Breaking the Language Barrier in Health-related Web Search

Pavel Pecina

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

June 7, 2019 – CLUBS Final Project Workshop, DFKI, Saarbrücken

Outline

1. Introduction
2. Query translation reranking
3. Term selection for query expansion
4. Document translation vs. query translation
5. Conclusions

Context of EU projects

khresmoi

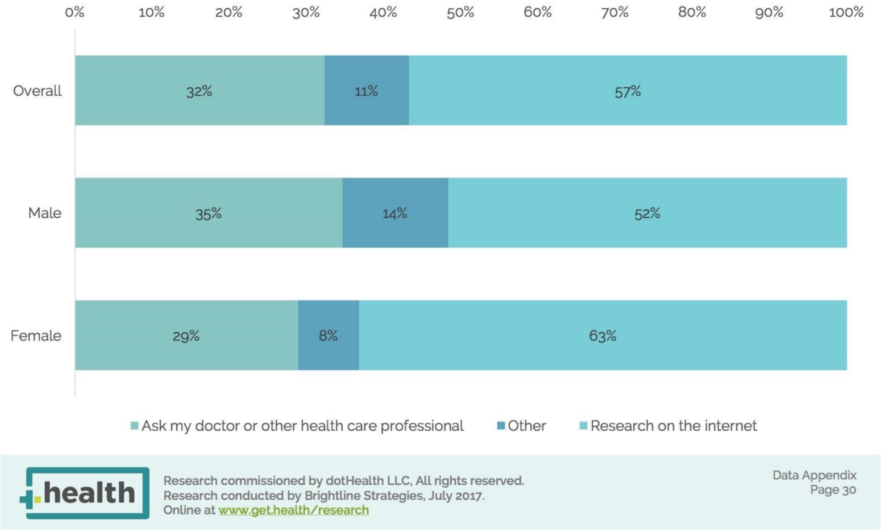
MEDICAL INFORMATION ANALYSIS & RETRIEVAL

- Multi-lingual multi-modal search and access to biomedical information and documents.
- EU FP7, 2010–2014, 12 partners, 10 mil EUR

KCONNECT

- Semantic annotation, search and machine translation of electronic health records and medical publications.
- EU H2020, 2015–2017, 10 partners, 4 mil EUR

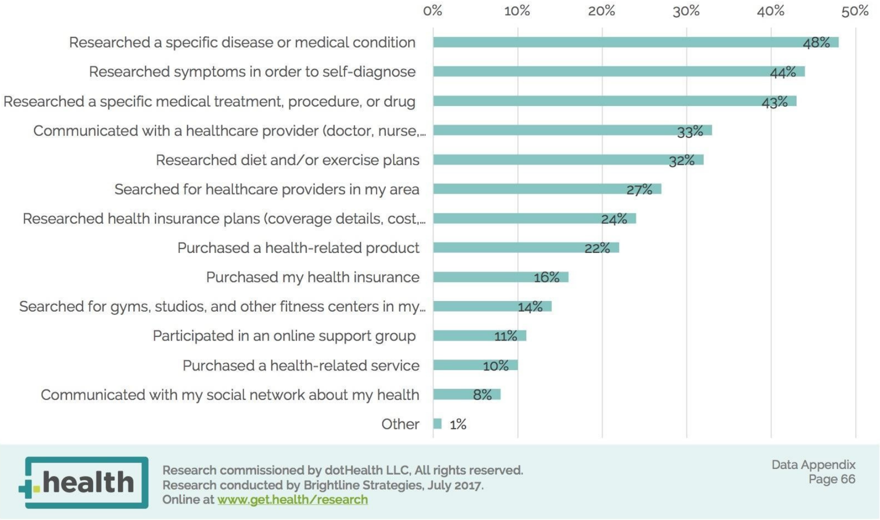
Internet as a source of health-related information



Research commissioned by dotHealth LLC, All rights reserved.
Research conducted by Brightline Strategies, July 2017.
Online at www.get.health/research

Data Appendix
Page 30

Health-related activities online

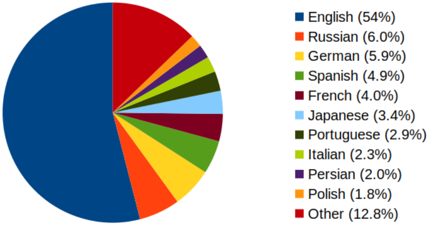


Research commissioned by dotHealth LLC, All rights reserved.
Research conducted by Brightline Strategies, July 2017.
Online at www.get.health/research

