Often the most relevant, up-to-date and clear medical information is not available in low-resourced languages, but is rather

And the other way around:

* more experience with certain types of diseases more common in the regions of the low-resourced languages

In some cases speakers of a low-resourced language may be also familiar with another language for certain historical or political reasons, e.g., many Ukrainian and Belarusian speakers know Russian. However, even in these cases high-quality MT for such low-resourced languages is critical for two reasons:

The motivation for high-quality Neural MT

Availability of high quality Neural MT for medical domain for different under-resourced languages is critical

Even though the speakers of some low-resourced languages may be familiar with another better-resourced language (e.g., Ukrainian and Belarusian speakers know Russian), the development of

The need for a high-quality neural MT for Ukrainian and Belarusian speakers is motivated by two groups of reasons. From the reception perspective, these speakers will gain access to more up-to-date, trustworthy, relevant and clearly explained medical information, which is produced and frequently updated in English, while the information which may be available in their own languages or in another widely-understood language such as Russian may often be not up to the same standard, outdated or unclear. Secondly, from the production perspective, such specialized NMT systems would allow the English-speaking medical community to more easily access information published in Ukrainian or Belarusian about certain diseases that are more wide-spread in these countries, where doctors and scientists have more expertise compared to their colleagues in other contries, but normally publish in their local language. For example, since Belarus and Ukraine have been severely affected by the radioactive fallout from the Chornobyl nuclear accident in 1986, so there are many more cases and publications about the long-term diseases of the endocrine system, but the majority of these articles are never translated into English (e.g., publications in the Ukrainian Medical Journal <https://www.umj.com.ua/>). Availability of a specialized NMT for the medical domain would bridge this gap both for the users and for the professional medical community.

// technical details

* challenge: (a) under-resourced & highly inflected; specialized domain with emphasis on terminology
* addressing this: lemmatization, rule-based framework…
* domain adaptation
* rule-based generation… >> two-staged

Training Neural MT engines normally requires large parallel and monolingual corpora to achieve acceptable translation quality, In addition, NMT into highly-inflected languages could require even larger corpora to compensate for the data sparseness due to the higher type/token ratio. However, for our set of low-resourced languages in the medical domain such large parallel corpora are not available. To address this problem, we will develop the technology for the project that will rely on much smaller available training resources, comparable corpora, morphosyntactic pre- and post-processing of the data, domain adaptation, pivot translation via closely-related languages and the use of related technologies, such as terminology extraction, cognate identification, identification of translation equivalents in comparable corpora, Named Entity recognition (NER), rule-based and hybrid Natural Language Generation (NLG). We will also explore a multimodal path for low-resource NMT, collecting and using medical images in the translation process.

Specifically, (1) in the corpus-creation stage, we will collect a set of larger monolingual comparable corpora in the medical domain with related medical images; we will collect a set of smaller parallel out-of-domain corpora and in-domain corpora, and crowd-source translation of ≈ 35k in-domain controlled parallel sentences, selected to cover most typical lexicogrammatic constructions in the medical domain. (2) In the data preparation stage we will lemmatize and morphologically analyze training corpora, identify medical terms and their translation equivalents using cognate identification, vector space models on non-parallel comparable corpora and rule-based heuristics for terms structure. (3) In the training stage we will create separate in-domain and out-of-domain models from lemmatized corpora, models for terminology translation and generation models for producing correct inflected forms from lemma strings. (4) In the system integration stage we will develop a model combination architecture that will combine our out-of-domain, in-domain, terminology and morphological generation models, rule-based and hybrid heuristics to produce translated text and selecting related medical images. Finally, (5) in the evaluation stage we will carry out automated and human evaluation of translation quality and usability of the system, compared to the baseline systems built with methods for well-resource languages.

// relevant experience

The members of the project – Dr Bogdan Babych and Dr Atul Kr. Ojha have published, participated and lead collaborative research project in the areas of MT for under-resourced languages and comparable corpora (EU FP7 ACCURAT – “Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation” project), hybrid MT (EU FP7 HyghTra “High Quality Hybrid MT” project), terminology extraction and terminological translation equivalents for MT (EU FP7 TTC “Translation Terminology and Corpora” project), Named Entity recognition for Machine Translation, morphological resources for highly-inflected low-resourced languages, human and automated MT evaluation. Please see CVs of the project team for more details.

combination architecture

Innovative aspects ::

we will

explore

hybrid (rule-based, statistical and neural) morphological analysis and

// notes

Also, translation into highly-inflected languages usually require even larger training

and highly inflected

// relevant experience :

Work on comparable corpora ::

has been shown to deliver high qu

The main technical chal

// relevant experience

and publications on with long-term impact in the area of endocrinology, which , e.g., in the Ukrainian-speaking

developed experience in certain specialized areas compared to their colleagues in the English-speaking

Proposal with Ireland:

**Automated extraction of terminological translation equivalents in medical domain for low-resourced languages**

// Medical terminology extraction from comparable corpora for under-resourced languages

We address an important problem in the development of computer-assisted translation technologies for specialized domains, such as medical translation, specifically – the problem of automated extraction of translation equivalents for terminology, i.e., linguistic expressions that refer to clearly defined concepts in the target subject domain. Even though this problem has received much attention in recent years, still there remain serious gaps in coverage of medical terminology even for leading Machine Translation (MT) engines, e.g., Google Translate, Bing MT or DeepL Translator often mistranslate medical terminology. Human translators who work in medical domain also rely on term banks for source and target languages, which are in many cases incomplete or not available. This problem is especially serious for low-resourced languages, where standard methods of parallel terminology extraction cannot be used, which require large parallel corpora, large medical glossaries, wide-coverage morphological taggers and parsers. In the project we will develop a set of technologies which combine methods of discovering translation equivalents in non-parallel, comparable corpora, identification of cognate terms, extraction of multiword expressions and Named Entities, rule-based and hybrid heuristics for identification of terminological boundaries and aligning translation candidates, vector space models for terms in comparable corpora, Natural Language Generation. The project will deliver a methodology for monitoring monolingual corpora in medical domain, such as Medline and populate term banks for a several under-resourced languages in the medical domain.

Methods for extracting terminological translation equivalents have been primarily developed for well-resourced languages, but they do not work well for under-resourced …

There is a need to develop a methodology for discovering equivalents for new terms from comparable corpora…

-- medline…

Heuristics – cognates,

is not leading

translation of terminology remains

discovery of the terminological translation equivalents approaches assume

, especially for well-resourced world languages,

// currently – poorly covered…

The project aims to address an important problem for translation technologies and

translation technologies

the use o translation technology in

in translation technology for specialized translation domains, such as medical translation…

The project aims at developing a technology for automatically discovering terminological translation equivalents for low-resourced languages, such as Hindi

Ex

Plan:

* terminology as the most critical aspect of translation technology for domain
* parallel resources are not available
* low-resourced languages
* motivation: