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.
Genre Identification of Webpages

<table>
<thead>
<tr>
<th>This web page wants</th>
<th>mainly</th>
<th>also</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>to allow the visitor to discuss with others</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>to suggest pages to the visitor (ignore ads!)</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>to get information from the visitor</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>to sell to or buy from the visitor (ignore ads!)</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>to entertain the visitor</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>to inform the visitor</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>
Multi Annotator Competence

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

<table>
<thead>
<tr>
<th>Workers</th>
<th>Name</th>
<th>Assignments</th>
<th>Approved</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>AZXM77IPUDQ0S</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AZ00712WKGS0I</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AYZNYPT2TVDWF</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AYSTMCRE2AE7T</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AYJ2Z50W4IN8V</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AYHIH9NTPYFLY</td>
<td>179</td>
<td>179</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AYGOIYMWWWGF2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AYF300N0NPCMU</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AY5ZTLIRK9IOS</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Random Forest Classifier

Diagram:
- Webis-WebSeg-20
- Word2Vec
- Model (1 total)
- Word Embeddings (1 per page)
- Annotations (per page)
- Classifier
- Predictions
- Accuracy
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", "en", []
["stephan", "weil"]
["ayg", "l", "zkan"]
["mehr", "zum", "them"]

[-0.39892951 0.94143304 2.85872626 2.95971623 4.65714902 5.88993579
  22.90343666 25.1868642 25.23326331 26.27327681 27.92765573 29.00115197
  29.4320538 30.35311526]
THE END
Conversational Argument Search

A Creepy Search Engine?
Show me arguments for nuclear energy

Here is the first page of arguments for the topic nuclear energy
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
Hey. I was wondering are cars with internal combustion engines better than electric cars?
Sorry, I didn’t get that.
Did you mean electric cats?

https://9gag.com/gag/ajqeR4R
A Convo with an Argument Search Engine

User:

No. I want to inform myself about the disadvantages of electric cars.
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.
Okay. Based on your previous interest, here are the first three arguments against electric cats.
A Convo with an Argument Search Engine

I DON'T WANT YOUR DAMN LEMONS!
Get me pictures of Spider-Man!
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
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
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

> generates motion vectors for translated elements
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
IMPLEMENTATION OF THE FRAMEWORK

COMPARISON

Original Screenshot

Description of Structure

Archived Screenshot

Reconstruction

FFMPEG

Reconstructed Archived Screenshot

Comparison

Error

Model

Human Annotations

Correlation

translation data for each element
IMPLEMENTATION OF THE FRAMEWORK

COMPARISON

<table>
<thead>
<tr>
<th>Comparison</th>
<th>AVG RELATIVE PIXEL ERROR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archived Screenshots vs Original</td>
<td>23.11 %</td>
</tr>
<tr>
<td>Reconstructed Archived Screenshots</td>
<td>21.07 %</td>
</tr>
</tbody>
</table>

**REDUCTION OF PIXEL ERROR: 2.04 %**
IMPLEMENTATION OF THE FRAMEWORK

MODEL
### LINEAR REGRESSION RESULTS

<table>
<thead>
<tr>
<th></th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCHIVED SCREENSHOTS VS ORIGINAL</td>
<td>74.8 %</td>
</tr>
<tr>
<td>RECONSTRUCTED ARCHIVED SCREENSHOTS VS ORIGINAL</td>
<td>76.6 %</td>
</tr>
</tbody>
</table>

**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:**
  - ≥ **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:
  - ≥ 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:
  - ≥ 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:
  - ≥ 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

Maik Fröbe
Relevance Label Transfer

Relevance Transfer from ClueWeb09 to ClueWeb12

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

- Pilot Study:
  - ≥ 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:
  - ≥ 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!
Phoenix – Scientific Text Reuse

Lukas Gienapp
March 11, 2021
Leipzig University
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?
...uestionable, we decided to retain these items and further evaluate the unidimensionality of the scale based upon PCA of the residuals. PCA of the residuals identified that the variance explained by the measures for the empirical calculation (75.6%) was almost identical to the model (75.7%). The unexplained variance explained by the first contrast was 1.7 eigenvalue units (i.e., <2.0 eigenvalue units). Taken together, these results suggest unidimensions...

