Cross language information retrieval in CLEF eHealth task3

Shadi Saleh

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
saleh@ufal.mff.cuni.cz

Outline

1. About ShAre/CLEF eHealth and Task 3a,3b
2. Language Model-based information retrieval and Hiemstra Model
3. Experiments
    1. Processing dataset
    2. Terrier Information Retrieval System
    3. Optimizing Model parameters
    4. Spell checking
    5. Queries expansions
    6. Narrative usage
    7. CLIR
    8. Results
4. Conclusion & Future Work
5. Q&A

1. ShAre/CLEF eHealth and Task 3

- **The CLEF Initiative**
  - Series of Evaluation Labs, i.e. laboratories to conduct evaluation of information access systems and workshops to discuss and pilot innovative evaluation activities
- **ShARe/CLEF eHealth Evaluation**
  - Shared task focused on natural language processing (NLP) and information retrieval (IR) for clinical care

1. ShAre/CLEF eHealth Task 3

1. **Task 3a**
   - Approximately one million medical documents in English.
   - The goal is to retrieve the relevant documents for the user queries.
   - All queries are in English.

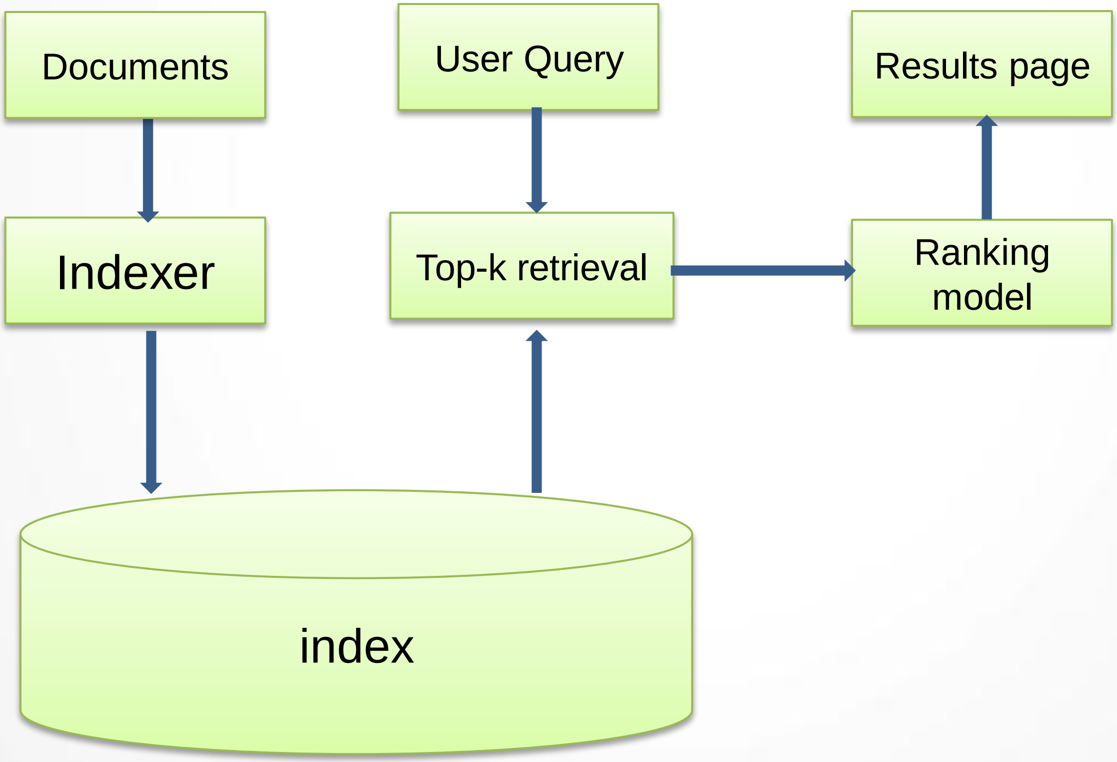
2. **Task 3b**
   - Task 3b extends Task 3a by providing a translation of the queries from Task 3a into German, French and Czech.
   - The goal in Task 3b is to develop techniques to translate these queries into English and then apply them to the retrieval task 3a.

