WOWS-Eval
Lexical Baselines for Relevance Label Transfer

ECIR 2025, April 6–10, Lucca, Italy

Daria Alexander, Maik Fröbe, and Gijs Hendriksen

ows.eu

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Motivation: Build a Search Engine for Your Favorite Local Sports Team

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Motivation: Build a Search Engine for Your Favorite Local Sports Team

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Motivation: Build a Search Engine for Your Favorite Local Sports Team

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Motivation: Build a Search Engine for Your Favorite Local Sports Team

Step 1: Slice and Dice the Open Web Index

- Load Sports Partition from the Open Web Index
- Retain only entries geo-tagged with Lucca

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Motivation: Build a Search Engine for Your Favorite Local Sports Team

Step 1: Slice and Dice the Open Web Index

- Load Sports Partition from the Open Web Index
- Retain only entries geo-tagged with Lucca

Step 2: Select Suitable Retrieval Model

- Select domains for which you think your search engine is used
- E.g., What retrieval models are effective for News and Medical search?

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Motivation: Build a Search Engine for Your Favorite Local Sports Team

Step 1: Slice and Dice the Open Web Index

- Load Sports Partition from the Open Web Index
- Retain only entries geo-tagged with Lucca

Step 2: Select Suitable Retrieval Model

- Select domains for which you think your search engine is used
- E.g., **What retrieval models are effective** for News and Medical search?

Goal of wows-eval:

- Transfer relevance labels from TREC topics to the Open Web Index
- Evaluate which retrieval models work well for downstream use-cases

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Task Formulation

Input

- Query

hydrogen liquid at what temperature?

- Known relevant Document from a TREC corpus

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

- Unknown document from the ClueWeb22

socratic.org/ … /questions/is-hydrogen-a-solid-liquid-or-gas-at-room-temperature

Is hydrogen a solid, **liquid**, or gas at room **temperature**? | Socratic

ClueWeb22 · Crawled Aug 2022

Is hydrogen a solid, **liquid**, or gas at room **temperature**? | Socratic **Is hydrogen** a solid, **liquid**, or gas at room **temperature**? Chemistry Matter Elements 1 Answer Meave60 Jun 6, 2018 **Hydrogen is** a gas at room **temperature**. It **is** in the air that you breathe.

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Task Formulation

Input

- Query

hydrogen liquid at what temperature?

- Known relevant Document from a TREC corpus

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

- Unknown document from the ClueWeb22

socratic.org/ … /questions/is-hydrogen-a-solid-liquid-or-gas-at-room-temperature

Is hydrogen a solid, **liquid**, or gas at room **temperature**? | Socratic

ClueWeb22 · Crawled Aug 2022

Is hydrogen a solid, **liquid**, or gas at room **temperature**? | Socratic **Is hydrogen** a solid, **liquid**, or gas at room **temperature**? Chemistry Matter Elements 1 Answer Meave60 Jun 6, 2018 **Hydrogen is** a gas at room **temperature**. It **is** in the air that you breathe.

Output

$$P(d_u \geq d_r | d_u, d_r, q)$$

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Approaches (1)

We collected autoqrels prompts during an OWS.eu hackathon

- We collected 5 prompts
- Executed all prompts for 3 autoqrels models (Flan-T5-small/base/large)

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Approaches (1)

We collected autoqrels prompts during an OWS.eu hackathon

- We collected 5 prompts
- Executed all prompts for 3 autoqrels models (Flan-T5-small/base/large)

Approaches (2)

- Relevance Feedback
- Relevant Document against Robust04 index
- Reformulated query used to score document with BM25

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Approaches (1)

We collected autoqrels prompts during an OWS.eu hackathon

- We collected 5 prompts
- Executed all prompts for 3 autoqrels models (Flan-T5-small/base/large)

Approaches (2)

- Relevance Feedback
- Relevant Document against Robust04 index
- Reformulated query used to score document with BM25

Approaches (3)

- Manual Snorkel rules (query term presence, BM25/BERT scores)

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Evaluation Setup

- Transferred 13 queries
  - Source corpora: Touche, Robust04, MS MARCO, ClueWeb09
- 1100 manually annotated ClueWeb22 documents
- Evaluation Measure: System ranking correlation when evaluated on human qrels vs. transferred relevance judgments

Results

Approach	τ
Flan-T5-large (best prompt)	**0.427**
Flan-T5-base (best prompt)	0.266
Flan-T5-small (best prompt)	0.067
BM25	0.151
RF (All)	0.266
RF (One)	**0.276**
Snorkel	0.230

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Conclusions

- We aim to transfer TREC relevance judgments to the OpenWebIndex
- Vision
  - Allow to select retrieval model suitable for a certain domain(s)
  - Dense judgments > sparse judgments to evaluate diverse models
- Current approaches work not good enough

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Conclusions

- We aim to transfer TREC relevance judgments to the OpenWebIndex
- Vision
  - Allow to select retrieval model suitable for a certain domain(s)
  - Dense judgments > sparse judgments to evaluate diverse models
- Current approaches work not good enough

Future Work

- Run more models: Umbrella, LLM4Eval approaches, . . .
- Transfer and annotate more queries
- Repetition of the task next year makes sense
  - Goal: reach $\tau \geq 0.7$

WOWS-Eval: Lexical Baselines for Relevance Label Transfer

Conclusions

- We aim to transfer TREC relevance judgments to the OpenWebIndex
- Vision
  - Allow to select retrieval model suitable for a certain domain(s)
  - Dense judgments > sparse judgments to evaluate diverse models
- Current approaches work not good enough

Future Work

- Run more models: Umbrella, LLM4Eval approaches, . . .
- Transfer and annotate more queries
- Repetition of the task next year makes sense
  - Goal: reach $\tau \geq 0.7$

Thank you!