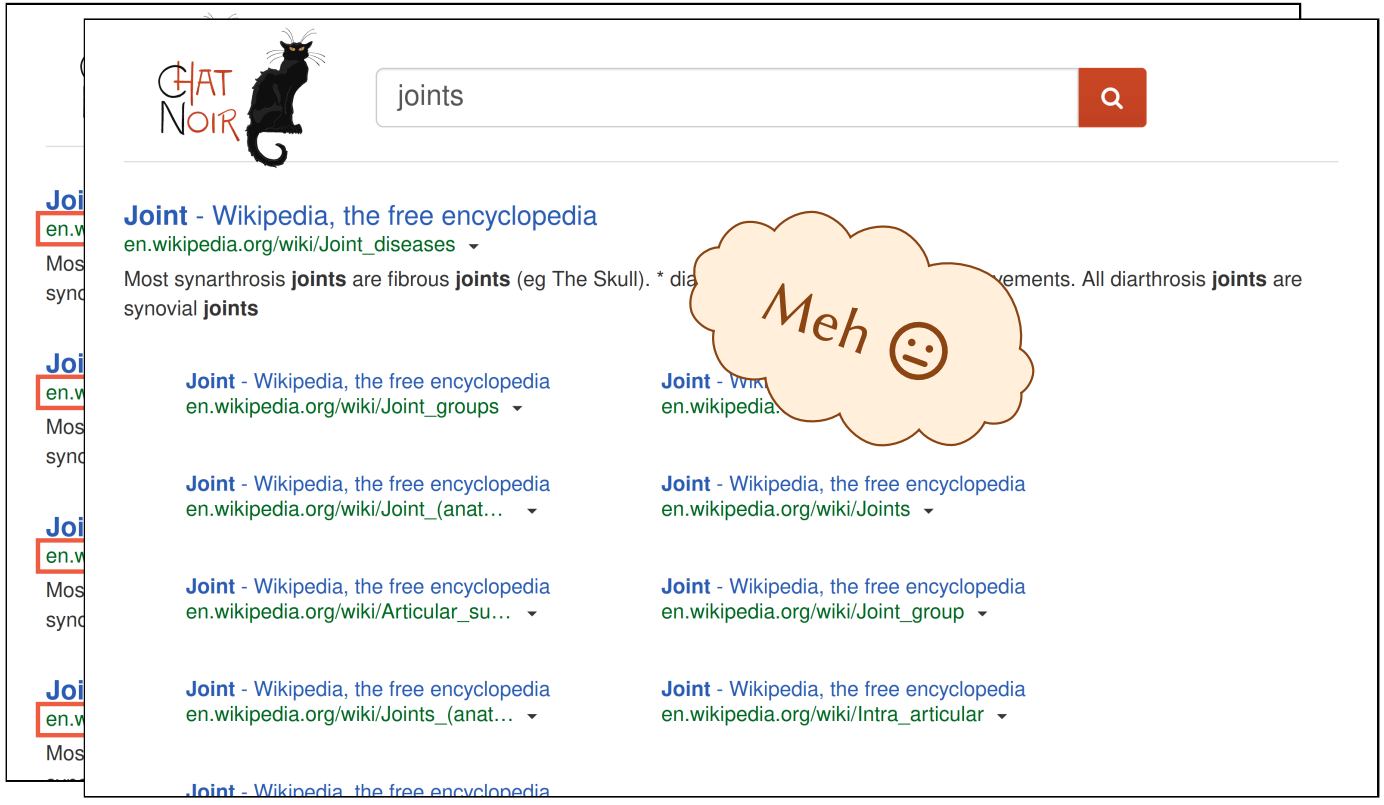
Martin Luther University Halle-Wittenberg

March 12, 2021

Have you been there?