Data Appendix
Page 66

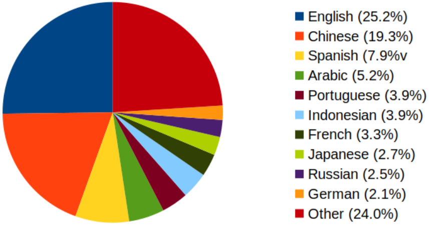
Languages used on the Internet

Content languages



(W3Techs, March 2018)

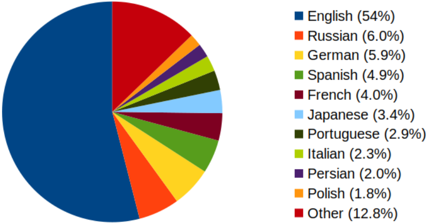
User languages



(InternetWorldStats, April 2019)

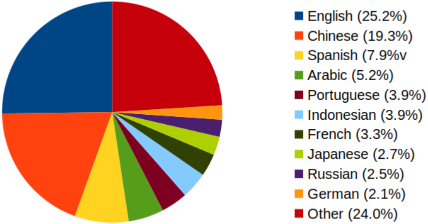
Languages used on the Internet

Content languages



(W3Techs, March 2018)

User languages

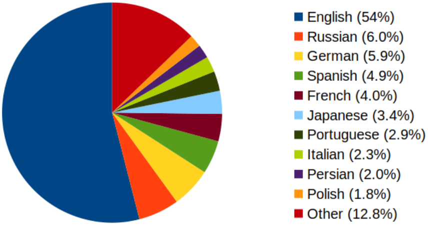


(InternetWorldStats, April 2019)

- Most Internet content is in English (54%)
- Most users don't speak English (75%) ...

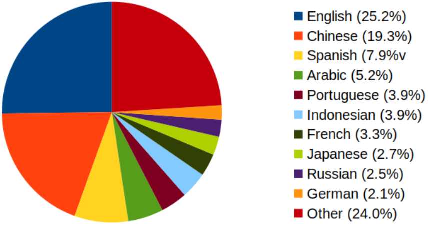
Languages used on the Internet

Content languages



(W3Techs, March 2018)

User languages



(InternetWorldStats, April 2019)

- Most Internet content is in English (54%)
- Most users don't speak English (75%) ...

... still increasing due to the development in third world countries

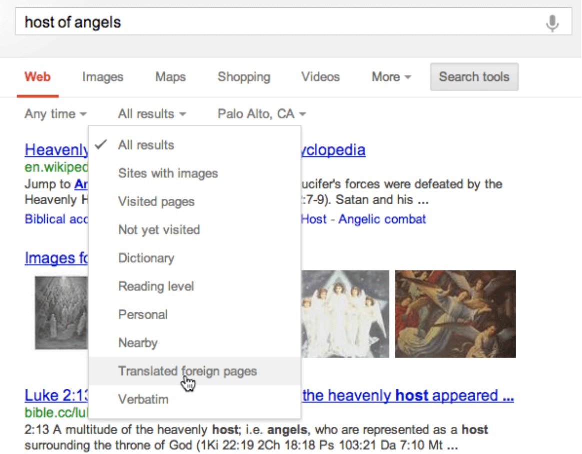
Need for cross-lingual web search in medical domain

- “Although Internet users are often well-educated, there was a strong preference for searching for health and food information in the local language, rather than English” *(Journal of Medical Internet Research, 2007)*

- “The trend towards monolingualism is far from decreasing, with the hegemonic use of one language, English.” *(Bulletin of the World Health Organization, 2015)*

- “Online Hispanics have a hard time finding health information in Spanish” *(comScore, 2011)*

Cross-lingual search option by Google



host of angels

Heavenly ... yclopedia
en.wikiped
Jump to Ar
Heavenly H ... ucifer's forces were defeated by the ... :7-9). Satan and his ...
Biblical acc ... Host - Angelic combat

Images fo

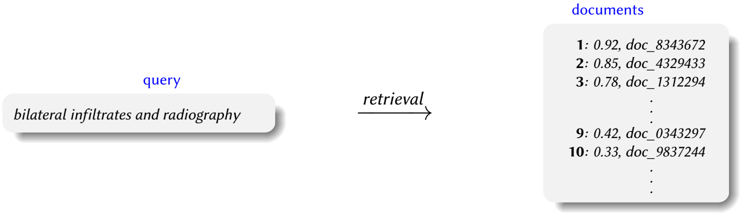
Luke 2:13 ... the heavenly **host** appeared ...
bible.cc/lu
2:13 A multitude of the heavenly **host**; i.e. **angels**, who are represented as a **host** surrounding the throne of God (1Ki 22:19 2Ch 18:18 Ps 103:21 Da 7:10 Mt ...

