Cross-lingual Information Retrieval

Pavel Pecina

Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic

Joint work with Shadi Saleh
Jan 17, 2019 - AFCEA

Outline

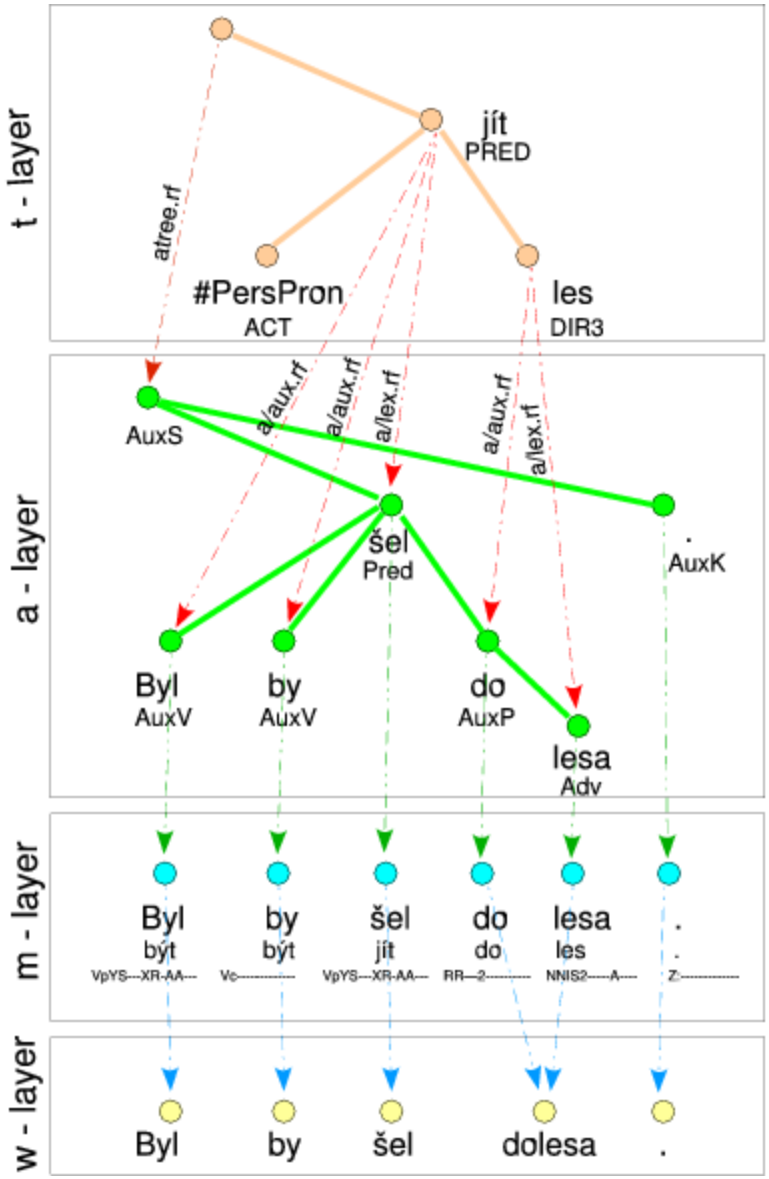
1. Introduction
2. (Cross-lingual) information retrieval
3. Query translation reranking
4. Term selection for query expansion
5. Document-translation vs. Query-translation approach

Institute of Formal and Applied Linguistics

- Founded in 1967 (independent institute since 1991)
- Part of Computer Science School (est. 1991), a part of Faculty of Mathematics and Physics (est. 1952), a part of Charles University
- Staff: 11 Professors, 20 RAs, 36 PhD students
- MSc program (approx. 10 graduates a year)
- Budget: 60 mil CZK/year (teaching + research, national, EU, US)
- Research areas:
  - linguistic theory („dependency“), formal description of language
  - corpus annotation (Prague Dependency Treebanks)
  - Natural Language Processing, Machine Translation, Deep Learning, ...

Prague Dependency Treebanks

- large amounts of texts with complex and interlinked **morphological**, **syntactic** and **tectogrammatical (complex semantic)** annotation
- based on the long-standing Praguian linguistic tradition, adapted for the current computational linguistics research needs
- used for training tools for automatic linguistic processing:
  - https://ufal.mff.cuni.cz/tools
  - https://lindat.mff.cuni.cz



Information Retrieval (IR)

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Towards Cross-Lingual Information Retrieval (CLIR)

Towards Cross-Lingual Information Retrieval (CLIR)

query

bilateral infiltrates and radiography

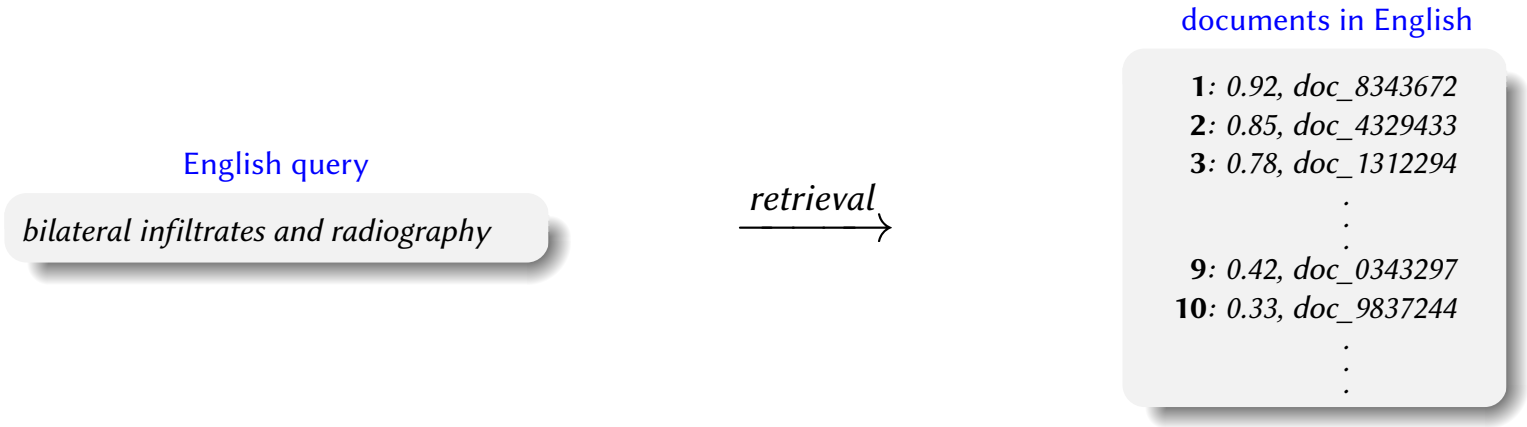
$\xrightarrow{\textit{retrieval}}$

documents

1: *0.92, doc_8343672*
2: *0.85, doc_4329433*
3: *0.78, doc_1312294*
⋮
9: *0.42, doc_0343297*
10: *0.33, doc_9837244*
⋮

Mono-lingual Information Retrieval

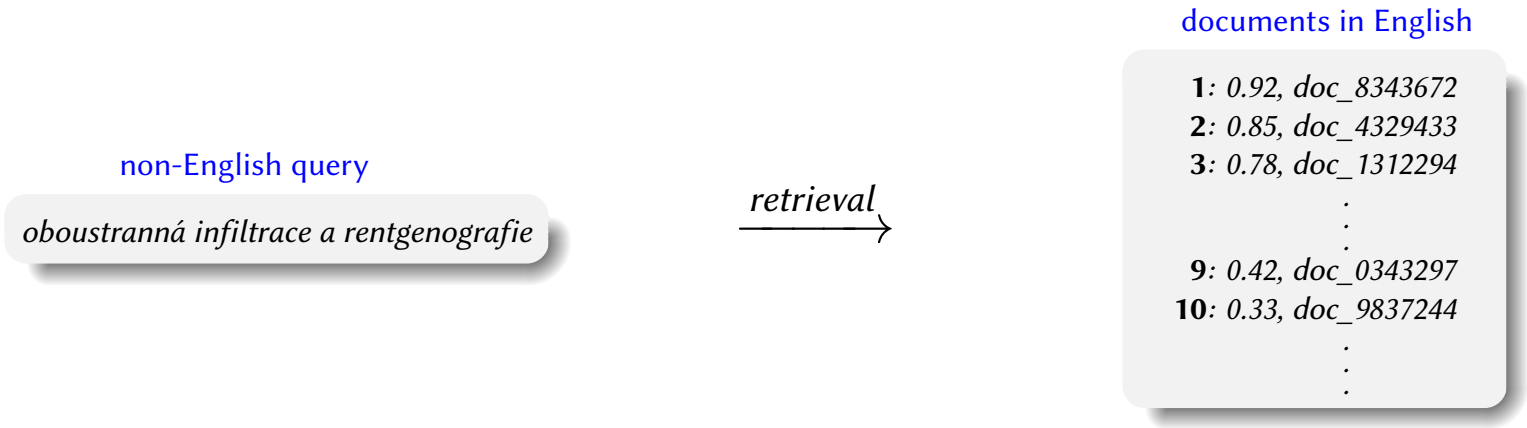
Towards Cross-Lingual Information Retrieval (CLIR)



Mono-lingual Information Retrieval

- Queries and documents in the same language.

