Multimodal Neural MT in Medical for Less-resourced Languages

----

1. Motivation

The need for a high-quality neural MT for Indian, Ukrainian and Belarusian speakers is motivated by two groups of reasons. From the reception perspective, these speakers will gain

a. access to more up-to-date, credible, relevant and clearly explained medical information, which is produced and frequently updated in English as compared to other languages, both generally and academically,

b. facilitation of communication between a physician and a patient leading to an efficient diagnosis and satisfactory prognosis,

c. ease in carrying out legal disputes and documentation.

From the production perspective, a specialized NMT system will also allow the English-speaking medical community to more easily access information published in Indian or Ukrainian or Belarusian languages about certain diseases that are more widespread in these countries, where doctors and scientists have more expertise compared to their colleagues in other countries, but normally publish in their local language. For example, since Belarus and Ukraine have been severely affected by the radioactive fallout from the Chernobyl nuclear accident in 1986, there are many cases and publications about chronic diseases afflicting 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/).

In India, English is the *de facto* mode of communication for health professionals, academicians, and legal personnel, but understood by only 9% (approx.) of the population. This makes it difficult for the non-English speakers to benefit from public medical information, such as a description of symptoms, prophylactics, availability of professional help, the results of recent research and developments in medical treatment and diagnosis of specific conditions of diseases. As per the National Health Policy (NHP) 2018 report[1], Andhra Pradesh (Telugu), Tamil Nadu (Tamil), and northern states of India (Hindi and Bhojpuri) suffer from several diseases and death ratio is higher than other states. One of the main reasons is the lack of medical awareness and unavailability of medical information in their respective languages. Therefore, the availability of a specialized, multimodal NMT for the medical domain would bridge this gap both for the users and for the professional medical community.

2. Scientific & Technical Details

Machine Translation (MT) systems have attained acceptable translation quality for many tasks, especially in well-resourced subject domains and for languages where large collections of human translations and wide-coverage language processing resources are available. A success in several domains, the application of automated translation in the health/medical domain is still limited, in particular – due to the lack of large-scale resources for medical terminology and Named Entities, especially for under-resourced languages, and due to gaps in state-of-the-art methods and tools for their systematic development and integration into MT.

Even high-quality MT services and systems, such as Google Translate, suffer from problems of reduced quality, hence usage, for less-resourced resourced languages and in specialized domains (such as medical patient-oriented information) due to missing or improper lexical mapping.

[1] https://cdn.downtoearth.org.in/pdf/NHP-2018.pdf

English-Hindi

English as source sentence- *Apart from this, the organisms like e-coli and pseudomonas have also been found because of whose infection the ill effect on the body is maximum.*

Google MT output - *इसके अलावा ई-कोलाई और स्यूडोमोनस जैसे जीव भी पाए जाते हैं, जिनके संक्रमण के कारण शरीर पर बीमार प्रभाव अधिकतम होता है।*

Bing MT output - *इसके अलावा ई जैसे जीवों-कोलाई और* pseudomonas *भी है क्योंकि जिसका संक्रमण शरीर पर बीमार प्रभाव की वजह से पाया गया है अधिकतम है ।*

REF - *इसके अलावा ई-कोली और सिडोमोनॉस जैसे जीवाणु भी मिले , जिनके संक्रमण से स्वास्थ्य पर सबसे ज्यादा कुप्रभाव पड़ता है ।*

The error illustrated above is the inadequate term identification leading to segmentation issues, e.g., "pseudomonas (*स्यूडोमोनस*)", which can be corrected using our proposed approach which includes Medical Entity recognition in the first stage and then a separate translation strategy for the identified units.

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.

3. Prior relevant work

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.

Dr Atul Kr Ojha has also worked with Akanksha Bansal for a project titled Indian Languages Corpora Initiative (ILCI), started in 2009 and completed in 2017. This project aimed at creating a labeled text corpus (parallel and monolingual) in 17 Indian languages in four domains - Health, Tourism, Agriculture, and Entertainment.

Please see CVs of the project team for more details.

4. Work Plan

From July 2019 to June 2020 (University of Leeds and Panlingua Language Processing LLP)		
Year	**Months**	**Targeted Outcomes**
July-December 2019	**0-3**	● Setting up the Project (including hiring and training to manpower) ● Creation of 10K parallel sentences and 2K parallel images ● Creation of terminological and lexicon database
	4-6	● Creation of 20K parallel sentences and 5K parallel images ● Update terminological and lexicon database
January-June 2020	**7-9**	● Creation of 5K parallel sentences and 3K parallel images ● Update terminological and lexicon database ● Development of baseline IS-LRL Neural-based Multilingual Multimodal MT system for Medical Text
	9-12	● Corpus and Lexicon Validation ● Evaluation of MT system ● Improving the efficiency and accuracy of the MT system ● Final evaluation and hosting ● Project Submission

5. Expected project outcomes

The deliverables are:

1. 35,000 parallel sentences for Indian and Slavonic less resourced languages (IS-LRL) for English-Belarusian, Bhojpuri, Hindi, Tamil, Telugu and Ukrainian in the medical text
2. 15,000 parallel images for the IS-LRL in the medical text
3. Development of the multilingual Educational Multimodal Medical Machine Translation System for IS-LRL.

Additional deliverable is :

4. Lexicons, and terminological database of IS-LRL in the medical text.