(http://searchresearch1.blogspot.com)

- ... dropped in 2013 due to lack of use.

Cross-Lingual Information Retrieval (CLIR)

Cross-Lingual Information Retrieval (CLIR)



query

bilateral infiltrates and radiography

$\xrightarrow{retrieval}$

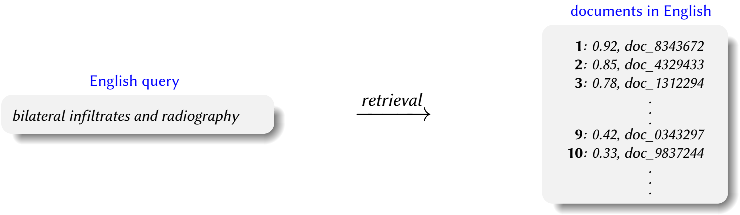
documents

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*
⋮

Information Retrieval

- Searching for relevant documents within a large collection

Cross-Lingual Information Retrieval (CLIR)



English query

bilateral infiltrates and radiography

retrieval

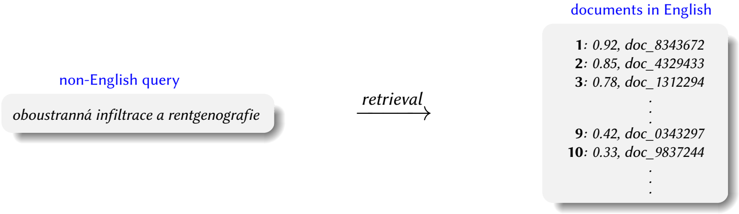
documents in English

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*
⋮

Information Retrieval

- Searching for relevant documents within a large collection
- Mono-lingual: queries and documents in the same language.

Cross-Lingual Information Retrieval (CLIR)



non-English query

oboustranná infiltrace a rentgenografie

$\xrightarrow{retrieval}$

documents in English

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*
⋮

Information Retrieval

- Searching for relevant documents within a large collection
- Mono-lingual: queries and documents in the same language.

Cross-lingual Information Retrieval

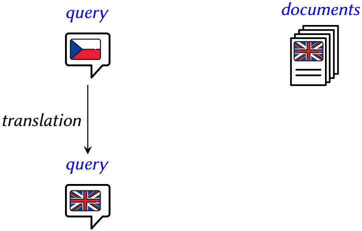
- Query language differs from the document language.
- Useful for:
  - a) searching in multilingual collections
  - b) users with no/little knowledge of the document language