Towards Cross-Lingual Information Retrieval (CLIR)



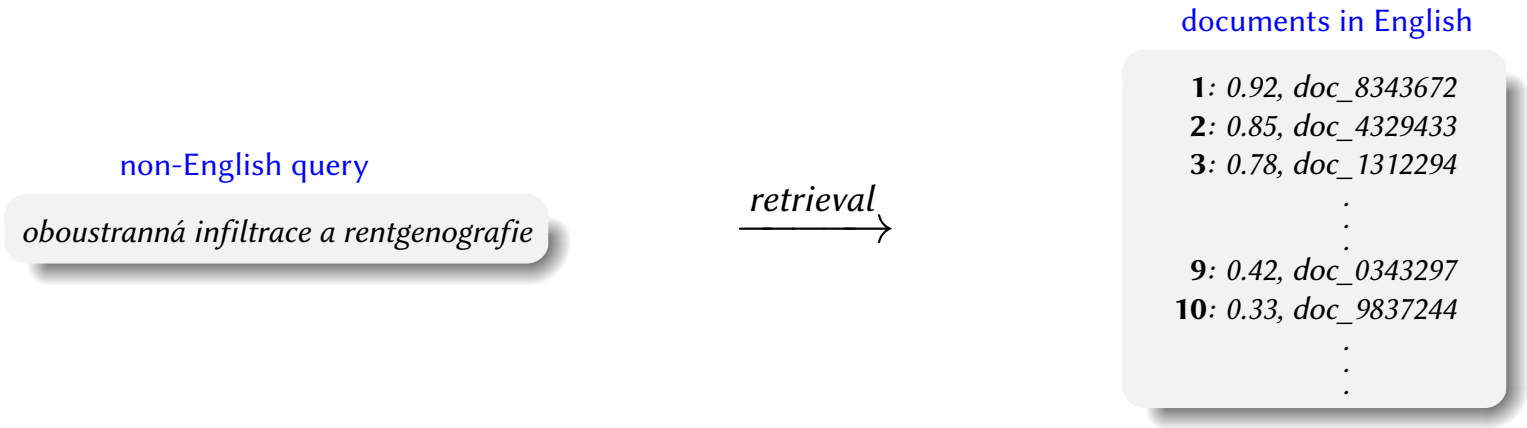
Mono-lingual Information Retrieval

- Queries and documents in the same language.

Cross-lingual Information Retrieval

- Query language differs from the document language.

Towards Cross-Lingual Information Retrieval (CLIR)



Mono-lingual Information Retrieval

- Queries and documents in the same language.

Cross-lingual Information Retrieval

- Query language differs from the document language.
- Useful for:
  a) searching in multilingual collections
  b) users with no/little knowledge of the document language