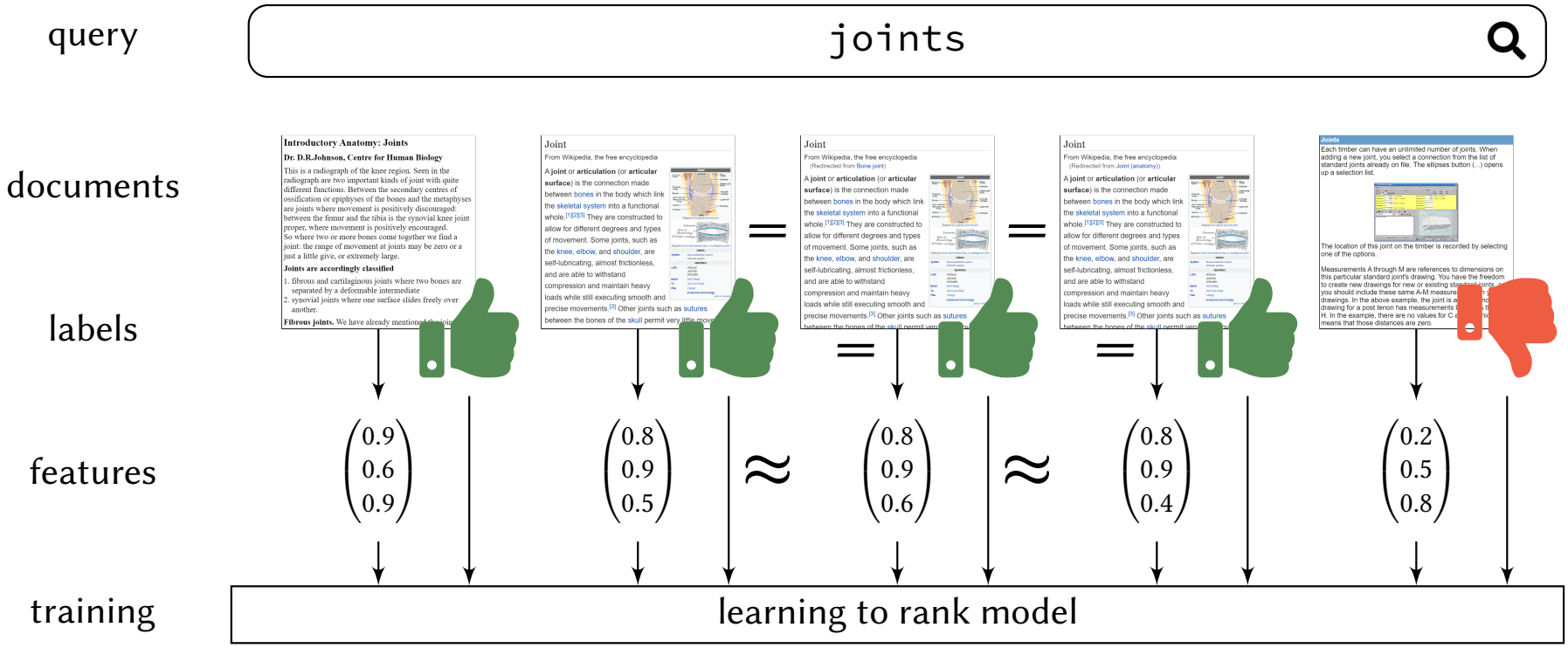
- redundant search results at top ranks

Have you been there?



- redundant search results at top ranks

What's the trouble with Learning to Rank?

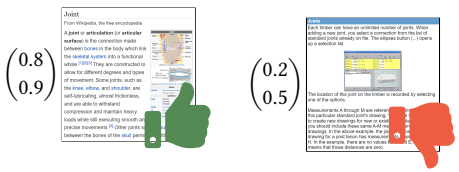


1. identical relevance labels (Cranfield paradigm)
2. similar features, e.g., same TF/IDF
3. oversampling → double impact on loss → overfitting

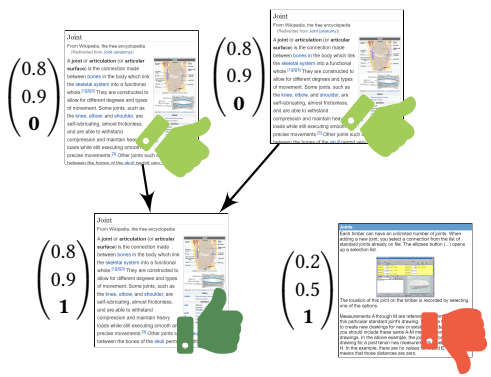
Can we do anything about it?

- reuse methods for counteracting overfitting → undersampling
- canonical link relations [OK12]