2.Language Model-based information retrieval and Hiemstra Model

2.1 Information Retrieval

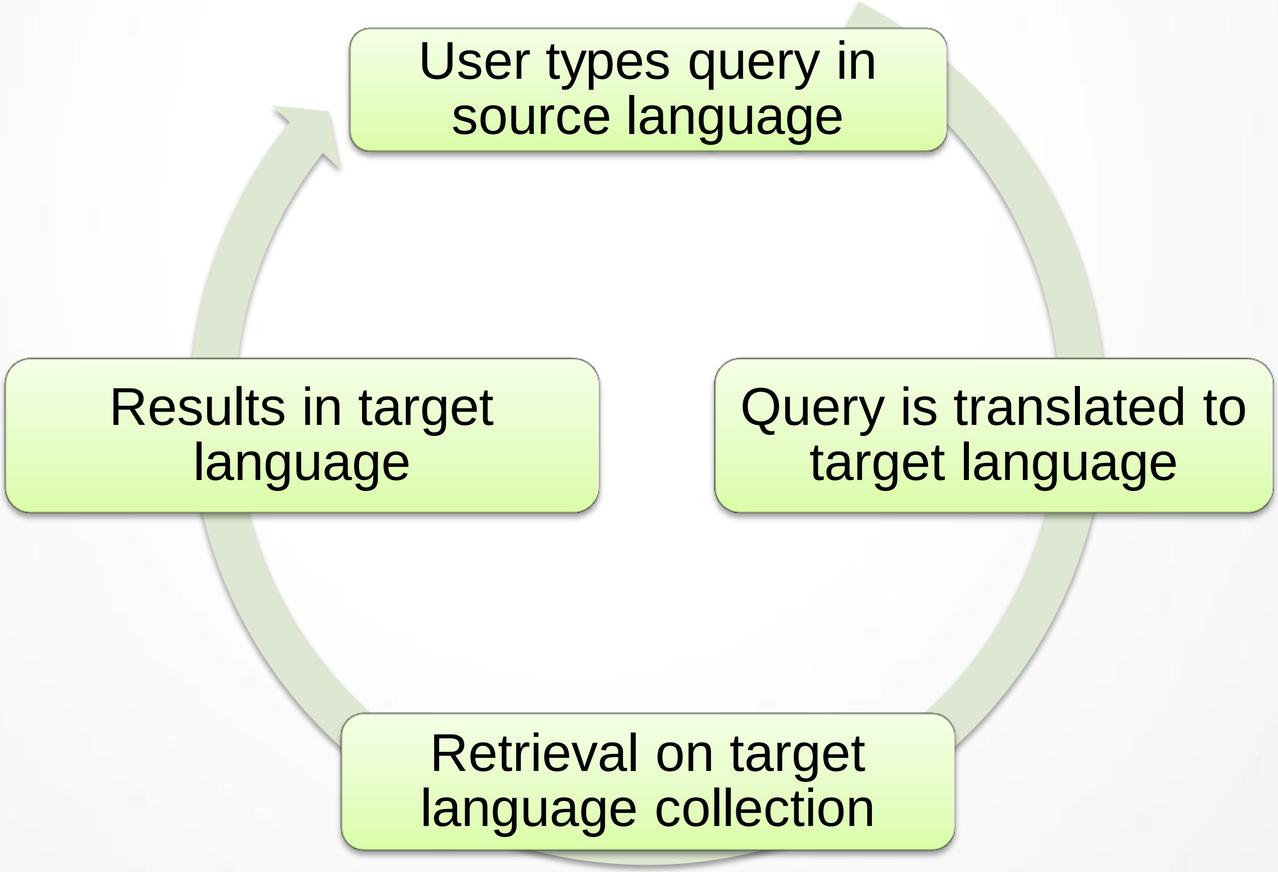
What is IR?

Information retrieval (IR) may be defined, in general, as the problem of the selection of documentary information from storage in response to search questions provided by the user.



2.2 Cross Languages Information Retrieval CLIR

CLIR search engines enable users to retrieve documents in a language different from the language used to formulate the query



User types query in source language

Query is translated to target language

Retrieval on target language collection

Results in target language

2.3 Information retrieval models

- Boolean model
- Vector Space model
- Probabilistic model
- Language model

2.4 LM-based information retrieval

- LM assigns a probability to a sequence of m words $P(W_1,....,W_m)$ by means of a probability distribution.