Machine Translation for CLIR

Machine Translation for CLIR

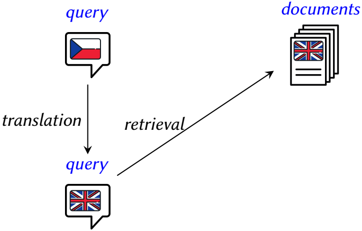
query
documents

Machine Translation for CLIR



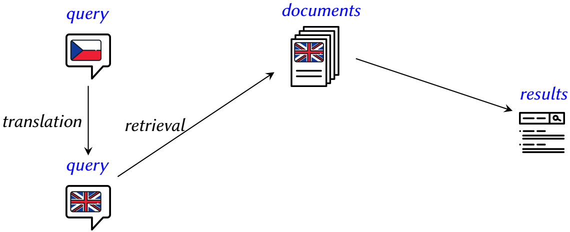
Query translation

Machine Translation for CLIR



Query translation

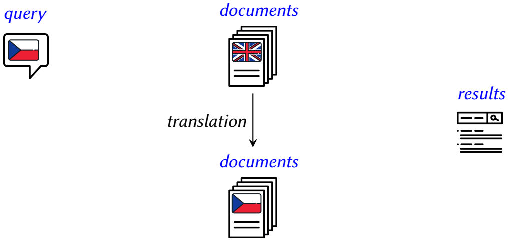
Machine Translation for CLIR



Query translation

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

Machine Translation for CLIR

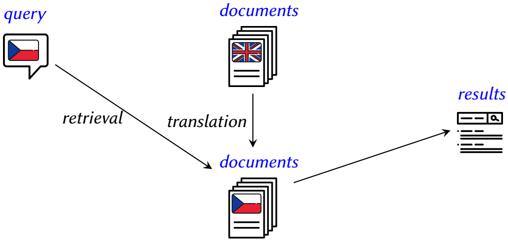


Query translation

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

Document translation

Machine Translation for CLIR



Query translation

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

Document translation

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

CLEF eHealth IR task 2013–2015

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

CLEF eHealth IR task 2013–2015

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

Documents

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

CLEF eHealth IR task 2013–2015

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

Documents

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

Queries

- Generated by medical experts in English to mimic queries of lay people
- Based on: *clinical reports* (50 queries, 2013), *discharge summaries* (50 queries, 2014), *symptoms/conditions* (66 queries, 2015)

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*

CLEF eHealth IR task 2013–2015: CLIR subtask

Official activity

- Part of CELF eHealth in 2014 and 2015
- English queries manually translated by medical experts to other languages:
  - Czech, French, German (2014)
  - Czech, French, German, Arabic, Farsi (2015)

Our extension

- All queries in Czech, French, German, Hungarian, Polish, Spanish, Swedish
- Random (ballanced) split into 100 training queries and 66 test queries.
- Additional rel. assessment of documents highly ranked in our experiments
- Transaltion and assessment done by medical experts

	2013	2014	2015	extension
relevant docs.	1,174	3,209	2,515	2,517
irrelevant docs.	3,676	3,591	9,576	11,851

available from: http://hdl.handle.net/11234/1-2925

Khresmoi Translator

- developed within the Khresmoi project
- based on phrase-bsed SMT (Moses)
- provides MT for search and access systems for biomedical information
- languages supported:
  - English ↔ 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

khresmoi
MEDICAL INFORMATION ANALYSIS & RETRIEVAL

- developed within the Khresmoi project
- based on phrase-bsed SMT (Moses)
- provides MT for search and access systems for biomedical information
- languages supported:
  - English ↔ 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 on parallel sentences to maximize BLEU
  2. search queries – tuned on parallel queries to maximize PER (fluency ignored)

Khresmoi Translator

khresmoi
MEDICAL INFORMATION ANALYSIS & RETRIEVAL

- developed within the Khresmoi project
- based on phrase-bsed SMT (Moses)
- provides MT for search and access systems for biomedical information
- languages supported:
  - English ↔ 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 on parallel sentences to maximize BLEU
  2. search queries – tuned on parallel queries to maximize PER (fluency ignored)

$$\boxed{s(\mathbf{e},\mathbf{f}) = \sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e},\mathbf{f})}$$

Retrieval system

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

Reranking Query Translations

SMT for query translation

- Standard approach:
    - use SMT as a “black box”
    - i.e. use the single best query translation
- Problem:
    - Queries are not “standard” text (short, ungrammatical)
    - SMT trained towards translation quality (e.g. BLEU).
    - CLIR evaluated based on retrieval quality (e.g. P@10).
    - Translation quality (BLEU) may not correlate well with retrieval quality (P@10).

Examples of query translation options by SMT (20-best-list)

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 by SMT (20-best-list)

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 by SMT (20-best-list)

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 by SMT (20-best-list)

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 by SMT (20-best-list)

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 by SMT (20-best-list)

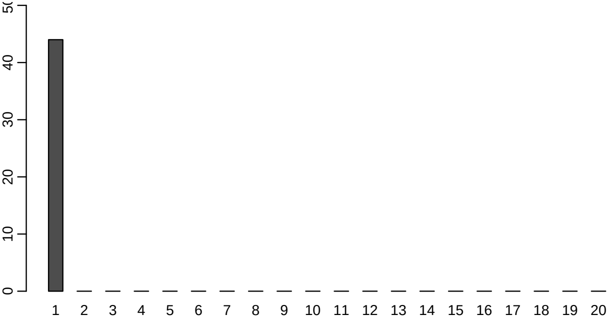
query id: 2015.18.cs

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

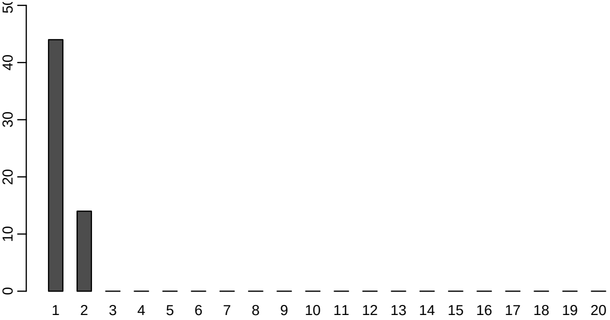
src:	*bílé povlaky v dutině ústní*
ref:	*white patchiness in mouth*
1	*white coating mouth*
2	*white coating oral*
3	*white coating the mouth*
4	*oral white coating*
5	*white coating in oral cavity*
6	*white coating in mouth*
7	*white sheets oral*
8	*white coatings oral*
9	*white coating in oral*
10	*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 comparison



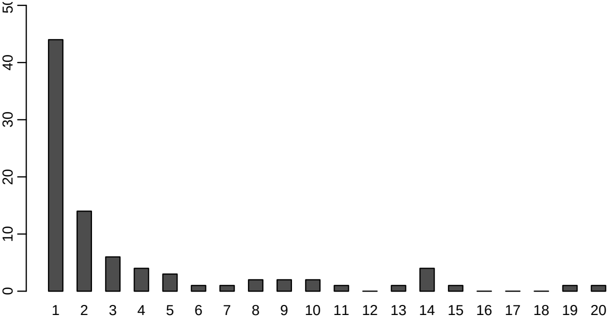
Distribution of the IR-optimal query translations among top 20 SMT translations

Translation quality vs. retrieval quality comparison



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

Translation quality vs. retrieval quality comparison



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

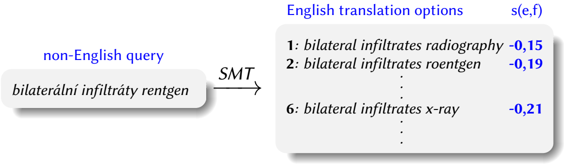
Query translation reranking

Query translation reranking

non-English query

bilaterální infiltráty rentgen

Query translation reranking



non-English query

bilaterální infiltráty rentgen

$\xrightarrow{SMT}$

English translation options s(e,f)

1: *bilateral infiltrates radiography* **-0,15**

2: *bilateral infiltrates roentgen* **-0,19**

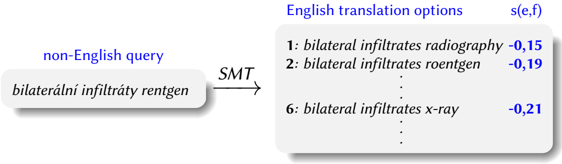
⋮

6: *bilateral infiltrates x-ray* **-0,21**

⋮

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

Query translation reranking



non-English query

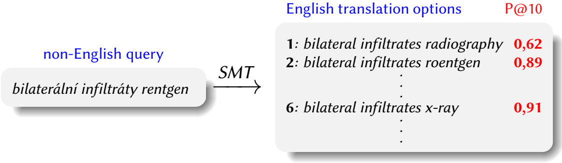
bilaterální infiltráty rentgen

$\xrightarrow{SMT}$

English translation options	s(e,f)
1: *bilateral infiltrates radiography*	**-0,15**
2: *bilateral infiltrates roentgen*	**-0,19**
⋮	
6: *bilateral infiltrates x-ray*	**-0,21**
⋮	

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

Query translation reranking



non-English query

bilaterální infiltráty rentgen

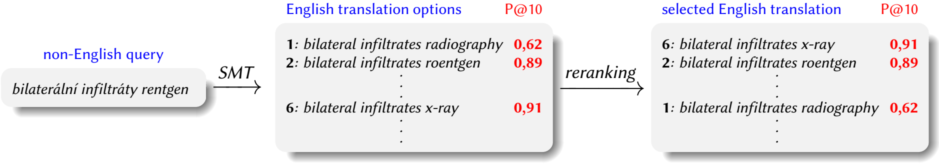
$\xrightarrow{SMT}$

English translation options

English translation options	P@10
1: *bilateral infiltrates radiography*	**0,62**
2: *bilateral infiltrates roentgen*	**0,89**
⋮	
6: *bilateral infiltrates x-ray*	**0,91**
⋮	

1. SMT 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



non-English query

bilaterální infiltráty rentgen

$\xrightarrow{SMT}$

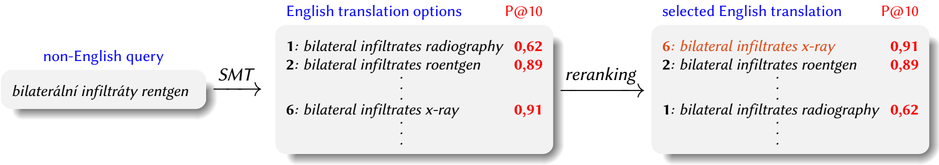
English translation options	P@10
1: *bilateral infiltrates radiography*	**0,62**
2: *bilateral infiltrates roentgen*	**0,89**
⋮	
6: *bilateral infiltrates x-ray*	**0,91**
⋮	

$\xrightarrow{reranking}$

selected English translation	P@10
6: *bilateral infiltrates x-ray*	**0,91**
2: *bilateral infiltrates roentgen*	**0,89**
⋮	
1: *bilateral infiltrates radiography*	**0,62**
⋮	

1. SMT 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



non-English query

bilaterální infiltráty rentgen

$\xrightarrow{SMT}$

English translation options — P@10

1: *bilateral infiltrates radiography* **0,62**
2: *bilateral infiltrates roentgen* **0,89**
⋮
6: *bilateral infiltrates x-ray* **0,91**
⋮

$\xrightarrow{reranking}$

selected English translation — P@10

6: *bilateral infiltrates x-ray* **0,91**
2: *bilateral infiltrates roentgen* **0,89**
⋮
1: *bilateral infiltrates radiography* **0,62**
⋮

1. SMT 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

- SMT model features + the total SMT score
- Retrieval status value
- Inverse document frequency from the collection
- Term frequency in SMT 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 (baseline)	45.61	47.73	42.42
SMT features	**44.70**	**48.79**	**42.73**
All features	**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 (baseline)	45.61	47.73	42.42
SMT features	**44.70**	**48.79**	**42.73**
All features	**50.15**	**51.06**	**45.30**
Google Translate	50.91	49.70	49.39
Bing Translator	47.88	48.64	46.52

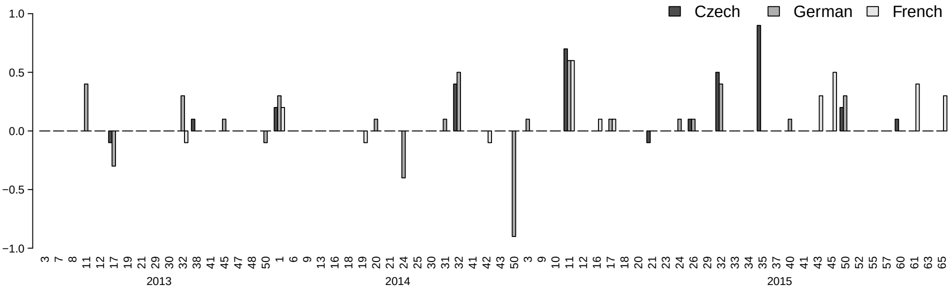
Query translation reranking: Overall results (2016)

P@10 (%) on test queries

system	Czech	French	German
Monolingual	50.30	50.30	50.30
1-best (baseline)	45.61	47.73	42.42
SMT features	**44.70**	**48.79**	**42.73**
All features	**50.15**	**51.06**	**45.30**
Google Translate	50.91	49.70	49.39
Bing Translator	47.88	48.64	46.52

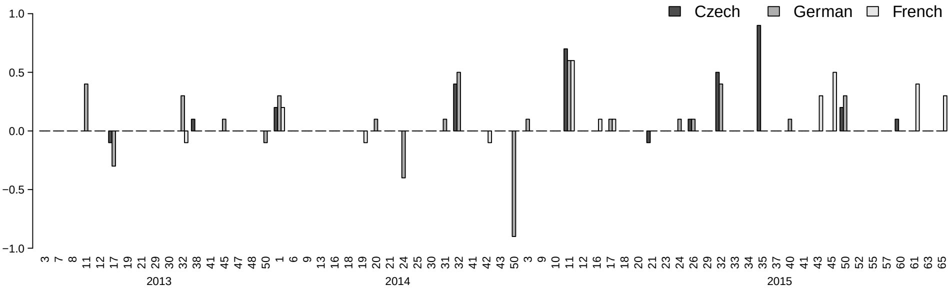
- A single model trained on data for all source languages
- $3 \times 100 \times 20 \sim 4000$ training instances (duplicities removed)

Query translation reranking: Results per query



ΔP@10

Query translation reranking: Results per query



ΔP@10

- ΔP@10 > 0: Czech: 9, French: 14, German: 16
- ΔP@10 < 0: Czech: 2 , French: 3, German: 4

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.