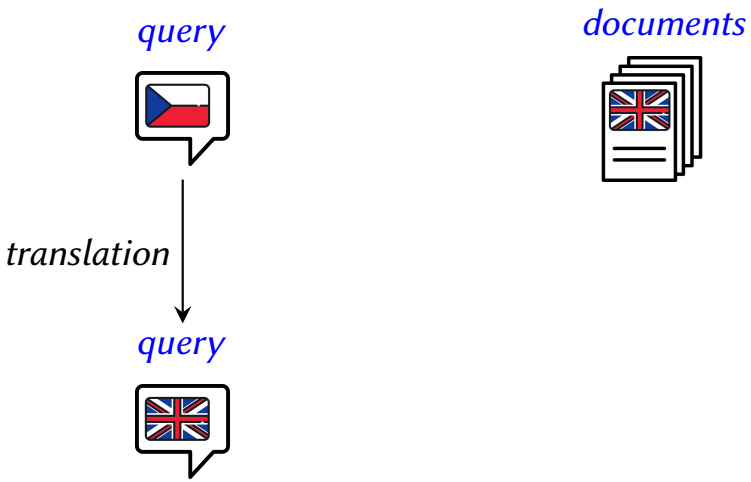
Machine Translation (MT) for CLIR

Machine Translation (MT) for CLIR

query

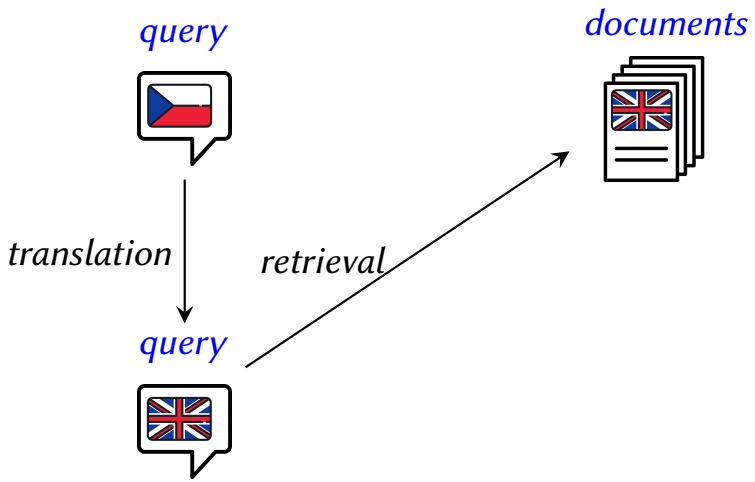
documents

Machine Translation (MT) for CLIR



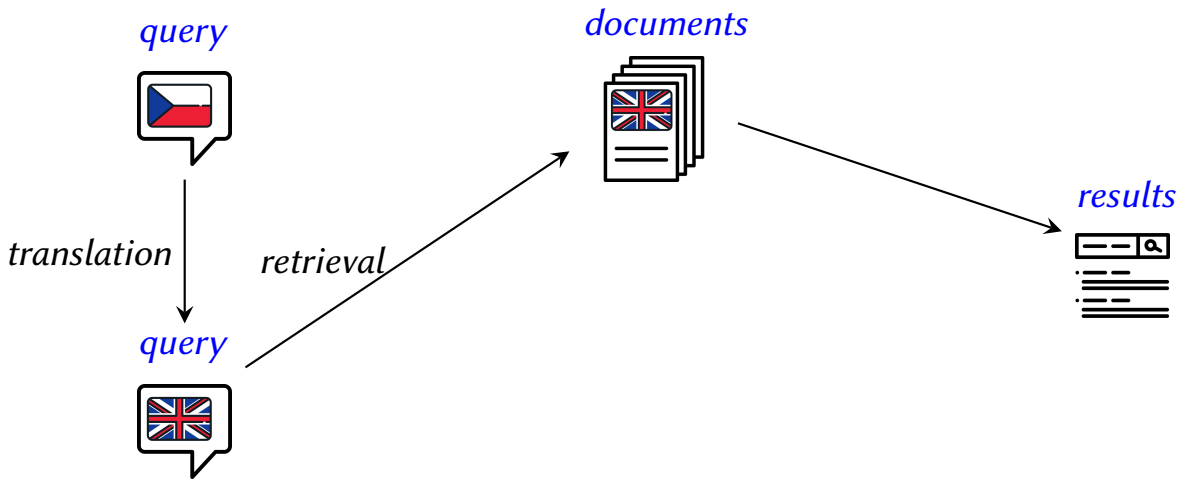
Query translation

Machine Translation (MT) for CLIR



Query translation

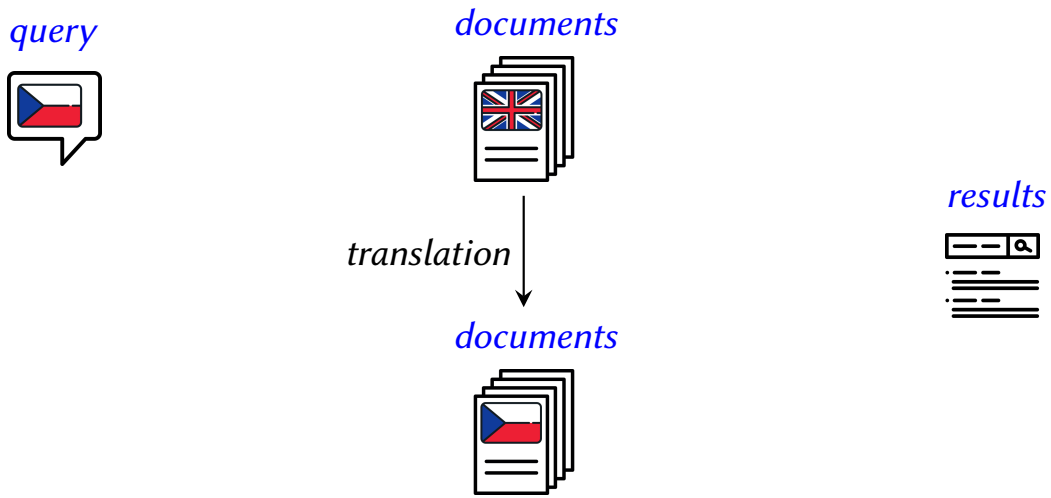
Machine Translation (MT) for CLIR



Query translation

- Query language → document language(s)
- Done at query time
- Multilingual collections: translation into all languages, results merged.

Machine Translation (MT) for CLIR

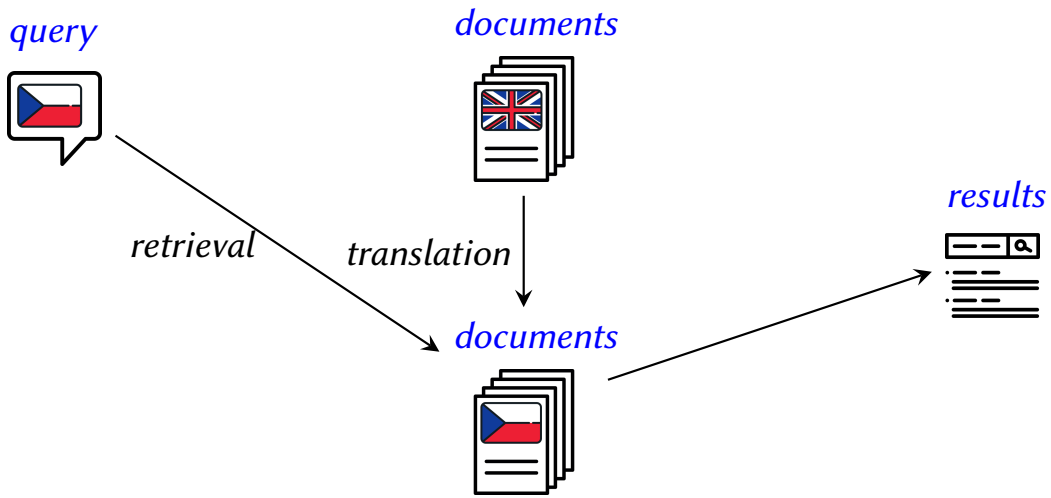


Query translation

- Query language → document language(s)
- Done at query time
- Multilingual collections: translation into all languages, results merged.

Document translation

Machine Translation (MT) for CLIR



Query translation

- Query language → document language(s)
- Done at query time
- Multilingual collections: translation into all languages, results merged.

Document translation

- Document language → query language(s)
- Done prior indexing for all documents
- Index size increases
- Assumed to outperform query translation due to greater context of MT

CLIR in the medical domain (CLEF eHealth IR task)

A series of shared tasks (since 2013) focused on patient-centered IR.
Precision oriented evaluation (evaluation measure: P@10)

CLIR in the medical domain (CLEF eHealth IR task)

A series of shared tasks (since 2013) focused on patient-centered IR.
Precision oriented evaluation (evaluation measure: P@10)

Documents

- single collection used in 2013–2015
- ∼1 million web-pages crawled from English medical websites

CLIR in the medical domain (CLEF eHealth IR task)

A series of shared tasks (since 2013) focused on patient-centered IR.
Precision oriented evaluation (evaluation measure: P@10)

Documents

- single collection used in 2013–2015
- ∼1 million web-pages crawled from English medical websites

Queries

- Generated by medical experts in English to mimic queries of lay people
- Based on: *clinical reports*, *discharge summaries*, *symptoms/conditions*
- Translated into Czech, French, German, Hungarian, Polish, Spanish, Swedish

query id	title
2013.38	*MI and hereditary*
2013.41	*right macular hemorrhage*
2014.1	*coronary artery disease*
2014.6	*aortic stenosis*
2015.1	*many red marks on legs after traveling from US*
2015.57	*infant labored breathing and tight wheezing cough*

Khresmoi Translator

- developed within the Khresmoi project (EU FP7)
- based on phrase-bsed SMT
- provides MT for search and access systems for biomedical information
- languages supported:
  - English $\leftrightarrow$ Czech, French, German, Hungarian, Polish, Spanish, Swedish
- trained on large training data
  - tens of millions of parallel sentences
  - billions of words of monolingual data
- general-domain models interpolated with in-domain models
- in-domain data selected by the perplexity-based method of Moore & Lewis

Khresmoi Translator

- developed within the Khresmoi project (EU FP7)
- based on phrase-bsed SMT
- provides MT for search and access systems for biomedical information
- languages supported:
  - English $\leftrightarrow$ Czech, French, German, Hungarian, Polish, Spanish, Swedish
- trained on large training data
  - tens of millions of parallel sentences
  - billions of words of monolingual data
- general-domain models interpolated with in-domain models
- in-domain data selected by the perplexity-based method of Moore & Lewis
- Specific models for translation of:
  1. full documents – tuned to maximize translation quality (adequacy + fluency)
  2. search queries – tuned to maximize adequacy only

Retrieval system

- based on Terrier (http://terrier.org/)
- language model with Dirichlet prior smoothing
- documents filtered for HTML mark-up
- main evaluation measure: P@10 (ratio of relevant documents among top 10)

Retrieval system

- based on Terrier (http://terrier.org/)
- language model with Dirichlet prior smoothing
- documents filtered for HTML mark-up
- main evaluation measure: P@10 (ratio of relevant documents among top 10)

query

bilateral infiltrates and radiography

retrieval →

top documents

1: *0.92, doc_8343672,*
2: *0.85, doc_4329433,*
3: *0.78, doc_1312294,*
4: *0.72, doc_7511255,*
5: *0.61, doc_3312294,*
6: *0.58, doc_8354296,*
7: *0.57, doc_9312598,*
8: *0.49, doc_5314294,*
9: *0.42, doc_0343297,*
10: *0.33, doc_9837244,*

Retrieval system

- based on Terrier (http://terrier.org/)
- language model with Dirichlet prior smoothing
- documents filtered for HTML mark-up
- main evaluation measure: P@10 (ratio of relevant documents among top 10)

query

bilateral infiltrates and radiography

retrieval →

top documents

1: *0.92, doc_8343672, rel*
2: *0.85, doc_4329433, not*
3: *0.78, doc_1312294, rel*
4: *0.72, doc_7511255, rel*
5: *0.61, doc_3312294, not*
6: *0.58, doc_8354296, rel*
7: *0.57, doc_9312598, not*
8: *0.49, doc_5314294, rel*
9: *0.42, doc_0343297, rel*
10: *0.33, doc_9837244, not*

Retrieval system

- based on Terrier (http://terrier.org/)
- language model with Dirichlet prior smoothing
- documents filtered for HTML mark-up
- main evaluation measure: P@10 (ratio of relevant documents among top 10)

query

bilateral infiltrates and radiography

$\xrightarrow{\textit{retrieval}}$

top documents

1: *0.92, doc_8343672, rel*
2: *0.85, doc_4329433, not*
3: *0.78, doc_1312294, rel*
4: *0.72, doc_7511255, rel*
5: *0.61, doc_3312294, not*
6: *0.58, doc_8354296, rel*
7: *0.57, doc_9312598, not*
8: *0.49, doc_5314294, rel*
9: *0.42, doc_0343297, rel*
10: *0.33, doc_9837244, not*

$$\text{P@10} = \frac{6}{10} = 60\%$$

Reranking Query Translations

MT for query translation

- Standard approach:
  - use MT as a “black box”
  - i.e. use the single best query translation
- Problem:
  - Queries are not “standard” text (short, ungrammatical)
  - MT trained towards translation quality.
  - CLIR evaluated based on retrieval quality.
  - Translation quality may not correlate well with retrieval quality.

Examples of query translation options

query id: 2015.18.cs

src:	*ischemická choroba srdeční*
ref:	*coronary artery disease*
1	*ischaemic heart disease*
2	*ischemic heart disease*
3	*heart disease*
4	*coronary heart disease*
5	*ischaemic disease*
6	*ischemic cardiac disease*
7	*coronary disease*
8	*ischaemic cardiac disease*
9	*ischemic disease*
10	*coronary artery disease*
11	*ischemic cardiac*
12	*cardiac disease*
13	*stroke heart*
14	*heart disease*
15	*ischaemic cardiac*
16	*stroke cardiac*
17	*heart ischaemic disease*
18	*cardiac ischemic disease*
19	*cardiac stroke*
20	*cardiac ischemic*

Examples of query translation options

query id: 2015.18.cs

src:	*ischemická choroba srdeční*
ref:	*coronary artery disease*
1	*ischaemic heart disease*
2	*ischemic heart disease*
3	*heart disease*
4	*coronary heart disease*
5	*ischaemic disease*
6	*ischemic cardiac disease*
7	*coronary disease*
8	*ischaemic cardiac disease*
9	*ischemic disease*
10	*coronary artery disease*
11	*ischemic cardiac*
12	*cardiac disease*
13	*stroke heart*
14	*heart disease*
15	*ischaemic cardiac*
16	*stroke cardiac*
17	*heart ischaemic disease*
18	*cardiac ischemic disease*
19	*cardiac stroke*
20	*cardiac ischemic*

Examples of query translation options

query id: 2015.18.cs

src:	*ischemická choroba srdeční*
ref:	*coronary artery disease*
1	*ischaemic heart disease*
2	*ischemic heart disease*
3	*heart disease*
4	*coronary heart disease*
5	*ischaemic disease*
6	*ischemic cardiac disease*
7	*coronary disease*
8	*ischaemic cardiac disease*
9	*ischemic disease*
10	*coronary artery disease*
11	*ischemic cardiac*
12	*cardiac disease*
13	*stroke heart*
14	*heart disease*
15	*ischaemic cardiac*
16	*stroke cardiac*
17	*heart ischaemic disease*
18	*cardiac ischemic disease*
19	*cardiac stroke*
20	*cardiac ischemic*

Examples of query translation options

query id: 2015.18.cs

src:	*ischemická choroba srdeční*
ref:	*coronary artery disease*
1	*ischaemic heart disease*
2	*ischemic heart disease*
3	*heart disease*
4	*coronary heart disease*
5	*ischaemic disease*
6	*ischemic cardiac disease*
7	*coronary disease*
8	*ischaemic cardiac disease*
9	*ischemic disease*
10	*coronary artery disease*
11	*ischemic cardiac*
12	*cardiac disease*
13	*stroke heart*
14	*heart disease*
15	*ischaemic cardiac*
16	*stroke cardiac*
17	*heart ischaemic disease*
18	*cardiac ischemic disease*
19	*cardiac stroke*
20	*cardiac ischemic*

query id: 2015.11.cs

src:	*bílé povlaky v dutině ústní*
ref:	*white patchiness in mouth*
1	*white coating mouth*
2	*white coating oral*
3	*white coating the mouth*
4	*oral white coating*
5	*white coating in oral cavity*
6	*white coating in mouth*
7	*white sheets oral*
8	*white coatings oral*
9	*white coating in oral*
10	*the white coating mouth*
11	*white coating of mouth*
12	*white sheets mouth*
13	*white coatings mouth*
14	*mouth white coating*
15	*oral white sheets*
16	*white coatings in oral cavity*
17	*white coatings in mouth*
18	*white sheets in oral cavity*
19	*the white coating oral*
20	*white sheets in mouth*

Examples of query translation options

query id: 2015.18.cs

src:	*ischemická choroba srdeční*
ref:	*coronary artery disease*
1	*ischaemic heart disease*
2	*ischemic heart disease*
3	*heart disease*
4	*coronary heart disease*
5	*ischaemic disease*
6	*ischemic cardiac disease*
7	*coronary disease*
8	*ischaemic cardiac disease*
9	*ischemic disease*
10	*coronary artery disease*
11	*ischemic cardiac*
12	*cardiac disease*
13	*stroke heart*
14	*heart disease*
15	*ischaemic cardiac*
16	*stroke cardiac*
17	*heart ischaemic disease*
18	*cardiac ischemic disease*
19	*cardiac stroke*
20	*cardiac ischemic*

query id: 2015.11.cs

src:	*bílé povlaky v dutině ústní*
ref:	*white patchiness in mouth*
1	*white coating mouth*
2	*white coating oral*
3	*white coating the mouth*
4	*oral white coating*
5	*white coating in oral cavity*
6	*white coating in mouth*
7	*white sheets oral*
8	*white coatings oral*
9	*white coating in oral*
10	*the white coating mouth*
11	*white coating of mouth*
12	*white sheets mouth*
13	*white coatings mouth*
14	*mouth white coating*
15	*oral white sheets*
16	*white coatings in oral cavity*
17	*white coatings in mouth*
18	*white sheets in oral cavity*
19	*the white coating oral*
20	*white sheets in mouth*