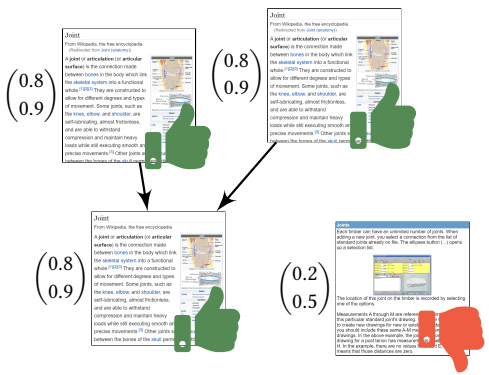
Remove



Discount & flag



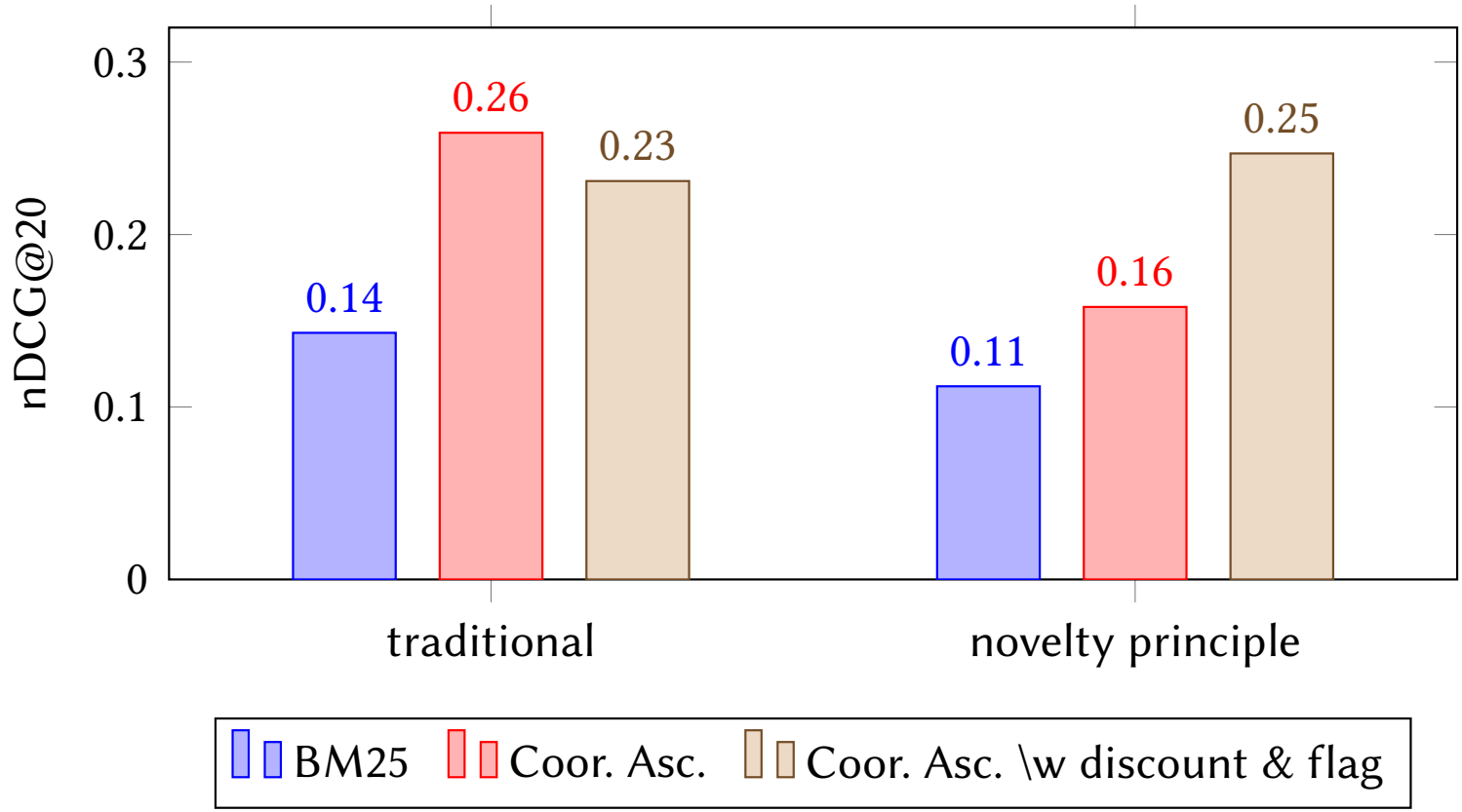
No deduplication



- removing discards training data
- discounting breaks label consistency
- …but works best

How bad is it? Does deduplication work?

Performance for Coordinate Ascent [MC07] on ClueWeb09



- performance decreases under novelty principle [Frö+20]
- discount & flag compensates impact

Conclusion

- near-duplicates reduce retrieval performance in LTR
- **De-duplicate your learning-to-rank training data!**

SIGIR '20 paper

DOI: 10.1145/3397271.3401212

Thank you!

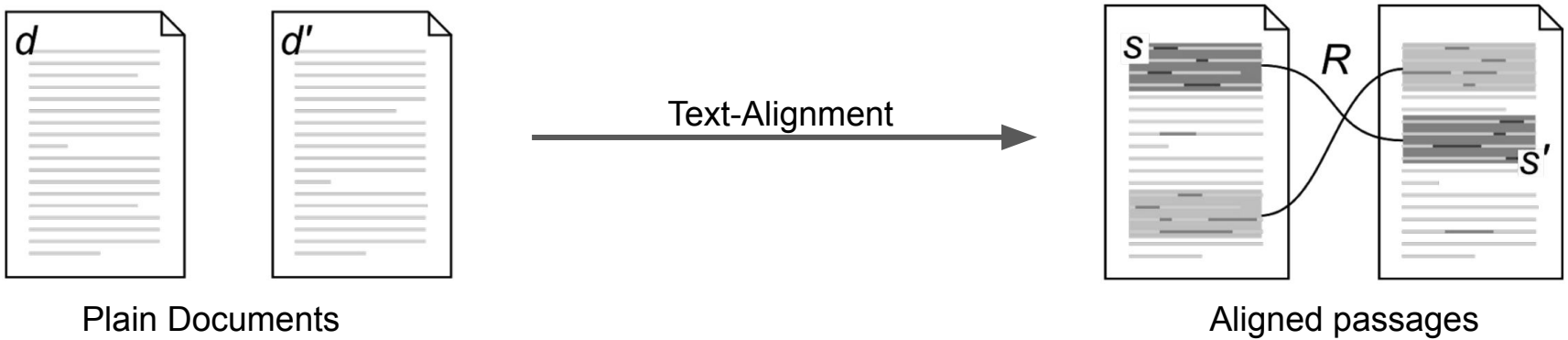
PLAGIARISM Detection

Student
Ankit Satpute

Supervisors
Michael Völske
Matti Wiegmann


PLAGIARISM
*Text-Alignment
YOU SHALL NOT PASS !

What is Text-Alignment?



Available Corpora

- Small scale
  - Answer pairs
- Large scale
  - PAN datasets

Vroniplag

VroniPlag Wiki

Eine kritische Auseinandersetzung mit der Dissertation von Prof. Dr. Diana Timova: Konzeptualisieren und Verbalisieren von Raum – kognitive und sprachliche Bewältigung von Raum in Schülertexten

Dissertation zur Erlangung des akademischen Grades Doctor philosophiae (Dr. phil.) der Philosophischen Fakultät der Universität Rostock. Tag der mündlichen Prüfung: 8. Dezember 2010. 1. Gutachter: Prof. Dr. Wolfgang Sucharowski, 2. Gutachter: Prof. Dr. Karl-Heinz Ramers, 3. Gutachter: Prof. Dr. Christina Gansel. Veröffentlicht: Rostock 2010.

→ Nachweis und Downloadmöglichkeit Universität Rostock

→ Nachweis Deutsche Nationalbibliothek.

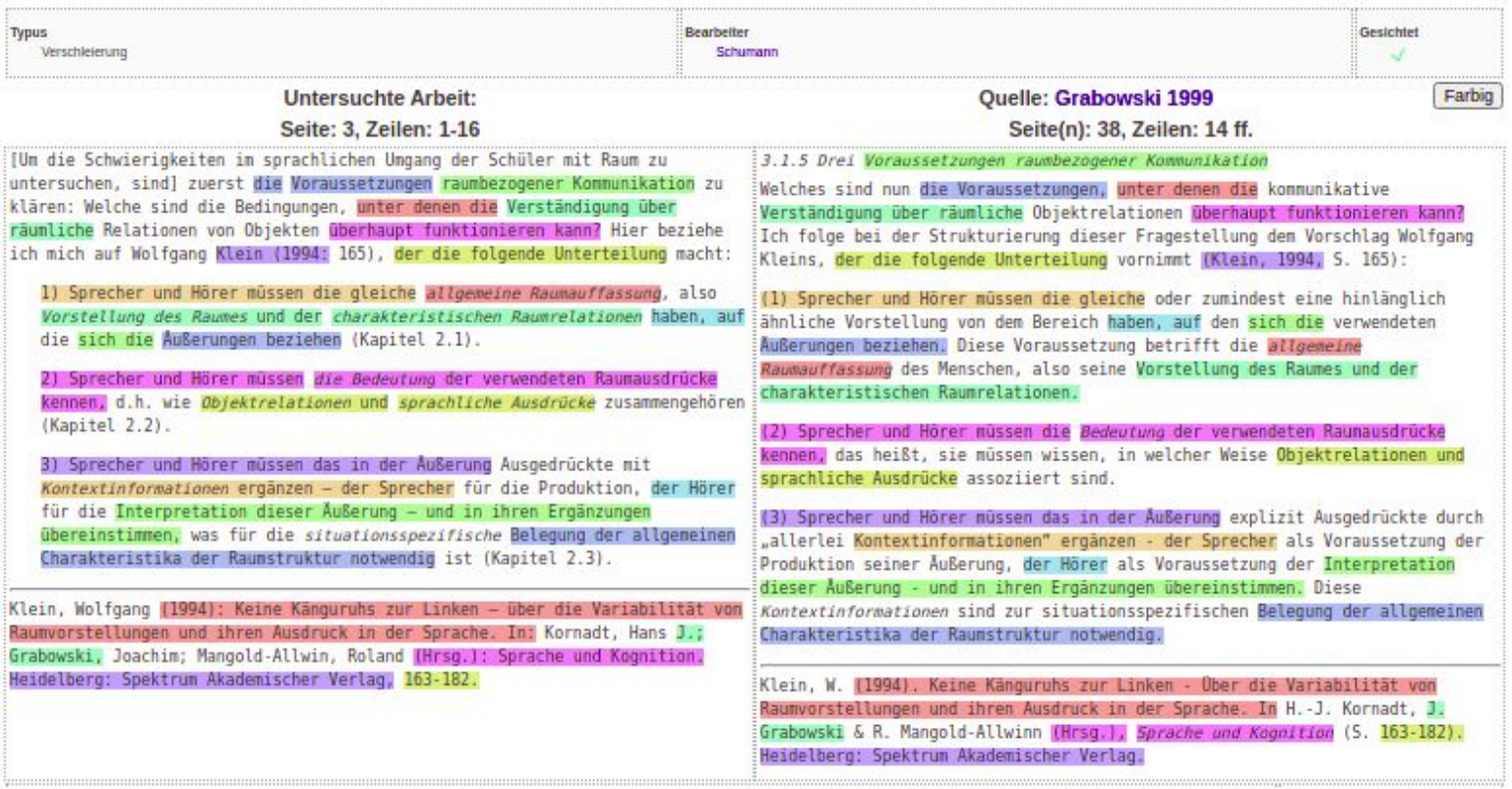
Seiten

Haupttext

001 **002 003** 004 **005** **007 008 009 010 011 012 013 014 015 016 017 018 019 020**
021 022 023 024 025 026 027 028 029 030 031 032 **033 034 035 036 037 038 039 040**
041 042 043 044 045 046 047 048 049 050 051 052 053 054 055 056 057 058 059 060
061 062 063 064 065 066 067 068 069 **070** 071 **072 073 074 075 076 077 078 079 080**
081 082 **083 084 085** 086 088 089 090 **091** 092 **093 094** 095 **096** 097 **098** 099 100
101 **102 103 104 105 106** 107 108 109 **110 111 112** 113 **114 115 116** 117 118 **119** 120
121 122 123 124 125 126 127 **128** 129 130 131 132 **133 134** 135 **136 137 138** 139 **140**
141 142 143 144 145 146 147 148 149 150 151 153 154 155 156 157 158 159 **160**
161

Sample case

Vroniplag: Sample fragment



Typus: Verschleierung

Bearbeiter: Schumann

Gesichtet: ✓

Farbig

Untersuchte Arbeit:
Seite: 3, Zeilen: 1-16

[Um die Schwierigkeiten im sprachlichen Umgang der Schüler mit Raum zu untersuchen, sind] zuerst die Voraussetzungen raumbezogener Kommunikation zu klären: Welche sind die Bedingungen, unter denen die Verständigung über räumliche Relationen von Objekten überhaupt funktionieren kann? Hier beziehe ich mich auf Wolfgang Klein (1994: 165), der die folgende Unterteilung macht:

1) Sprecher und Hörer müssen die gleiche *allgemeine Raumauffassung*, also *Vorstellung des Raumes* und der *charakteristischen Raumrelationen* haben, auf die sich die Äußerungen beziehen (Kapitel 2.1).

2) Sprecher und Hörer müssen *die Bedeutung* der verwendeten Raumausdrücke kennen, d.h. wie *Objektrelationen* und *sprachliche Ausdrücke* zusammengehören (Kapitel 2.2).

3) Sprecher und Hörer müssen das in der Äußerung Ausgedrückte mit *Kontextinformationen* ergänzen – der Sprecher für die Produktion, der Hörer für die Interpretation dieser Äußerung – und in ihren Ergänzungen übereinstimmen, was für die *situationsspezifische* Belegung der allgemeinen Charakteristika der Raumstruktur notwendig ist (Kapitel 2.3).