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 (2018)

P@10 (%) on test queries

system	Czech	French	German
Monolingual	53.03	53.03	53.03
1-best (baseline)	47.27	48.03	44.24
1-best+expansion	**52.58**	**49.55**	**47.12**

Query expansion: Overall results (2018)

P@10 (%) on test queries

system	Czech	French	German
Monolingual	53.03	53.03	53.03
1-best (baseline)	47.27	48.03	44.24
1-best+expansion	**52.58**	**49.55**	**47.12**
reranking	49.09	53.64	46.67
reranking+expansion	**53.18**	**50.00**	**46.52**

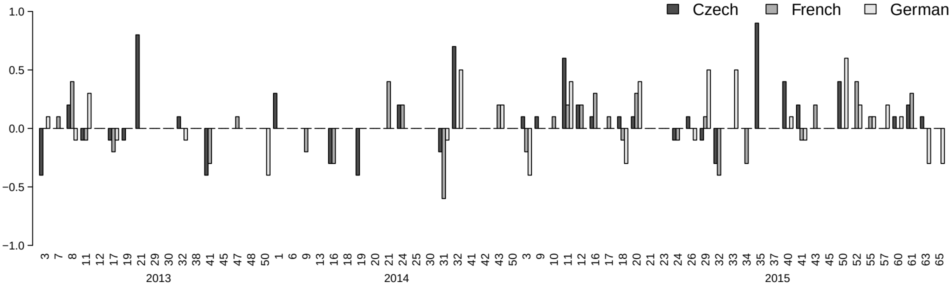
Query expansion: Overall results (2018)

P@10 (%) on test queries

system	Czech	French	German
Monolingual	53.03	53.03	53.03
1-best (baseline)	47.27	48.03	44.24
1-best+expansion	**52.58**	**49.55**	**47.12**
reranking	49.09	53.64	46.67
reranking+expansion	**53.18**	**50.00**	**46.52**

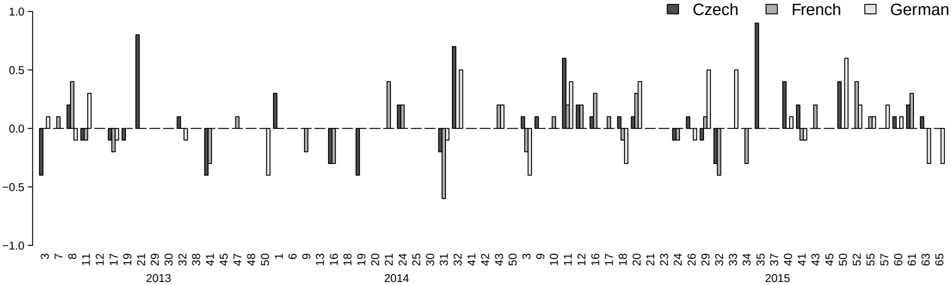
- A single model trained on data for all source languages
- ~ 4000 training instances

Query expansion: Per query results



ΔP@10

Query expansion: Per query results



ΔP@10

- ΔP@10 > 0: Czech: 21, French: 17, German: 14
- ΔP@10 < 0: Czech: 11 , French: 12, German: 11

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*

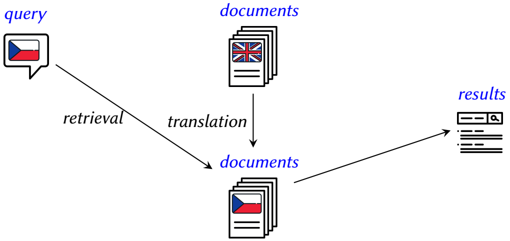
query id: 2015.50.cs		P@10
src:	*červená skvrna obličej kojenec*	
ref:	*red spot baby face*	*0.4*
smt:	*red face infants*	*0.2*
exp:	*+baby +stain +spot*	*0.6*