... the measurement, they were retained in the questionnaire and unidimensionality was further evaluate based upon PCA of the residuals. PCA of the residuals identified that the variance explained by the measures for the empirical calculation (34.3%) was very similar to the model (35.0%). However, the unexplained variance explained by the first contrast was 2.7 eigenvalue units (i.e. 42.0 eigenvalue units), suggesting the possible presence of a secondary...

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>( \approx 6 \times 10^6 )</td>
</tr>
<tr>
<td>Possible Comparisons</td>
<td>( \approx 36 \times 10^{12} )</td>
</tr>
<tr>
<td>Conducted Comparisons</td>
<td>( \approx 9 \times 10^9 )</td>
</tr>
<tr>
<td>Found Reuse Cases(^1)</td>
<td>( \approx 80 \times 10^6 )</td>
</tr>
</tbody>
</table>

\(^1\)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
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:

what is more, the relevant cost dynamic is completely

im übrigen ist die diesbezügliche kostenentwicklung völlig

wir sind es den steuerzahlern schuldig die kosten

we owe it to the taxpayers to keep the costs

under control

unter kontrolle

unter kontrolle zu haben

in check

Bannard et al., 2005, ACL 2005
Most automatic paraphrase acquisition methods are based on this approach:

- Only phrasal paraphrases
Motivation

Most automatic paraphrase acquisition methods are based on this approach:

- Only phrasal paraphrases
- Require large parallel corpora

Bannard et al., 2005, ACL 2005
Motivation

Most automatic paraphrase acquisition methods are based on this approach:

- Only phrasal paraphrases
- Require large parallel corpora

Bannard et al., 2005, ACL 2005
Novel Paraphrase Approach
Novel Paraphrase Approach

Novel Paraphrase Approach


https://en.wikipedia.org/wiki/Goethe%E2%80%93Schiller_Monument
Johann Wolfgang von Goethe was a German poet, playwright, novelist, scientist, statesman, theatre director, critic, and amateur artist.

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

- 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

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

• Simulating Interactive Information Retrieval

Problem:
• Evaluating changes to a search engine/digital library is:
  • Cost intensive
  • Slow
  • You need actual users 😞

Solution:
JUST SIMULATE IT.
Idea

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

• Aspects to cover:
  - Query Formulation
  - Eye Tracking!
  - Clicking Behaviour
  - Reading Time?
  - Stopping Behavior
  - What about missclicks!?
  - Knowledge Model
  - Typos!!1!
  - Cost/Gain
  - Query Re-Formulation
  - ...
The Complex Searcher Model

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

Source: D. Maxwell [1]

12.03.2021
Webis Flash Talks 2021
So far: It’s all about Queries!

Task:
• Build an artificial session
  • Queries only
  • Containing multiple related queries
  • On a certain topic

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


Title image: by Oladimeji Ajegbile (https://www.pexels.com/de-de/foto/mann-der-mit-einem-laptop-arbeitet-2696299/)
How to get FAME in 7 steps

Leipzig, 12.03.2021
Ahmad Dawar Hakimi

ASV
Automatische Sprachverarbeitung
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 \(\iff\) 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 \(\iff\) manual encoding/mapping into ACE representation
Figure 1: FAME architecture
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
What does my Language Model know?
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?
But...

Wrong Legrand(e)!

Pierre Legrand

Nonviolent Civil Disobedience and Police Enforcement Policy (1968)

J. L. LeGrande
- 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.
The Iron Law of AI:
All productivity gains from Artificial Intelligence will be lost to sitting around looking at weird stuff produced by Artificial Intelligence.

Source: https://www.smbc-comics.com/comic/gan
Thanks for listening!

Image sources


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)

What is safer, beer or milk? and why?

Because beer is boiled during the brewing process, this liquid was safer than water, milk, and other perishable liquids.
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?
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.
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?

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!

SAY CLIMATE CHANGE

ONE MORE TIME
TEXT QUALITY IN SEARCH - SUPPORTED WRITING

Supervised by: Michael Völske and Magdalena Wolska

Presenter: Bibek Khadayat

12/03/2020
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.
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
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
YOU NEED DAYS TO LEARN ALEXA?

I'LL TELL YOU IN 5 MINUTES
A stack of fruits

- Push(🍒)  
  “Add a cherry”
- Pop() → 🍒  “Remove the last one”
- Clear() →  “Eat it all”
THE TECH IS REALLY COMPLICATED

BUT THIS SHOWS EVERYTHING YOU NEED TO KNOW
Speech

Logic

Example skill for Alexa: a stack of fruits.

data
- skill
  - en
    - intents
    - launch
    - types
Speech

Logic

Add \{fruit\}
Speech

Example skill for Alexa: a stack of fruits.

Logic

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

@KieselJohannes
Example skill for Alexa: a stack of fruits.

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

Example skill for Alexa: a stack of fruits.

Logic

```java
public class PushIntent extends GenericIntent {
    public PushIntent(...) throws ... {
        super(new Configuration("push", ...)
            .addSlot(Slot.builder("fruit", "fruit")
                .isRequired(true)
            ...
        );
    }
}
```
public class PushIntent extends GenericIntent {
    protected Response.Builder onRequest(...
            Map<String, SlotValue> slots, User user) {
        user.pushFruit(slots.get("fruit"));
    }
}
public class PushIntent extends GenericIntent {
    protected Response.Builder onRequest(...
            Map<String, SlotValue> slots, User user) {
        user.pushFruit(slots.get("fruit"));
    }
}
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), ...);
            }
    }
}

@KieselJohannes
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), ...);
        }
    }
}
Example skill for Alexa: a stack of fruits.

```
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), ...);
            }
        }
```

Speech
```
I added a marvelous {fruit}.
```

Logic
```
Whoa, a stack of {size} fruits with a {fruit} on top!
```
PushIntent extends GenericIntent {
    Response.Builder onRequest(String, SlotValue> slots, User user) {
        String fruit = slots.get("fruit")
        int size = user.getStack().size()
        if (size < 3) {
            this.respond("success", slots,
        } else {
            this.respond("successLargeStack", slots,
    }
}
}
public class PushIntent extends GenericIntent {
    Response.Builder onRequest...
    Map<String, SlotValue> slots, User user) {
        user.pushFruit(slots.get("fruit"));
        int size = user.getStack().size();
        if (size < 3) {
            this.respond("success", slots, ...);
        } else {
            this.respond("successLargeStack", slots, Map.of("size", size), ...);
        }
    }
}
THERE YOU HAVE IT

SUPER EASY PEASY
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

- How to add an intent
- How to add an intent confirmation prompt
- How to add an intent reprompt
- How to add a skill launch response
- How to add a slot confirmation prompt
- How to add a slot elicitation prompt
- How to add a slot type
- How to add a slot type with automatically generated values
- How to add a slot utterance
- How to add a slot validation rule
- How to add the command line interface
- How to install the Amazon Skill Kit CommandLine Interface (ASK-CLI)
- How to run the skill server
- How to run the skill server in our cloud
- How to run the skill server locally for testing
- How to run the skill setup
- How to run the skill update
- How to write a speech file
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-fruits-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

```java
$ 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
Your Fruits

- peach
- lemon
- fig
- pear
- pomegranate

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

Approach:

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

Alonso Palomino
Bias

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?

<table>
<thead>
<tr>
<th>Argument A</th>
<th>Argument B</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>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 - &quot;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.&quot;</td>
</tr>
</tbody>
</table>
Differential bias

Argument Corpora

WAF 19 [1]

Argument A

Argument B

Unlabeled Argument instances

1

...  

N

Large scale crowdsourcing labeling

Differential bias

Argument Corpora

WAF 19 [1]

Labeled Argument instances

Large scale crowdsourcing labeling

**Which one is more biased? (stance)**

**Banning Zoos**

<table>
<thead>
<tr>
<th>Argument A</th>
<th>Argument B</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>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 - &quot;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.&quot;</td>
</tr>
</tbody>
</table>
### Argument A

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]

### Argument B

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]
### Argument A

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.

### Argument B

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

<table>
<thead>
<tr>
<th>Argument A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zoos are not capable of sustaining all <strong>endangered</strong> species. According to the world conservation union which keeps records of <strong>endangered</strong> 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.</td>
</tr>
<tr>
<td>Most of the animals that you see in zoos aren't <strong>endangered</strong>. While some argue that zoos are a means to protecting <strong>endangered</strong> species, the reality is that very few animals in zoos are actually <strong>endangered</strong>. In other words, this is really not the reason why zoos exist and so should not be put forward as a justification for them.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Argument B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zoos can raise awareness of <strong>endangered</strong> species. Visitors to zoos may raise their awareness of <strong>endangered</strong> species by being directly exposed to them.</td>
</tr>
<tr>
<td>If nature was appropriately preserved, we would not need zoos. Michael fox, sierra, november-december 1990 - &quot;Zoos are becoming facsimiles - or perhaps caricatures - of how animals once were in their <strong>natural</strong> habitat. if the right policies toward <strong>nature</strong> were pursued, we would need no zoos at all.&quot;</td>
</tr>
</tbody>
</table>
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


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 \(\rightarrow\) double impact on loss \(\rightarrow\) 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
YOU SHALL NOT PASS!

PLAGIARISM
What is Text-Alignment?

Plain Documents

Aligned passages

Text-Alignment
Available Corpora

- Small scale
  - Answer pairs
- Large scale
  - PAN datasets
Eine kritische Auseinandersetzung mit der Dissertation von Prof. Dr. Diana Timova: Konzeptualisieren und Verbalisieren von Raum – kognitive und sprachliche Bewältigung von Raum in Schüllertexten
Vroniplag: Sample fragment

[Text content from the image]

Vroniplag: Sample fragment
Vroniplag: Data Acquisition

Collect all the documents by yourself
<table>
<thead>
<tr>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collect all the documents by yourself</td>
</tr>
<tr>
<td>Ask someone else to do it who has access</td>
</tr>
</tbody>
</table>
Vroniplag: Data Acquisition

<table>
<thead>
<tr>
<th>Collect all the documents by yourself</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask someone else to do it who has access</td>
</tr>
</tbody>
</table>

How many documents you said you want?

amm.. some 500
Vroniplag corpus creation

Raw data
Vroniplag corpus creation

Raw data

Structured data

web scraping, fragment offset matching,
unified encoding, missing text…….
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
• High precision extraction of 10M+ cause - effect relations from ClueWeb12

humidity → permanent damage to wood

spyware present on the computer → security risks

industrial soy production → environmental damage in south america

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

- Evaluation on 1000 causal claims
  F1 Score: 0.73

  \[ \frac{2,902,249}{11,592,969} = 25.03\% \text{ medical claims} \]
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
- Stethoscope: https://clipground.com/stethoscope-clipart.html
- Magnifying glass: http://www.emoji.co.uk/view/11072/
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.
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

**education start discussion**

**revision timetable**

**difficult subject**

**revision stress**

**report text version**

**query not in clear text**

queries to Google

less sensitive search queries („keyqueries“)

how to cheat in an exam sensitive search query
The realization

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

<table>
<thead>
<tr>
<th>Sensitive Search Query</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>car radar detectors</td>
<td>40%</td>
</tr>
<tr>
<td>cheating husbands</td>
<td>20%</td>
</tr>
<tr>
<td>hacking yahoo passwords</td>
<td>54%</td>
</tr>
<tr>
<td>how to take optygen</td>
<td>25%</td>
</tr>
<tr>
<td>leukemia symptoms teens</td>
<td>14%</td>
</tr>
<tr>
<td>post traumatic stress</td>
<td>68%</td>
</tr>
<tr>
<td>symptoms of bone infection</td>
<td>43%</td>
</tr>
<tr>
<td>unemployment office</td>
<td>13%</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>35%</strong></td>
</tr>
</tbody>
</table>

12.03.2021
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)
Google is watching YOU