Examples of query translation options

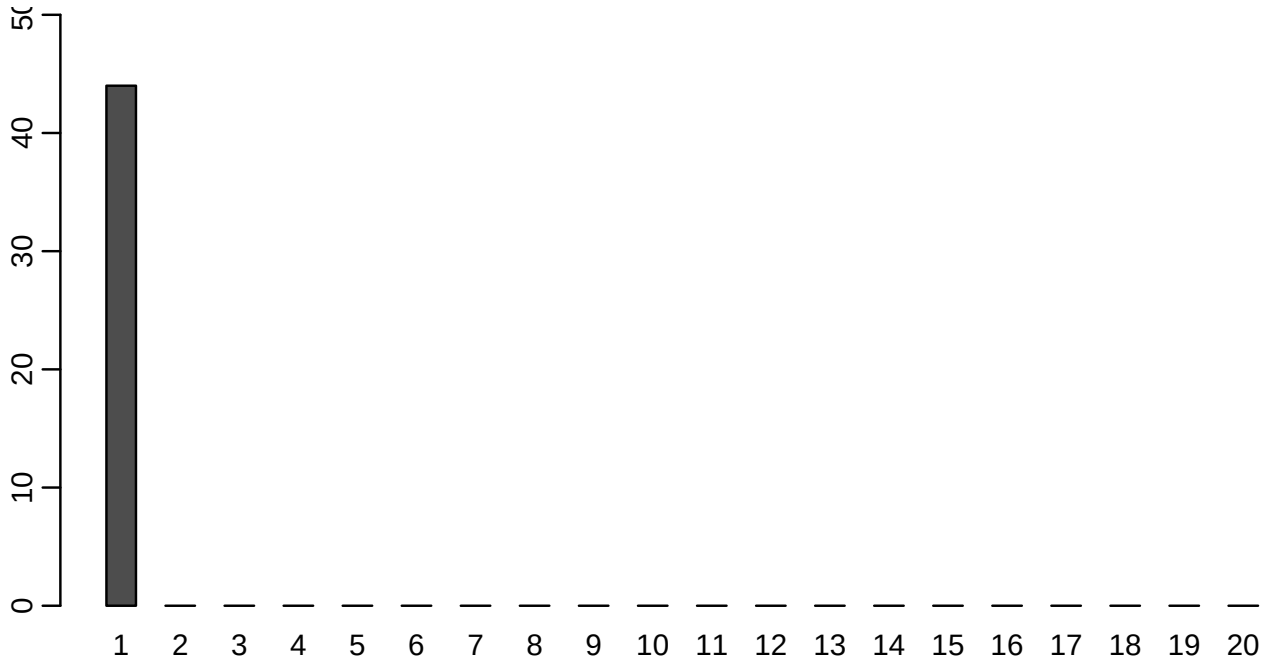
query id: 2015.18.cs

src:	*ischemická choroba srdeční*
ref:	*coronary artery disease*
1	*ischaemic heart disease*
2	*ischemic heart disease*
3	*heart disease*
4	*coronary heart disease*
5	*ischaemic disease*
6	*ischemic cardiac disease*
7	*coronary disease*
8	*ischaemic cardiac disease*
9	*ischemic disease*
10	*coronary artery disease*
11	*ischemic cardiac*
12	*cardiac disease*
13	*stroke heart*
14	*heart disease*
15	*ischaemic cardiac*
16	*stroke cardiac*
17	*heart ischaemic disease*
18	*cardiac ischemic disease*
19	*cardiac stroke*
20	*cardiac ischemic*

query id: 2015.11.cs

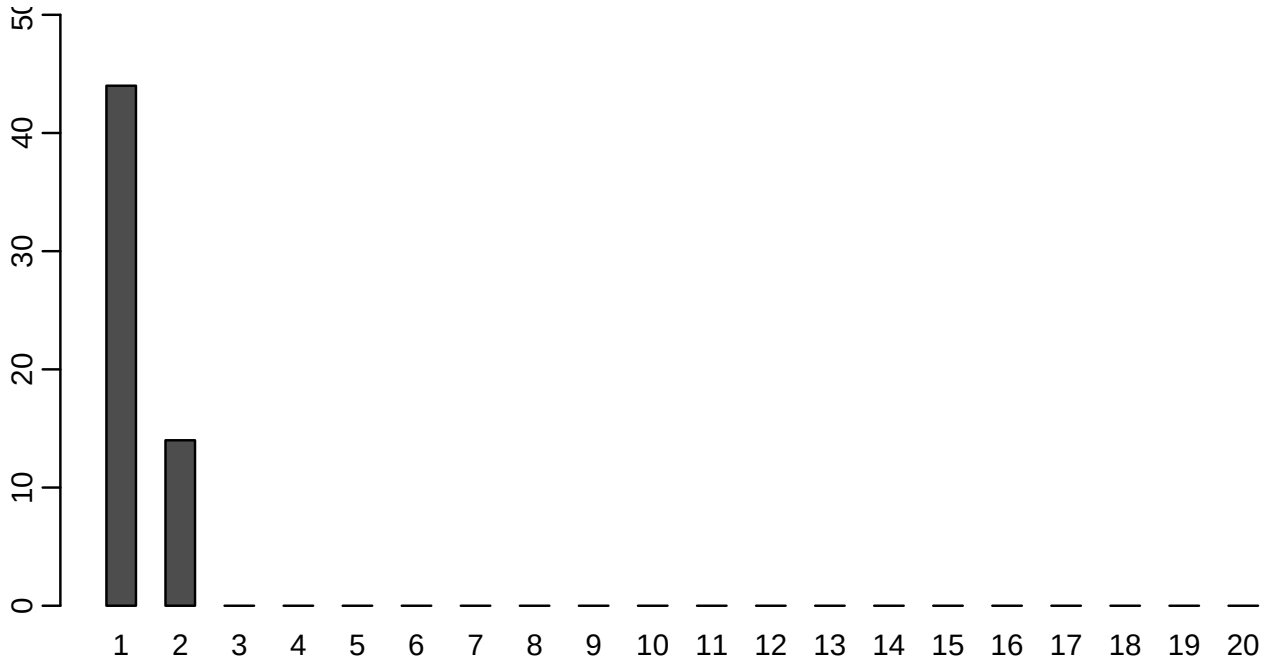
src:	*bílé povlaky v dutině ústní*
ref:	*white patchiness in mouth*
1	*white coating mouth*
2	*white coating oral*
3	*white coating the mouth*
4	*oral white coating*
5	*white coating in oral cavity*
6	*white coating in mouth*
7	*white sheets oral*
8	*white coatings oral*
9	*white coating in oral*
10	*the white coating mouth*
11	*white coating of mouth*
12	*white sheets mouth*
13	*white coatings mouth*
14	*mouth white coating*
15	*oral white sheets*
16	*white coatings in oral cavity*
17	*white coatings in mouth*
18	*white sheets in oral cavity*
19	*the white coating oral*
20	*white sheets in mouth*

Translation quality vs. retrieval quality



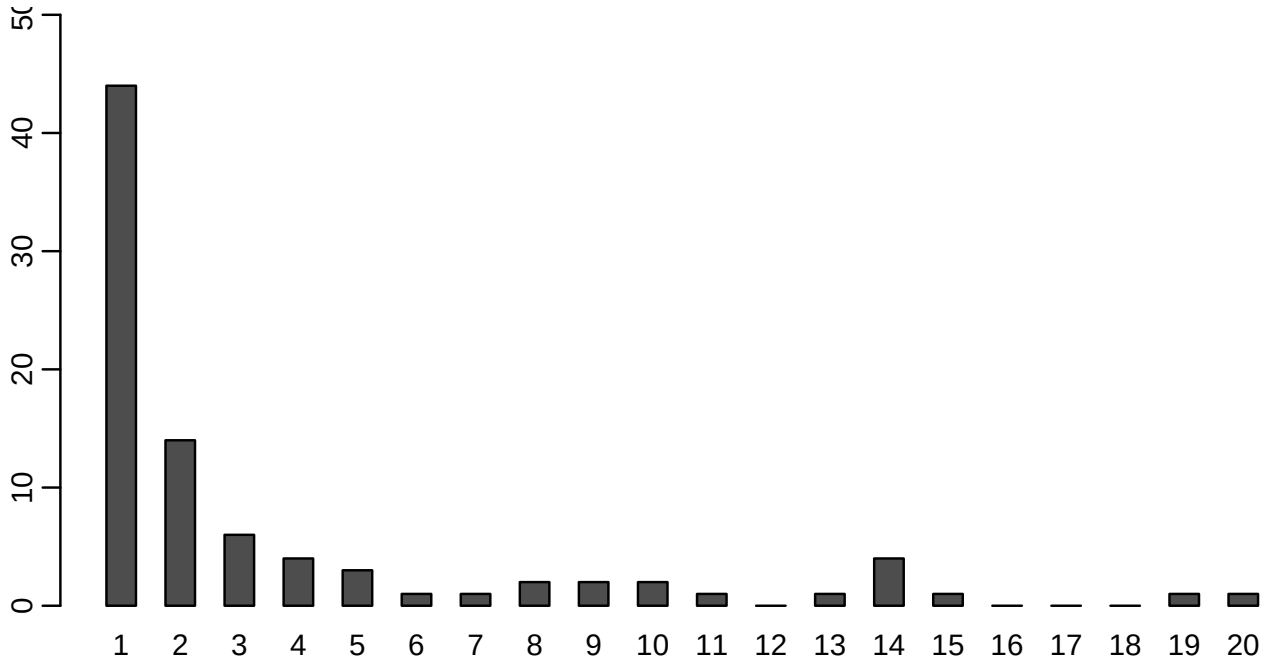
Distribution of the IR-optimal query translations among top 20 MT translations

Translation quality vs. retrieval quality



Distribution of the IR-optimal query translations among top 20 MT translations

Translation quality vs. retrieval quality



Distribution of the IR-optimal query translations among top 20 MT translations

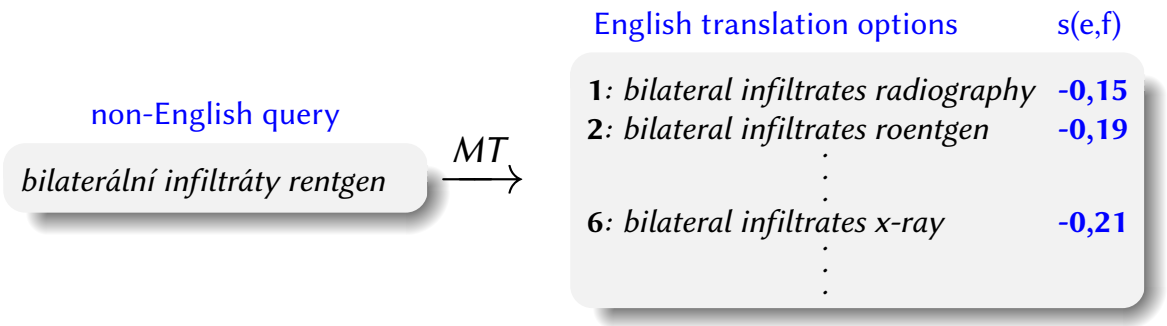
Query translation reranking

Query translation reranking

non-English query

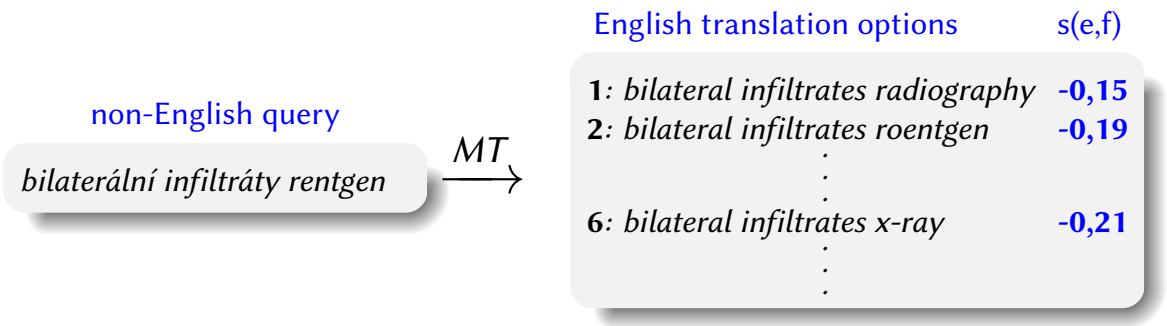
bilaterální infiltráty rentgen

Query translation reranking



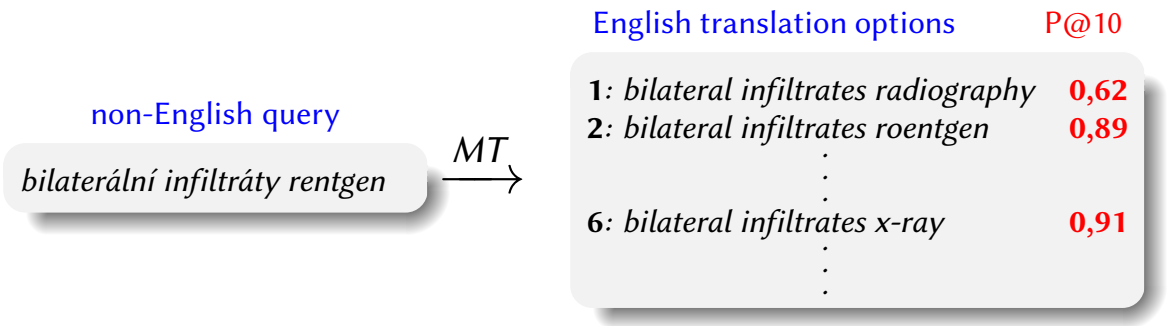
1. MT produces multiple translation options (e.g. 20)

Query translation reranking



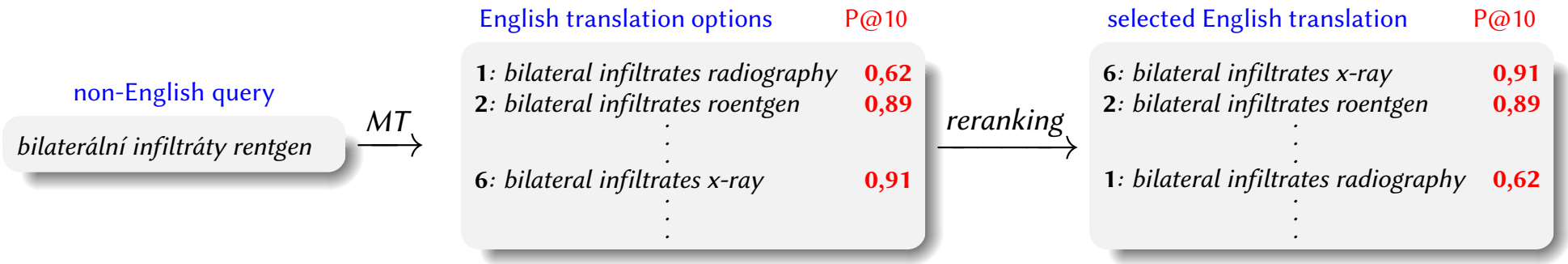
1. MT produces multiple translation options (e.g. 20)
2. Each translation option represented by a vector of features

Query translation reranking



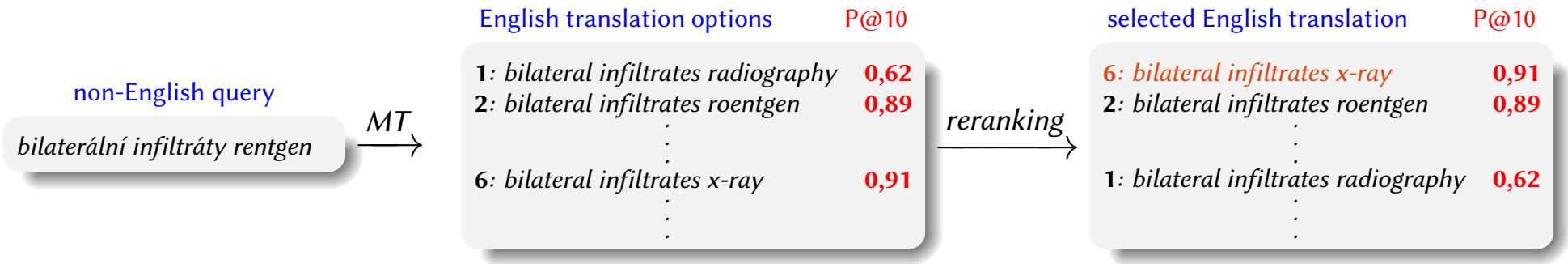
1. MT produces multiple translation options (e.g. 20)
2. Each translation option represented by a vector of features
3. Training instances assigned P@10 score (based on relevance assessment of training queries)
4. A regression model trained to predict P@10 for each translation option

Query translation reranking



1. MT produces multiple translation options (e.g. 20)
2. Each translation option represented by a vector of features
3. Training instances assigned P@10 score (based on relevance assessment of training queries)
4. A regression model trained to predict P@10 for each translation option
5. Reranking according to the predicted P@10 scores

Query translation reranking



1. MT produces multiple translation options (e.g. 20)
2. Each translation option represented by a vector of features
3. Training instances assigned P@10 score (based on relevance assessment of training queries)
4. A regression model trained to predict P@10 for each translation option
5. Reranking according to the predicted P@10 scores
6. The highest-scored translation selected

Regression model features

- MT model features + the total MT score
- Retrieval status value
- Inverse document frequency from the collection
- Term frequency in MT n-best lists
- Term frequency in UMLS thesaurus
- Term frequency in abstracts of 10 Wikipedia articles retrieved as a response to 1-best translation used to query the Wikipedia articles

Query translation reranking: Overall results (2016)

P@10 on test queries

system	Czech	French	German
Monolingual	50.30	50.30	50.30
1-best (“black-box”)	45.61	47.73	42.42
Reranking	**50.15**	**51.06**	**45.30**