- In IR, LM is associated with a document in a collection.

- A document is a good match to a query if the document model is likely to generate the query

- For a query Q, retrieved documents are ranked based on the probability $P(Q|M_d)$.

2.5 Hiemstra LM

The model:

$$\mathrm{Prob[Di|Q] = Prob[Di].} \prod_{\mathrm{t_j \in Q}}^{\mathrm{n}} [\boldsymbol{\lambda}.\, \mathrm{Prob}\left[\mathrm{t_j}\middle|\mathrm{Di}\right] + (1 - \boldsymbol{\lambda}).\, \mathrm{Prob[t_j|C]}]$$

- Lambda is the importance of term, it’s smoothing factor (constant for all indexing terms t_j)
- C is the collection

4. Experiments

4. Our Contribution

1. Process data and generate TREC-format dataset
2. Index dataset
3. Do retrieval on training queries using the most popular IR models and evaluate the results to choose the best model
4. Tune the model parameter

5. Task 3a:
   - Do the retrieval using best model and best parameter.
   - Do spell checking on training and test queries and apply the retrieval again.
   - Apply Query Expansion
   - Build queries using other fields (e.g TITLE + NARR)
6. Translate other queries to English and redo the retrieval again.

4.1 Test collection

- Approximately 1.1 million medical-related documents, provided by the Khresmoi project.
- The documents in the collection come from several online sources, including Health On the Net organization certified websites, as well as well-known medical sites and databases.
- All documents are in English

4.1 Test collection

Documents available as a set of HTML pages:

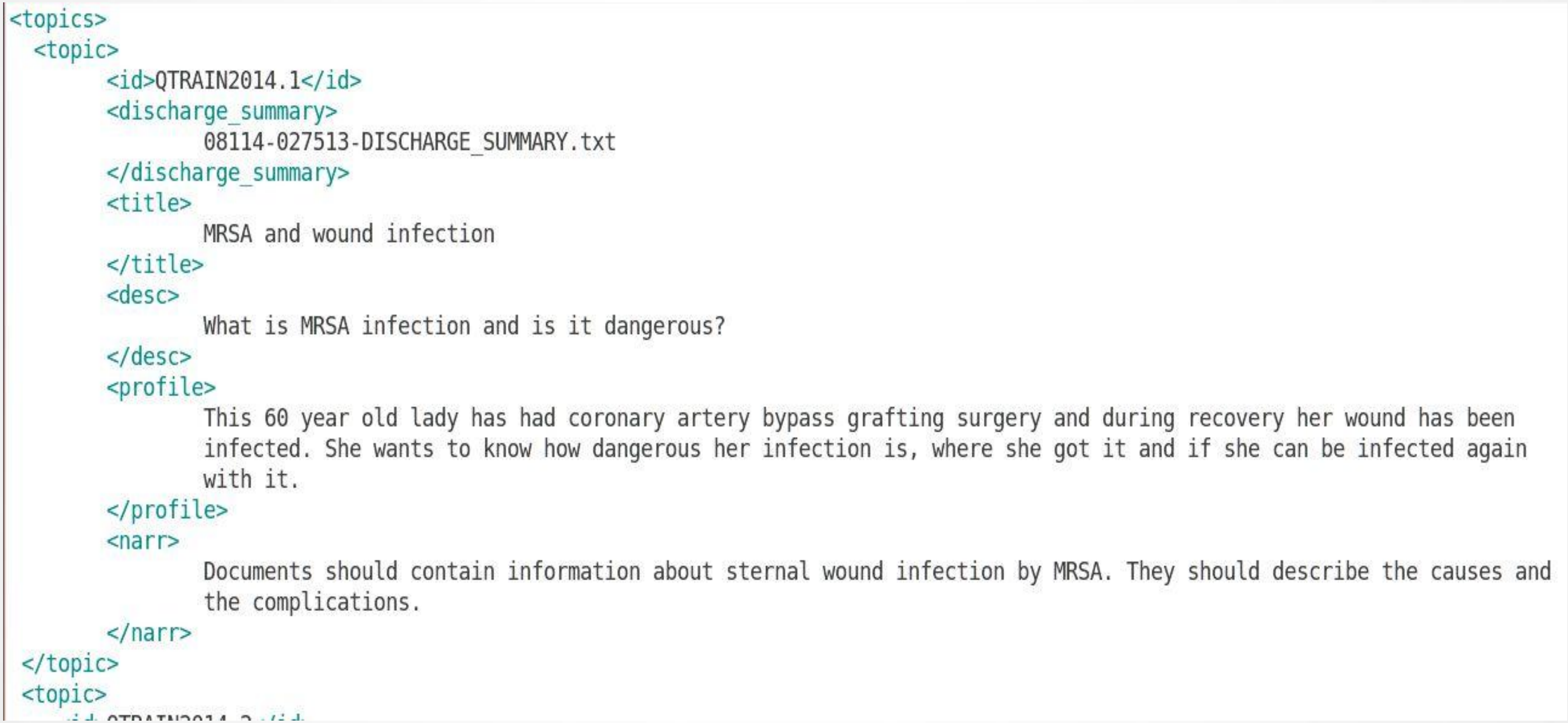
```
#UID:surfa4585_12_000001
#DATE:201204-06
#URL:http://www.surfacehippy.info/hipstories07/patrick07.php
#CONTENT:
<html>
…
</html>
#EOR
#UID:surfa4585_12_000002
#DATE:201204-06
#URL:http://www.surfacehippy.info/hipstories07/janeuk07.php
#CONTENT:
<html>
…
</html>
#EOR
```