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



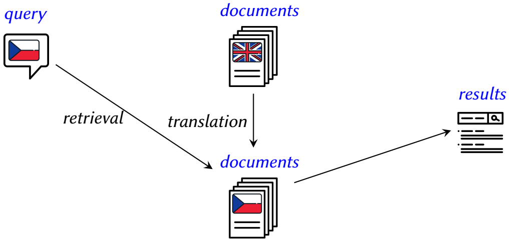
Query translation

- 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

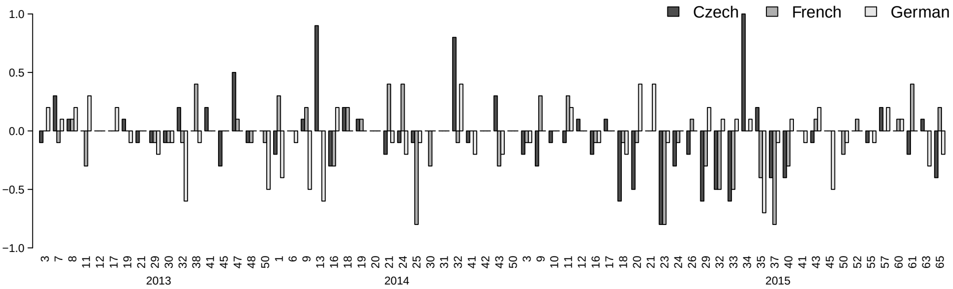
B: Post-lemmatization of output of A

C: Pre-lemmatization of training data of A

P@10 (%) on test queries.

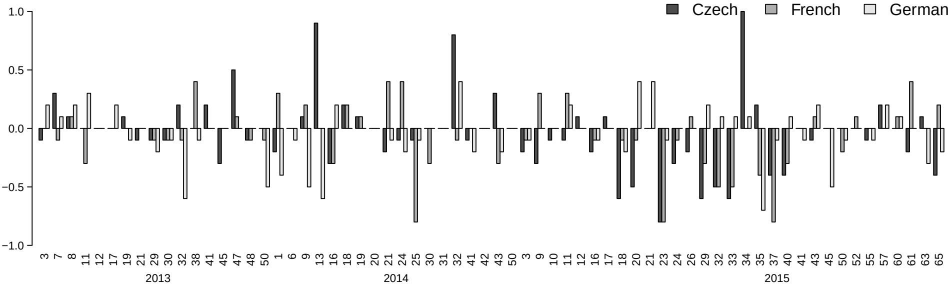
	Czech	French	German
QT-baseline	47.27	48.03	44.24
QT-reranker	48.03	51.67	46.21
DT-form (A)	**38.03**	**42.73**	**37.88**
DT-post-lemma (B)	**40.76**	**41.36**	**38.18**
DT-pre-lemma (C)	**42.88**	**43.18**	**39.85**

Query translation vs. document translation: Results per query



ΔP@10

Query translation vs. document translation: Results per query



ΔP@10

- ΔP@10 > 0: Czech: 18, French: 17, German: 18
- ΔP@10 < 0: Czech: 31 , French: 25, German: 27

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*

Conclusions

Conclusions

Findings

- SMT query translation can be improved by reranking of translation options
- Query translations can be expanded by terms extracted from SMT output and other sources
- Document translation approach is not better than query translation

Conclusions

Findings

- SMT query translation can be improved by reranking of translation options
- Query translations can be expanded by terms extracted from SMT output and other sources
- Document translation approach is not better than query translation

Future/ongoing work

1. Moving to Neural MT (single system for all language pairs)
2. Replace MT by cross-lingual embeddings

NMT query translation experiments

Architecture:

- Transformer + backtranslation of (large) monolingual data

Problem:

- noisy parallel training data
- large portion of parallel training data is extracted from dictonaries
- query language is very specific

Query translation errors examples:

- **src:** *neumonía por aspiración y disfagia faríngea*
  tgt: *Aspiration pneumonia and dysphagia of pharynx (disorder)*
- **src:** *trockene rote und schuppige Füße bei Kindern*
  tgt: *red itchy feet of infant not due to birth, not due to birth, not due to birth, not due to birth, not due to birth, not due to birth, not due to birth, not due to birth, not due to birth*