Klein, Wolfgang (1994): Keine Känguruhs zur Linken – über die Variabilität von Raumvorstellungen und ihren Ausdruck in der Sprache. In: Kornadt, Hans J.; Grabowski, Joachim; Mangold-Allwin, Roland (Hrsg.): Sprache und Kognition. Heidelberg: Spektrum Akademischer Verlag, 163-182.

Quelle: Grabowski 1999
Seite(n): 38, Zeilen: 14 ff.

3.1.5 Drei Voraussetzungen raumbezogener Kommunikation

Welches sind nun die Voraussetzungen, unter denen die kommunikative Verständigung über räumliche Objektrelationen überhaupt funktionieren kann? Ich folge bei der Strukturierung dieser Fragestellung dem Vorschlag Wolfgang Kleins, der die folgende Unterteilung vornimmt (Klein, 1994, S. 165):

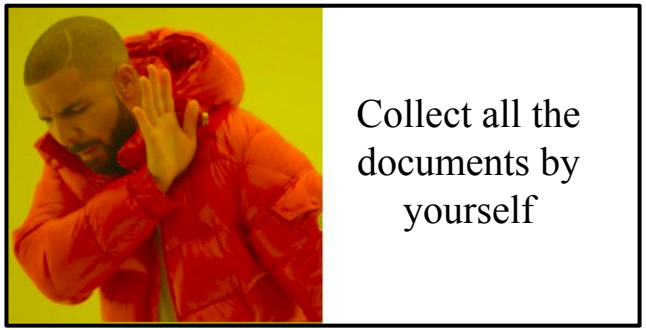
(1) Sprecher und Hörer müssen die gleiche oder zumindest eine hinlänglich ähnliche Vorstellung von dem Bereich haben, auf den sich die verwendeten Äußerungen beziehen. Diese Voraussetzung betrifft die *allgemeine Raumauffassung* des Menschen, also seine Vorstellung des Raumes und der charakteristischen Raumrelationen.

(2) Sprecher und Hörer müssen die *Bedeutung* der verwendeten Raumausdrücke kennen, das heißt, sie müssen wissen, in welcher Weise Objektrelationen und sprachliche Ausdrücke assoziiert sind.

(3) Sprecher und Hörer müssen das in der Äußerung explizit Ausgedrückte durch „allerlei Kontextinformationen" ergänzen - der Sprecher als Voraussetzung der Produktion seiner Äußerung, der Hörer als Voraussetzung der Interpretation dieser Äußerung - und in ihren Ergänzungen übereinstimmen. Diese *Kontextinformationen* sind zur situationsspezifischen Belegung der allgemeinen Charakteristika der Raumstruktur notwendig.

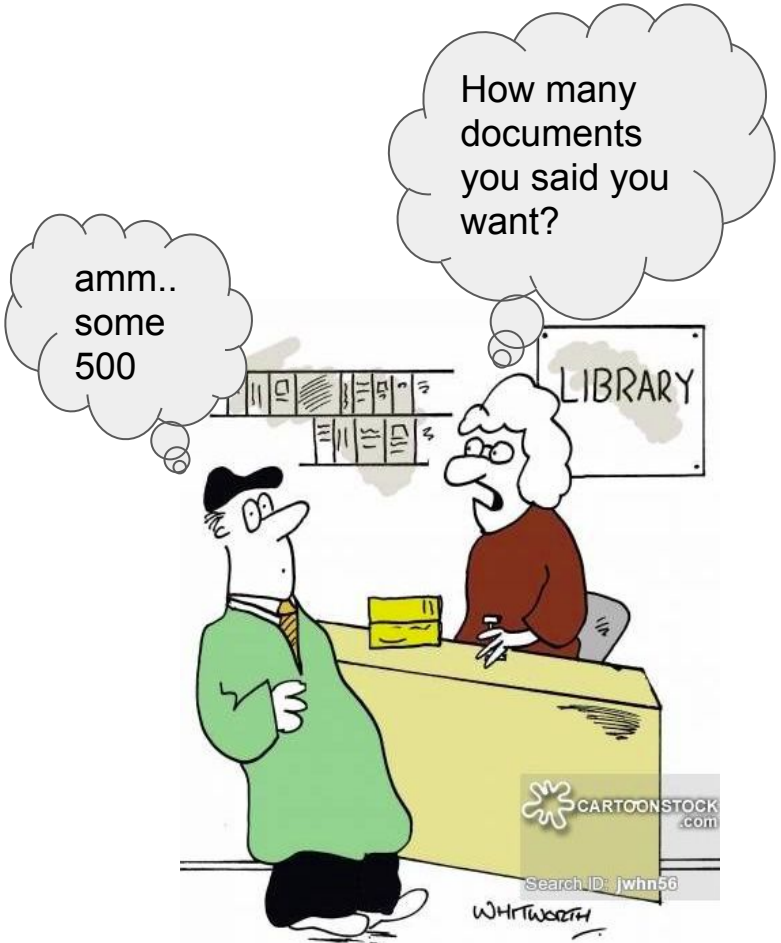
Klein, W. (1994). Keine Känguruhs zur Linken - Über die Variabilität von Raumvorstellungen und ihren Ausdruck in der Sprache. In H.-J. Kornadt, J. Grabowski & R. Mangold-Allwinn (Hrsg.), *Sprache und Kognition* (S. 163-182). Heidelberg: Spektrum Akademischer Verlag.

Vroniplag: Data Acquisition



Vroniplag: Data Acquisition

Vroniplag: Data Acquisition



Vroniplag corpus creation

Raw data

Vroniplag corpus creation

Raw data

web scraping, fragment offset matching,

unified encoding, missing text…….