4.1 Queries

- We had 5 train queries + 50 test queries in English, Czech, French and German.
- Queries were in TREC format.
- The Text REtrieval Conference (TREC) is an on-going series of workshops focusing on a list of different information retrieval (IR) research areas.
- TREC provides the infrastructure necessary for large-scale evaluation of text retrieval methodologies(e.g. MAP mean average precision).

4.1 Queries

Query Sample:



```
<topics>
  <topic>
        <id>QTRAIN2014.1</id>
        <discharge_summary>
                08114-027513-DISCHARGE_SUMMARY.txt
        </discharge_summary>
        <title>
                MRSA and wound infection
        </title>
        <desc>
                What is MRSA infection and is it dangerous?
        </desc>
        <profile>
                This 60 year old lady has had coronary artery bypass grafting surgery and during recovery her wound has been
                infected. She wants to know how dangerous her infection is, where she got it and if she can be infected again
                with it.
        </profile>
        <narr>
                Documents should contain information about sternal wound infection by MRSA. They should describe the causes and
                the complications.
        </narr>
 </topic>
 <topic>
     <id>QTRAIN2014.2</id>
```

4.1 Document format

TREC documents have the following format:

```
<DOCS>
        <DOC>
        <DOCID>ID1</DOCID>
        <TITLE>Title1</TITLE>
        <TEXT>Text1</TEXT>
        </DOC>

        <DOC>
        <DOCID>ID2</DOCID>
        <TITLE>Title2</TITLE>
        <TEXT>Text2</TEXT>
        </DOC>
</DOCS>
```

4.1 Processing Dataset

- TREC file mainly has TITLE and TEXT tags, Those tags were filled as following:

- TITLE tag contained: <title>THIS TEXT </title>
- If title tag didn't exist in some documents, then <H1>THIS TEXT</H1> as appended to title tag.

- TEXT tag contains: All text in other HTML tags Plus, Meta keywords and description tags.