Query translation reranking: Overall results (2016)

P@10 on test queries

system	Czech	French	German
Monolingual	50.30	50.30	50.30
1-best (“black-box”)	45.61	47.73	42.42
Reranking	**50.15**	**51.06**	**45.30**
Google	50.91	49.70	49.39
Bing	47.88	48.64	46.52

Query translation reranking: Overall results (2016)

P@10 on test queries

system	Czech	French	German
Monolingual	50.30	50.30	50.30
1-best ("black-box")	45.61	47.73	42.42
Reranking	**50.15**	**51.06**	**45.30**
Google	50.91	49.70	49.39
Bing	47.88	48.64	46.52

- A single model trained on data for all source languages
- $3 \times 100 \times 15 \sim 4000$ training instances (duplicities removed)

Query translation reranking: Examples

- Reranked translation (**rnk**) better than the one selected by SMT (**smt**):

query id: 2014.1.fr		P@10
src:	*maladie coronarienne*	
ref:	*coronary artery disease*	*0.8*
smt:	*CHD*	*0.5*
rnk:	*coronary artery disease*	*0.8*

query id: 2014.1.cs		P@10
src:	*ischemická choroba srdeční*	
ref:	*coronary artery disease*	*0.8*
smt:	*ischaemic heart disease*	*0.7*
rnk:	*coronary heart disease*	*0.8*

Query translation reranking: Examples

- Reranked translation (**rnk**) better than the one selected by SMT (**smt**):

query id: 2014.1.fr		P@10
src:	*maladie coronarienne*	
ref:	*coronary artery disease*	*0.8*
smt:	*CHD*	*0.5*
rnk:	*coronary artery disease*	*0.8*

query id: 2014.1.cs		P@10
src:	*ischemická choroba srdeční*	
ref:	*coronary artery disease*	*0.8*
smt:	*ischaemic heart disease*	*0.7*
rnk:	*coronary heart disease*	*0.8*

- Reranked translation (**rnk**) better than the reference translation (**ref**):

query id: 2015.11.cs		P@10
src:	*bílé povlaky v dutině ústní*	
ref:	*white patchiness in mouth*	*0.1*
smt:	*white coating mouth*	*0.1*
rnk:	*white coating in oral cavity*	*0.8*

query id: 2015.16.fr		P@10
src:	*taches de sang rouges sur les jambes*	
ref:	*red patchy bruising over legs*	*0.1*
smt:	*red blood spots on legs*	*0.1*
rnk:	*blood spots on legs*	*0.2*

Term Selection for Query Expansion

Query translation and expansion: Motivation

query id: 2015.18.cs

src:	*ischemická choroba srdeční*
ref:	*coronary artery disease*
1	*ischaemic heart disease*
2	*ischemic heart disease*
3	*heart disease*
4	*coronary heart disease*
5	*ischaemic disease*
6	*ischemic cardiac disease*
7	*coronary disease*
8	*ischaemic cardiac disease*
9	*ischemic disease*
10	*coronary artery disease*
11	*ischemic cardiac*
12	*cardiac disease*
13	*stroke heart*
14	*heart disease*
15	*ischaemic cardiac*
16	*stroke cardiac*
17	*heart ischaemic disease*
18	*cardiac ischemic disease*
19	*cardiac stroke*
20	*cardiac ischemic*

Query translation and expansion: Motivation

query id: 2015.18.cs

src:	*ischemická choroba srdeční*
ref:	*coronary artery disease*
1	*ischaemic heart disease*
2	*ischemic heart disease*
3	*heart disease*
4	*coronary heart disease*
5	*ischaemic disease*
6	*ischemic cardiac disease*
7	*coronary disease*
8	*ischaemic cardiac disease*
9	*ischemic disease*
10	*coronary artery disease*
11	*ischemic cardiac*
12	*cardiac disease*
13	*stroke heart*
14	*heart disease*
15	*ischaemic cardiac*
16	*stroke cardiac*
17	*heart ischaemic disease*
18	*cardiac ischemic disease*
19	*cardiac stroke*
20	*cardiac ischemic*

Query translation and expansion: Motivation

query id: 2015.18.cs

src:	*ischemická choroba srdeční*
ref:	*coronary artery disease*
1	*ischaemic heart disease*
2	*ischemic heart disease*
3	*heart disease*
4	*coronary heart disease*
5	*ischaemic disease*
6	*ischemic cardiac disease*
7	*coronary disease*
8	*ischaemic cardiac disease*
9	*ischemic disease*
10	*coronary artery disease*
11	*ischemic cardiac*
12	*cardiac disease*
13	*stroke heart*
14	*heart disease*
15	*ischaemic cardiac*
16	*stroke cardiac*
17	*heart ischaemic disease*
18	*cardiac ischemic disease*
19	*cardiac stroke*
20	*cardiac ischemic*

query id: 2015.11.cs

src:	*bílé povlaky v dutině ústní*
ref:	*white patchiness in mouth*
1	*white coating mouth*
2	*white coating oral*
3	*white coating the mouth*
4	*oral white coating*
5	*white coating in oral cavity*
6	*white coating in mouth*
7	*white sheets oral*
8	*white coatings oral*
9	*white coating in oral*
10	*the white coating mouth*
11	*white coating of mouth*
12	*white sheets mouth*
13	*white coatings mouth*
14	*mouth white coating*
15	*oral white sheets*
16	*white coatings in oral cavity*
17	*white coatings in mouth*
18	*white sheets in oral cavity*
19	*the white coating oral*
20	*white sheets in mouth*

Query translation and expansion: Motivation

query id: 2015.18.cs

src:	*ischemická choroba srdeční*
ref:	*coronary artery disease*
1	*ischaemic heart disease*
2	*ischemic heart disease*
3	*heart disease*
4	*coronary heart disease*
5	*ischaemic disease*
6	*ischemic cardiac disease*
7	*coronary disease*
8	*ischaemic cardiac disease*
9	*ischemic disease*
10	*coronary artery disease*
11	*ischemic cardiac*
12	*cardiac disease*
13	*stroke heart*
14	*heart disease*
15	*ischaemic cardiac*
16	*stroke cardiac*
17	*heart ischaemic disease*
18	*cardiac ischemic disease*
19	*cardiac stroke*
20	*cardiac ischemic*

query id: 2015.11.cs

src:	*bílé povlaky v dutině ústní*
ref:	*white patchiness in mouth*
1	*white coating mouth*
2	*white coating oral*
3	*white coating the mouth*
4	*oral white coating*
5	*white coating in oral cavity*
6	*white coating in mouth*
7	*white sheets oral*
8	*white coatings oral*
9	*white coating in oral*
10	*the white coating mouth*
11	*white coating of mouth*
12	*white sheets mouth*
13	*white coatings mouth*
14	*mouth white coating*
15	*oral white sheets*
16	*white coatings in oral cavity*
17	*white coatings in mouth*
18	*white sheets in oral cavity*
19	*the white coating oral*
20	*white sheets in mouth*

Term selection for query expansion

1. Candidate terms extracted from:
   - SMT query translation options (*n-best-list*)
   - Wikipedia (10 documents retrieved using the baseline traslation)
   - PubMed (10 documents) – *didn't work*
2. Each candidate term scored by a regression model to predict P@10
3. Candidates with the predicted score above a treshold used for expansion.