Structured data

State of the art

- Winner of PAN competition
  - Sanchez. et. al. approach
  - ~10 hyperparameters

Next steps

- Run state of the art
- Hyperparameters tuning
- Towards new approach to solve the problem

Thank
You!

Vaccines Cause Autism

Finding Evidence for Medical Causal Claims

12.03.2021

Ferdinand Schlatt
Martin-Luther-Universität Halle-Wittenberg

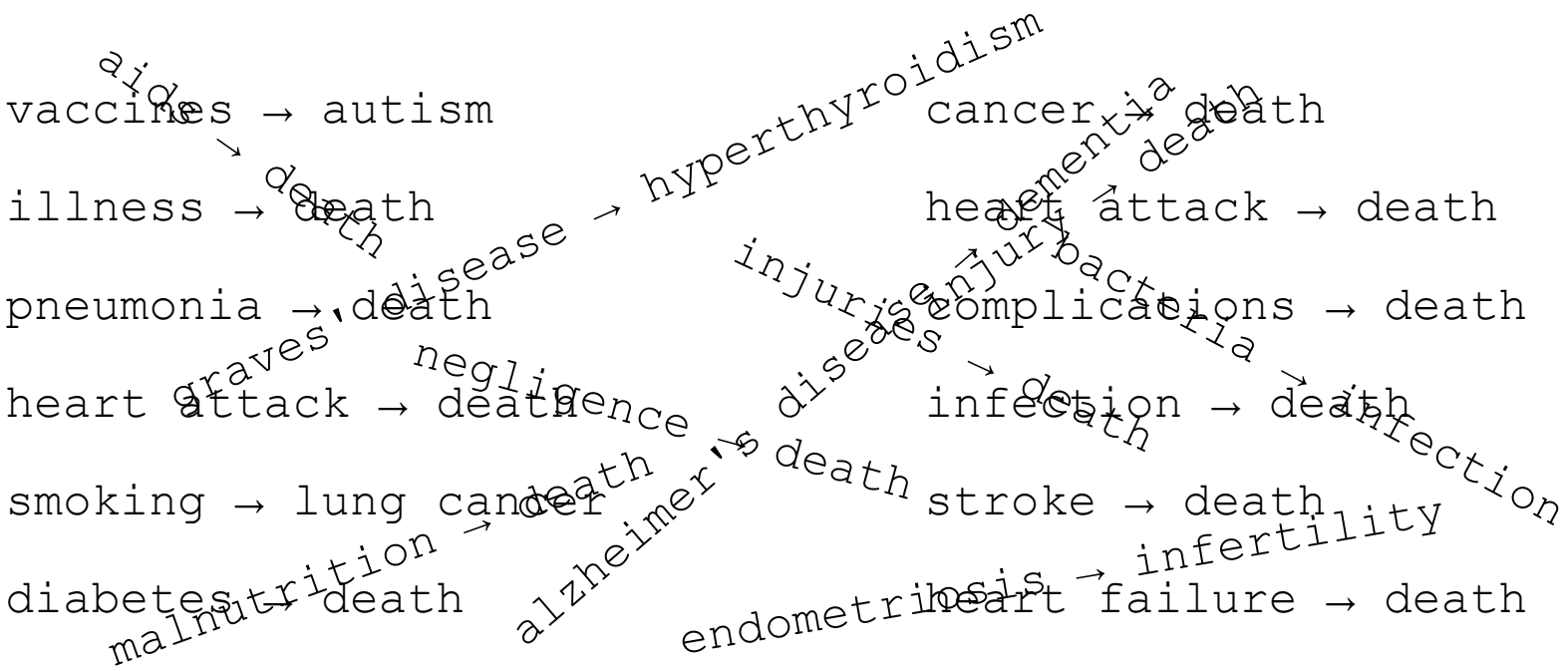
CauseNet

- High precision extraction of 10M+ cause - effect relations from ClueWeb12[1]

```
humidity → permanent damage to wood

spyware present on the computer → security risks

industrial soy production → environmental damage in south
america
```

1. Stefan Heindorf, et al. 2020. CauseNet: Towards a Causality Graph Extracted from the Web. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20). 3023--3030.}

Medical CauseNet



Medical CauseNet

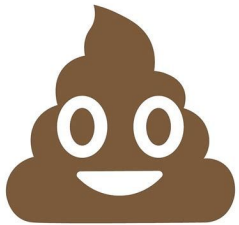
Questions:

1. What portion of the causal claims have a medical context?

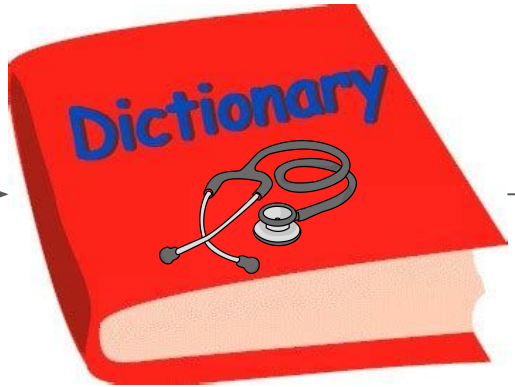
2. For what portion of medical causal claims is it possible to find evidence for?

3. ~~Does the found evidence support or contradict the claim?~~

   What portion of causal claims are bull****?