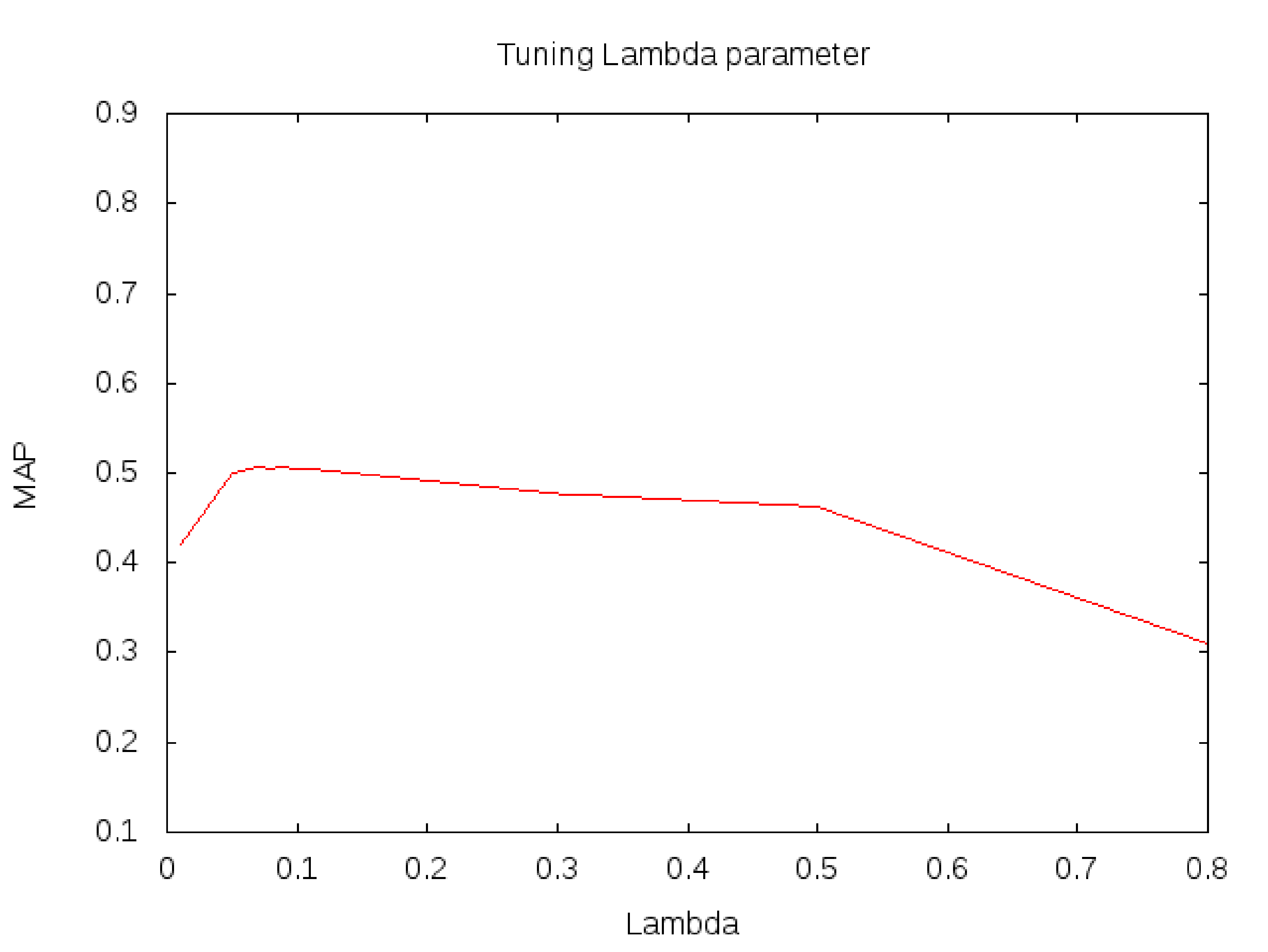
- Cleaning was done using HTML::Strip Perl module.

- Boilerpipe cleaning tool was used too, but didn't give good result compared to HTML::Strip (we got map equals to 0.4082 from Boilerpipe, while we got 0.5130 using HTML::Strip)

4.2 Terrier Search Engine

- Open source, written in Java
- High performance and scalable search engine
- Provides implementation of many weighting models.
- Offers Automatic Query Expansion
- More information: http://terrier.org/

4.3 Optimizing Model parameter (Lambda)



4.4 Spell checking

- Some errors were found in MT output:

  Input: Anoxické poškození mozku
  Output: anoxické brain damage
  Expected: Anoxic brain injury

