Multilingual Test Sets for Machine Translation of Search Queries for Cross-Lingual Information Retrieval in the Medical Domain

Zdeňka Urešová, Ondřej Dušek, Jan Hajič, Pavel Pecina {uresova,odusek,hajic,pecina}@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague

- 1,508 real **English** user search queries from the medical domain
- Translated into **Czech**, **French**, and **German**
- For development and testing of machine translation and cross-lingual information retrieval
- **Available for download at:**

http://bit.ly/khresmoi-query-set

- Used in the **WMT 2014 Medical Translation Task** (along with a similar data set of in-domain full sentences)

Domain and Genre Specific Translation

- Domain-specific: terminology, specialized readings of known words
- Genre: search query – short, specific (or no) grammar
- Machine Translation must be adapted to avoid performance loss

Examples

Preserving (non-)syntax

colon cancer *(noun phrase)*	rakovina tlustého střeva cancer du côlon Dickdarmkrebs
pain cancer *(separate words)*	bolest rakovina douleur cancer Schmerz Krebs

Translating abbreviations

EEG, CRP *(keep English abbreviation – international usage)*	EEG, CRP EEG, CRP EEG, CRP
ICU *(Intensive Care Unit – abbreviation translated)*	JIP USI ITS
RTU *(Real-Time Ultra-sonography – full expression used)*	ultrazvukové vyšetření v reálném čase ultrasons en temps réel Echtzeit-Ultraschall

Preserving query structure

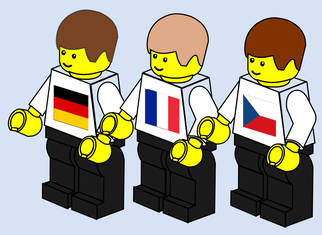
caustic AND stent *(logical operators)*	žíravý AND stent caustique AND stent kaustisch AND stent

Manual Translation of the Test Sets

English query logs

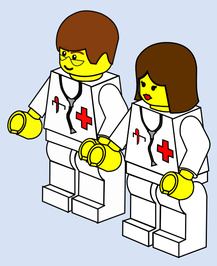
- Real web search queries by healthcare professionals and general public

1) Translation by native speakers



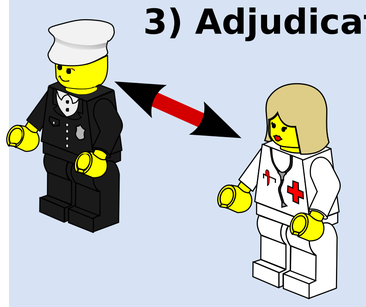
- Including spelling correction and basic filtering
- No specific guidelines (just “translate, do not explain”)

2) Check by medical professionals



- Marking and correcting errors (terminology etc.)
- Spellcheck review
- Much more **specific instructions**, regarding:
  - syntax
  - abbreviations
  - logical operators

3) Adjudication process

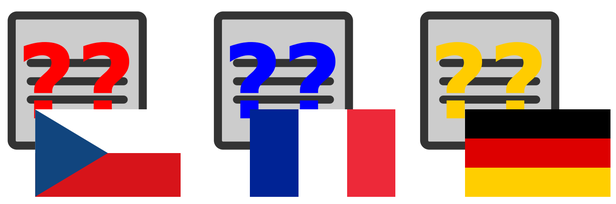


- Solving disagreement between translators and expert reviewers
- Consulting different medical experts

4) Final check

- Independent person
- Taking all opinions so far into consideration

Translated data sets



Khresmoi Project

Automated information extraction from biomedical documents

- Semantic search adapted to user requirements
- Automated analysis and indexing of medical images in 2D (X-Rays), 3D (MRI, CT), and 4D (MRI with a time component)
- Linking information extracted from biomedical texts and images to structured information in knowledge bases
- Support of **cross-language search**, including multilingual queries, and returning **machine-translated pertinent excerpts**
- Adaptive user interfaces to assist in formulating queries and interacting with results

Data Statistics

	devel	test
total queries	508	1,000
public	249	500
professionals	259	500
English words	1,084	2,067
Czech words	1,128	2,121
German words	1,041	1,951
French words	1,335	2,490
average words per query (English)	2.13	2.07

MT Experiment

- Phrase-based machine translation using the **Moses** toolkit
- Trained on general-domain texts:
  10M parallel sentences
  30M monolingual sentences for LM
- Plain tokenized texts, no factors
- **Comparing systems**:
  a) tuned on **general**-domain texts
  b) tuned on Khresmoi **query** devel set

BLEU score

	general	query
Czech-English	26.59±4.42	35.73±5.60
French-English	32.67±5.17	37.84±5.32
German-English	23.03±3.87	29.50±4.92

- Remarkable improvement solely due to tuning model weights (otherwise same training data)
- Problem: high variance

Presented at LREC, May 30th, 2014, Reykjavík, Iceland

http://khresmoi.eu

Supported by the EU FP7 grant n°257528 (KHRESMOI). Resources are distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2010013).