Medical Causal Relations

Causal Claims → 

→ Medical Causal Claims

- Evaluation on 1000 causal claims
  F1 Score: 0.73

⟶ 2,902,249 / 11,592,969 = 25.03% medical claims

Medical Evidence

- Find PubMed abstracts containing all medical concepts

  ⟶ Sample of 50,000 relations, 76.42 % at least 1 abstract

Image Credits

- Poo: https://ih0.redbubble.net/image.453133309.9960/flat,1000x1000,075,f.jpg
- Dictionary: http://learningideasgradesk-8.blogspot.com/2010/11/free-online-dictionary.html
- Stethoscope: https://clipground.com/stethoscope-clipart.html
- Pubmed: https://de.wikipedia.org/wiki/Datei:US-NLM-PubMed-Logo.svg
- Magnifying glass: http://www.emoji.co.uk/view/11072/

Crypsor

Obfuscation of sensitive search queries

Eric O. Schmidt, Georg-Cantor-Gymnasium Halle

If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place.

Eric Schmidt

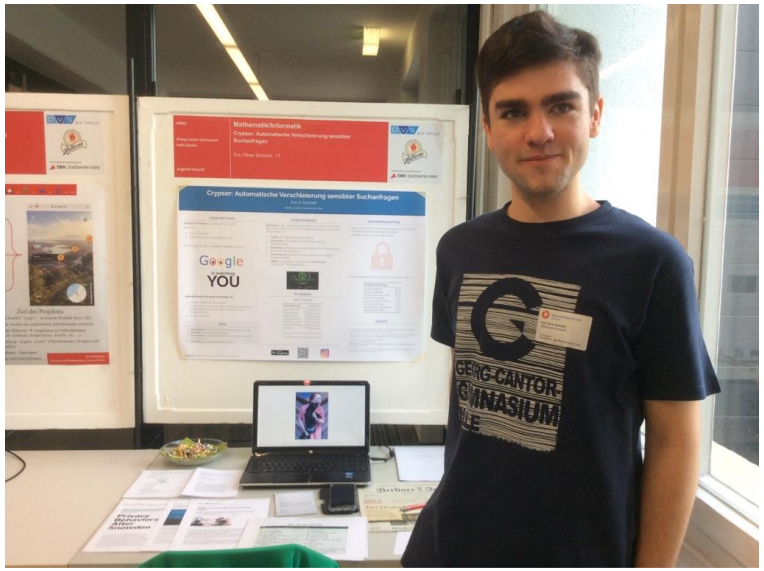
If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place.

Eric Schmidt (former Google CEO)

If you have something that you don't want anyone to know, ~~maybe you shouldn't be doing it in the first place~~ you should be able to do it anyway – so let’s go for it!

Eric Schmidt (former Google CEO)

Eric Schmidt (student)