- This happened because there were unknown words to MT system

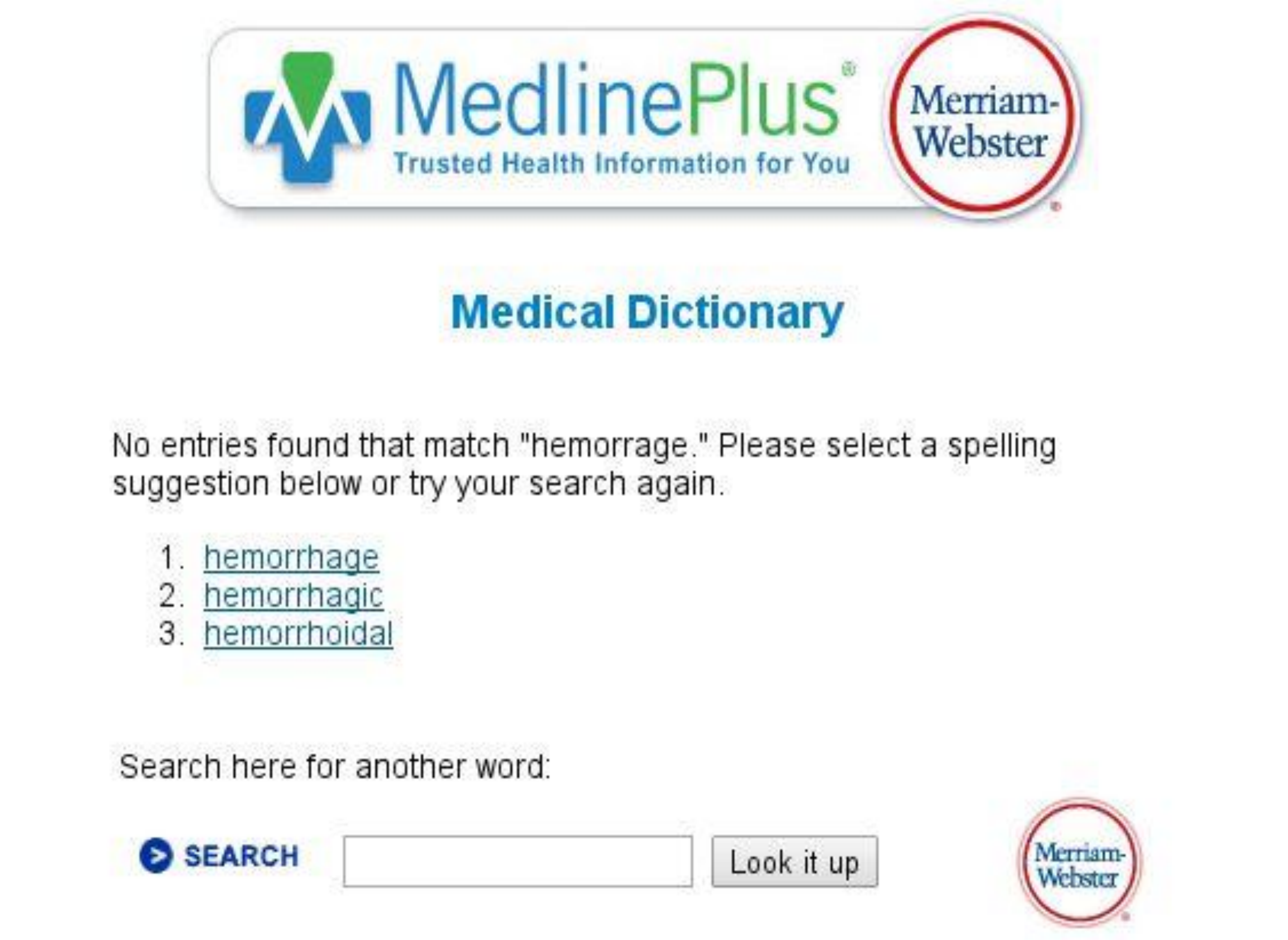
4.4 Spell checking

While looking at English queries, this query was appeared:

tretament for subarachnoid hemorrage

“tretament” is clearly spelling error, but what about hemorrage?

4.4 Spell checking



Medical Dictionary

No entries found that match "hemorrage." Please select a spelling suggestion below or try your search again.

1. hemorrhage
2. hemorrhagic
3. hemorrhoidal

Search here for another word:

Errors were found only in test queries, not in train queries.