Model features include:

- Inverse document frequency from the collection
- Term frequency in SMT n-best lists, Wikipedia results, PubMed results
- Retrieval status value
- Term frequency in UMLS thesaurus
- Word embedding similarity to the 1-best query translation terms

Query expansion: Overall results

P@10 on test queries

system	Czech	French	German
Monolingual	53.03	53.03	53.03
1-best (“black-box”)	47.27	48.03	44.24
1-best+QE	**52.58**	**49.55**	**47.12**

Query expansion: Overall results

P@10 on test queries

system	Czech	French	German
Monolingual	53.03	53.03	53.03
1-best (“black-box”)	47.27	48.03	44.24
1-best+QE	**52.58**	**49.55**	**47.12**
Reranking	49.09	53.64	46.67
Reranking+QE	**53.18**	**50.00**	**46.52**

Query expansion: Overall results

P@10 on test queries

system	Czech	French	German
Monolingual	53.03	53.03	53.03
1-best ("black-box")	47.27	48.03	44.24
1-best+QE	**52.58**	**49.55**	**47.12**
Reranking	49.09	53.64	46.67
Reranking+QE	**53.18**	**50.00**	**46.52**

- A single model trained on data for all source languages
- ~ 4000 training instances

Query expansion: Examples

query id: 2015.18.cs P@10

src:	*špatné držení těla a rovnováha s třesem*	
ref:	*poor gait and balance with shaking*	*0.5*
smt:	*bad posture and balance with tremor*	*0.6*
exp:	*+poor +shaking*	*0.7*

query id: 2015.50.cs P@10

src:	*červená skvrna obličej kojenec*	
ref:	*red spot baby face*	*0.4*
smt:	*red face infants*	*0.2*
exp:	*+baby +stain +spot*	*0.6*

query id: 2015.61.cs P@10

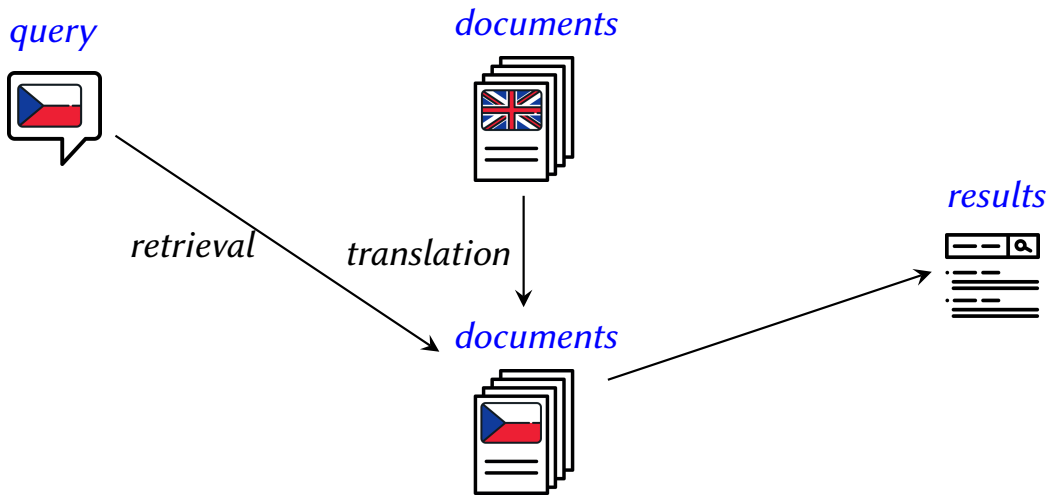
src:	*krvácení pod nehty*	
ref:	*fingernail bruises*	*0.4*
smt:	*bleeding under nails*	*0.4*
exp:	*+fingernails +blood*	*0.6*

query id: 2014.21.fr P@10

src:	*insuffisance rénale*	
ref:	*renal failure*	*0.1*
smt:	*renal impairment*	*0.0*
exp:	*+kidney +disease +function +dysfunction*	
	+failure +insufficiency +deficiency +poor	*0.3*

Query Translation vs. Document Translation

Query translation vs. document translation



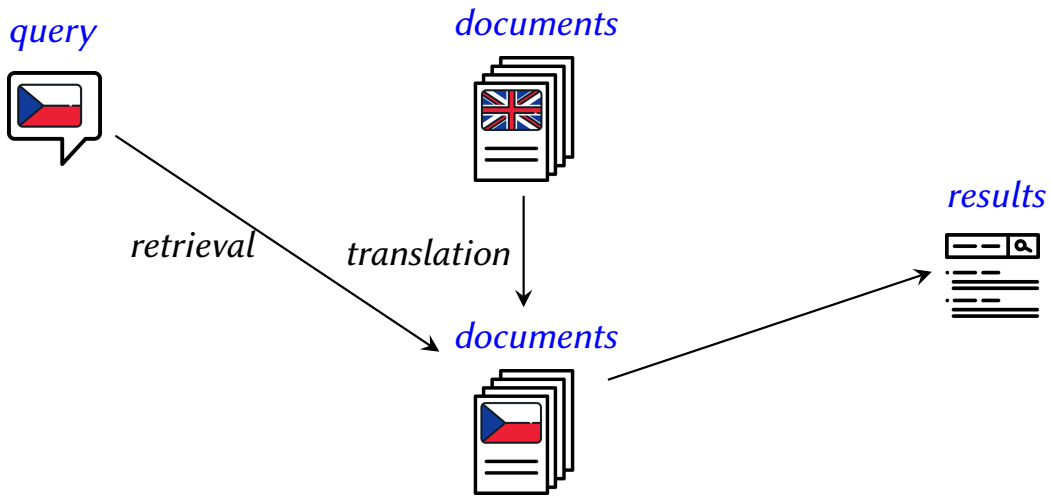
Query translation

- Query language → document language(s)
- Done at query time
- Multilingual collections: translation into all languages, results merged.

Document translation

- Document language → query language(s)
- Done prior indexing for all documents
- Index size increases
- Assumed to outperform query translation due to greater context of MT

Query translation vs. document translation



Query translation

- Query language → document language(s)
- Done at query time
- Multilingual collections: translation into all languages, results merged.

Document translation

- Document language → query language(s)
- Done prior indexing for all documents
- Index size increases
- Assumed to outperform query translation due to greater context of MT

Query translation vs. document translation: Results

Three systems based on Khresmoi Translator evaluated:

- A: Plain system in *document translation* mode
- B: Post-lemmatization of output of A
- C: Pre-lemmatization of training data of A

Query translation vs. document translation: Results

Three systems based on Khresmoi Translator evaluated:

- A: Plain system in *document translation* mode
- B: Post-lemmatization of output of A
- C: Pre-lemmatization of training data of A

P@10 on test queries.

	Czech	French	German
QT-baseline	47.27	48.03	44.24
QT-reranker	48.03	51.67	46.21
DT-form (A)	**38.03**	**42.73**	**37.88**
DT-post-lemma (B)	**40.76**	**41.36**	**38.18**
DT-pre-lemma (C)	**42.88**	**43.18**	**39.85**

Query translation vs. document translation: Examples

Query translation vs. document translation: Examples

- Lemmatized Document translation (**dt**) better than query translation (**qt**):

query id: 2013.47.fr		P@10
src:	*ulcère sacré et soins*	
ref:	*sacral ulcer and care*	*0.2*
qt:	*sacral ulcer care*	*0.2*
dt:	*ulcère sacré et soin*	*0.3*

query id: 2014.24.fr		P@10
src:	*diabète de type 1 et problèmes cardiaques*	
ref:	*diabetes type 1 and heart problems*	*0.4*
qt:	*type 1 diabetes and heart problems*	*0.4*
dt:	*diabète de type 1 et problème cardiaque*	*0.8*

Query translation vs. document translation: Examples

- Lemmatized Document translation (**dt**) better than query translation (**qt**):

query id: 2013.47.fr		P@10
src:	*ulcère sacré et soins*	
ref:	*sacral ulcer and care*	*0.2*
qt:	*sacral ulcer care*	*0.2*
dt:	*ulcère sacré et soin*	*0.3*

query id: 2014.24.fr		P@10
src:	*diabète de type 1 et problèmes cardiaques*	
ref:	*diabetes type 1 and heart problems*	*0.4*
qt:	*type 1 diabetes and heart problems*	*0.4*
dt:	*diabète de type 1 et problème cardiaque*	*0.8*

- Query translation (**qt**) better than lemmatized document translation (**dt**):

query id: 2013.7.fr		P@10
src:	*convulsions et syndrome de sevrage alcoolique*	
ref:	*seizures and alcohol withdrawal syndrome*	*0.3*
qt:	*seizures and alcohol withdrawal syndrome*	*0.3*
dt:	*convulsion et syndrome de sevrage alcoolique*	*0.2*

query id: 2015.33.fr		P@10
src:	*řezná rána a péče*	
ref:	*incision and care*	*0.5*
qt:	*cut and care*	*0.2*
dt:	*řezný rána a péče*	*0.1*

Thank you