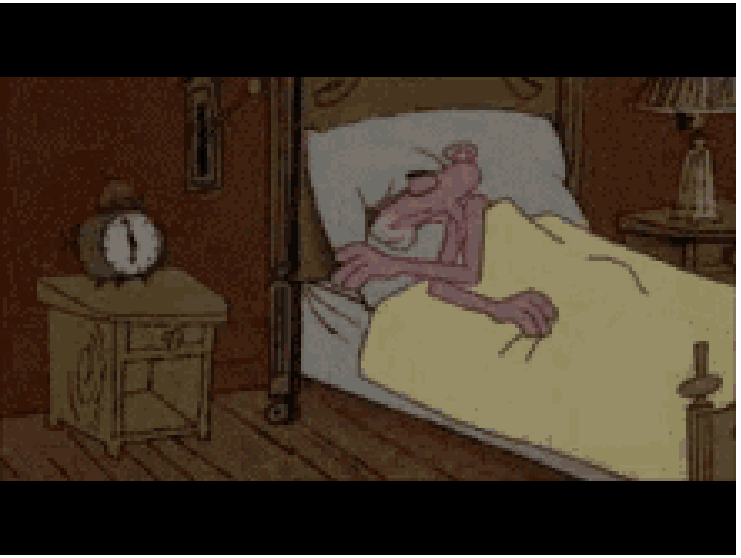


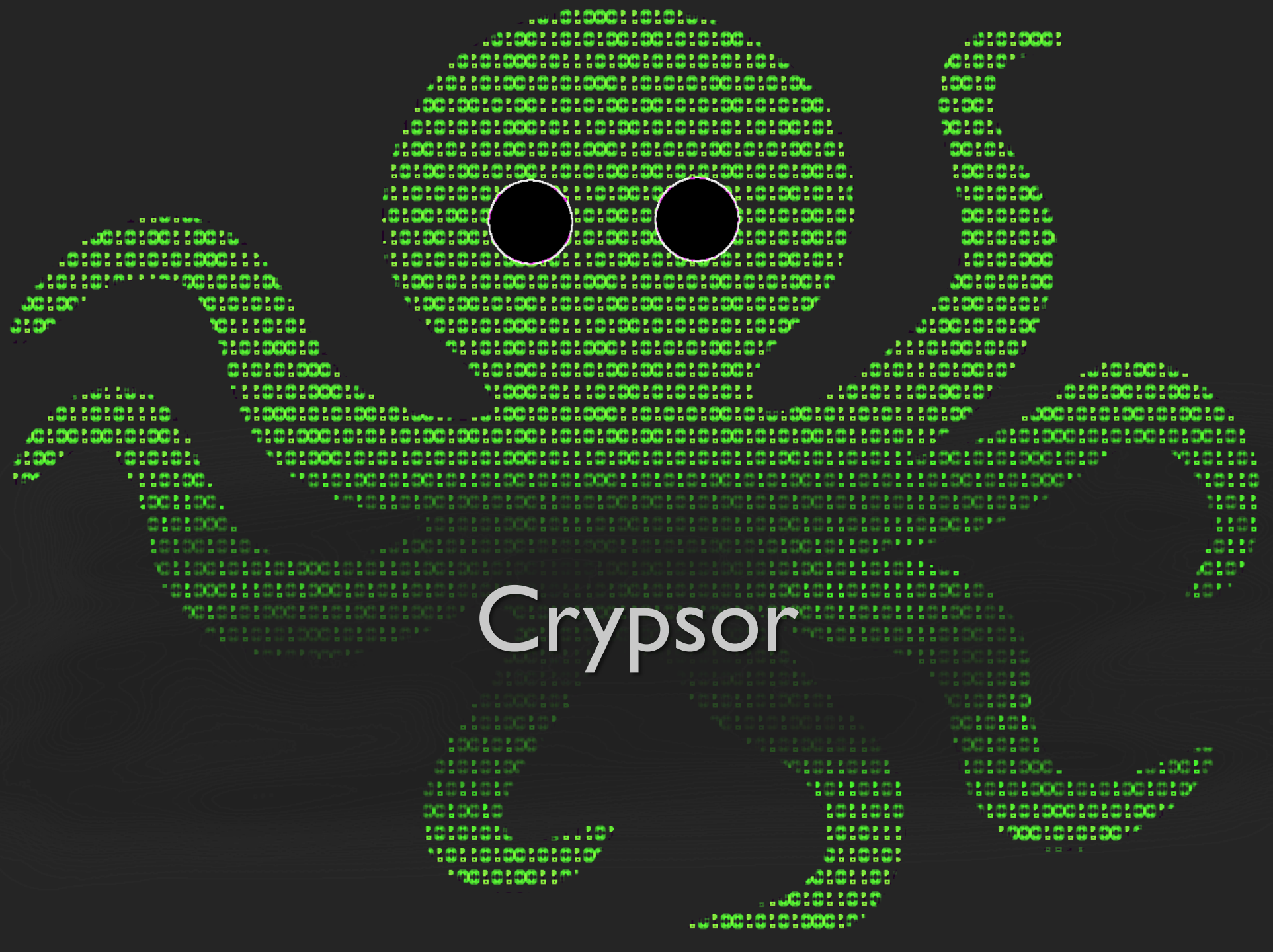
Goals


I don't use my blinkers because it's none of your business where I'm going.

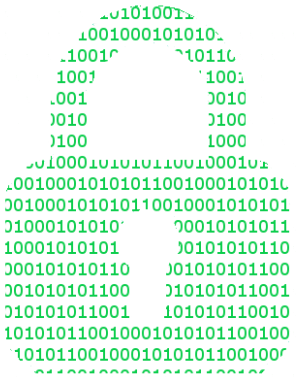

YOUR BROWSER HISTORY
WE NEEDS IT
memegenerator.net

Crypsor




Crypsor

The idea & use case

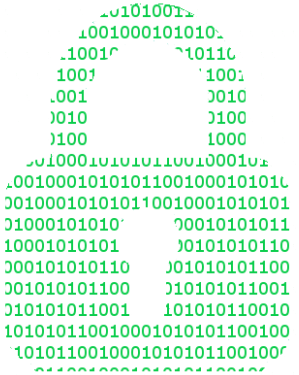


how to cheat in an exam

sensitive search query

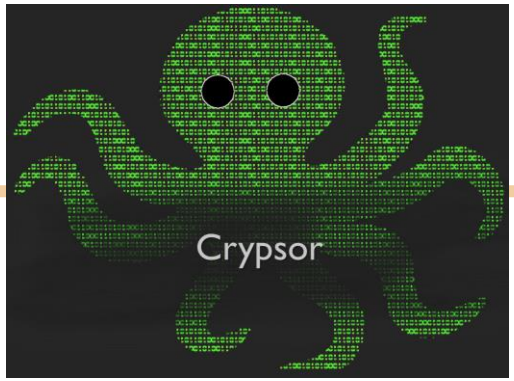
queries to Google

The idea & use case



how to cheat in an exam

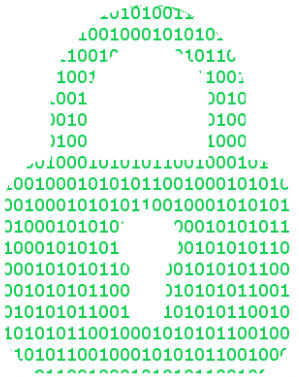
sensitive search query



query not in clear text

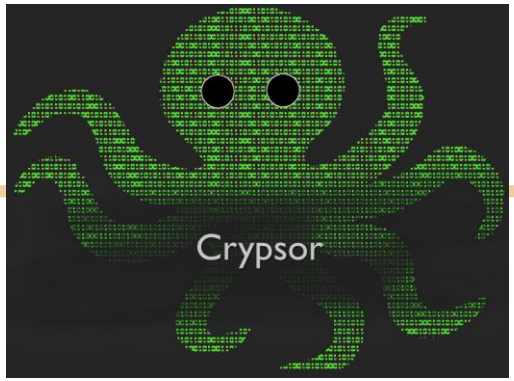
queries to Google

The idea & use case



how to cheat in an exam

sensitive search query



education start discussion

revision timetable

difficult subject

revision stress

report text version

...

less sensitive search queries („keyqueries“)

query not in clear text

queries to Google

The realization

- "Germans fear for their data"
- no trust to foreign servers
- Evaluation + empirical user study

Evaluation

sensitive search query	recall
car radar detectors	40%
cheating husbands	20%
hacking yahoo passwords	54%
how to take optygen	25%
leukemia symptoms teens	14%
post traumatic stress	68%
symptoms of bone infection	43%
unemployment office	13%
average	**35%**

We know where you are. We know where you've been. We can more or less know what you're thinking about.

Eric Schmidt (former Google CEO)

is watching

YOU