4.5 Query Expansion

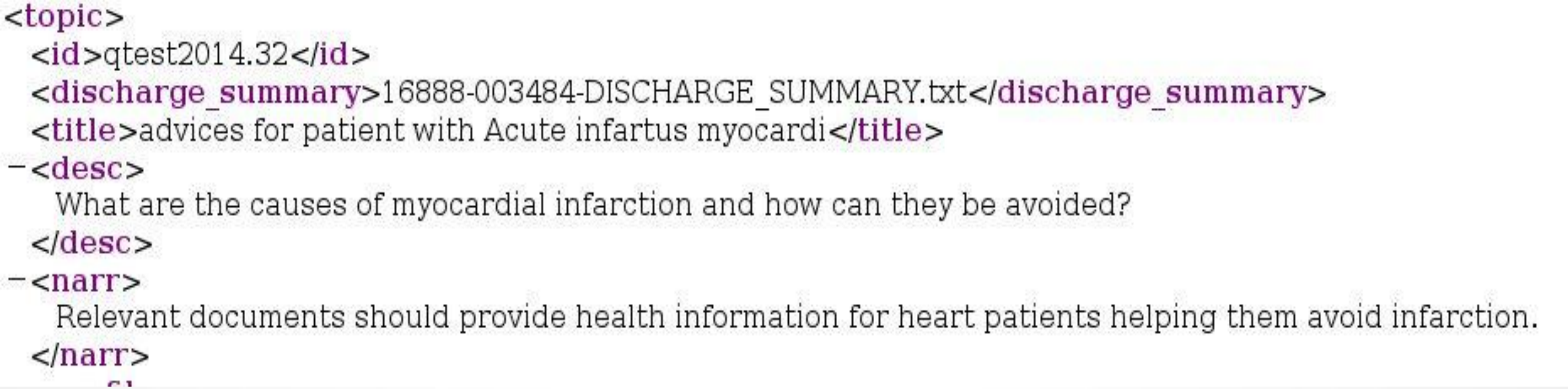
- Users Hate interaction: We can't depend on them to tell us if the result is good or not.
- But we can simulate user's interaction
- Do first retrieval to create an initial pool of top ranked relevant documents

- Top Ranked documents are usually relevant
- Add first X terms from top Y documents to the query and redo the retrieval.
- Y documents have the highest rank, What about first X Terms? … Get the highest tf.idf score
    - tf–idf, is often used as a weighting factor in information retrieval

- Any Danger in this approach?

4.6 Narrative usage

- Narr tags could contain extra important keywords
- Queries were built using text in title and narr tags



```
<topic>
  <id>qtest2014.32</id>
  <discharge_summary>16888-003484-DISCHARGE_SUMMARY.txt</discharge_summary>
  <title>advices for patient with Acute infartus myocardi</title>
-<desc>
   What are the causes of myocardial infarction and how can they be avoided?
  </desc>
-<narr>
   Relevant documents should provide health information for heart patients helping them avoid infarction.
  </narr>
```

4.7 Multilingual Medical information System

- Task 3b extends Task 3a.
- Translate the queries we were given from German, French and Czech into English using Khresmoi system.
- Then we applied the retrieval again.

4.8 Results

- The following results express the MAP (Mean Average Value)
- We used only train queries for these results, still waiting for official results.

Language	baseline	Spell checking	Query Expansion	Narrative usage
English	0.5130	0.5130	0.2866	0.2467
Czech	0.4124	0.4124	0.2096	0.0735
German	0.2088	0.2088	0.1126	0.0737
French	0.3633	0.3633	0.2426	0.2006

5. Conclusion

- We got many insights from our first year participation:
  - Using simple tool for data processing could be better than complicated one.
  - Optimal parameter value in general IR-domain may not be optimal in specific IR-domain
  - Sometimes MT can’t translate the different forms of a term (anoxic, anoxické)
  - We should always do manual checking of data

5. Future work

- Expand Queries using UMLS
  - The Unified Medical Language System (UMLS) is a compendium of many controlled vocabularies in the biomedical sciences.

- Use discharge summaries:
  - contains a brief summary of all important information from the entire hospitalization or stay in the institution, including the discharge diagnosis and often a plan for follow-up care.

- Work more on queries auto correction.

Thanks

Time for questions and answers