Machine Translation in Medicine

A quality analysis of statistical machine translation in the medical domain

Marta R. Costa-jussà
TALP Research Center
Universitat Politècnica de Catalunya
Barcelona, Spain
mruiz@gps.tsc.upc.edu

Mireia Farrús
N-RAS Research Center
Universitat Pompeu Fabra
Barcelona, Spain
mireia.farrus@upf.edu

Jordi Serrano Pons
UniversalDoctor Project
Barcelona, Spain
jserranopons@universaldoctor.com

Abstract—**Machine translation is evolving quite rapidly in terms of quality. Nowadays, we have several machine translation systems available in the web, which provide reasonable translations. However, these systems are not perfect, and their quality may decrease in some specific domains. This paper presents an analysis and evaluation of the quality of a state-of-the-art machine translation system in the medical domain. Evaluation is performed automatically and manually for seven language pairs. Results show that for most of them, machine translation is not ready to be applied in a domain like medicine where a translation with 100% adequacy is required.**

Keywords- medical communication tools; patient-doctor communication; medical translation; statistical machine translation; machine translation evaluation

I. INTRODUCTION

Recently, machine translation has become quite popular, especially because people are becoming used to the text translation systems freely available online. In addition to this, the scientific community is really involved in machine translation. We have just to point out the organization of major conferences and events and the great efforts that outstanding research groups and companies dedicate to its improvement.

One of the great advantages of the machine translation application is that most of the users do not require perfect translations. Users may be only interested in roughly understanding a text in order to simply get an idea of what the text is about. However, other users may not be that flexible. The correctness and the beauty of the text in Law and Medicine might not be important, but the precision and adequacy in the translated message can be crucial for the objective of the communication. In medical communication, a translation error between the patient and the doctor, or an error in the communication of a treatment or a diagnosis may lead to serious consequences in people's health.

Given the importance of a precise communication in Medicine, medical translations are currently realized by human translators (or *mediators*) or technological applications, which give support to mediators by providing a multimedia technology, which can back up between the communication healthcare providers and the patient in a standard medical visit.

Recently, due to the growing success and interest in new language technologies, machine translation has been brought to the Medicine field, but normally subject to a posterior post-edition process, i.e. the correction of a text after being translated by a machine translator. A study performed by [1], for instance, analyzed the feasibility of post-editing machine translated health-promotion English documents from local and national public-health websites in the USA. It was assumed, a priori, that machine translation would not provide enough quality for the documents to be used as official versions.

In spite of that, language technologies are constantly increasing their quality of performance, and we should expect that in a mid/long term, machine translation will be capable of translating any text in any domain with the required quality.

The current study focuses on analyzing a freely available, popular and state-of-the-art translation system in the web. We want to explore if this kind of system could be used by a hospital, doctor and patient, without extra economical or structure resources than Internet access. We experiment if this type of system could be used in cooperation or in substitution of human translators. For this purpose, the Google Translate application has been chosen as a statistical machine translator, which has been proven as state-of-the-art MT systems in relevant international campaigns such as NIST (http://www.itl.nist.gov/iad/mig/tests/mt) or WMT (http://www.statmt.org/wmt12), and it offers machine translation among a high number of language pairs.

The evaluation corpus has been obtained by one of the tools developed by UniversalDoctor project. This material has given us the opportunity of working with medical and real English questions and answers containing the terminology used in a medical consultation. Moreover, the corresponding correct translations into several target languages done by professional translators are provided, in order to use them as reference sentences. The performance quality is given through two different kinds of evaluation.

The remaining of this paper is organized as follows. Section II provides a brief introduction in medical translation, together with the description of the UniversalDoctor's tool. Section III reports the methods used and section IV introduces the

experimental framework. Section V presents the evaluation results, and, finally, section VI concludes.

II. BACKGROUND

In this section we report the related work in medical translation together with the description of the UniversalDoctor tool.

A. *Medical translation*

Immigration between people from different countries takes place for diverse reasons; but in general terms, it can be either voluntary or forced. Whatever the reason is, learning a new culture and language is an effort that becomes still more difficult in a forced immigration. In the meanwhile, many immigrants find themselves in severe difficulties when trying to communicate with local inhabitants, a fact that gets worse in medical situations. It is at this time when a medical translation plan to help communication between patients and doctors becomes of great importance [2].

Medical translation refers to the translation of several kinds of documents: technical, regulatory, clinical, marketing, etc., as well as software or training curriculum for the pharmaceutical, medical device or healthcare fields. As in other areas, literature and labeling of medical products and devices, and documents to conduct clinical trials that have to be read by both clinicians and patients, are usually required to be presented in the local language. However, despite all these requirements, medical translation needs to be done by professional translators with specific matter knowledge, able to deal with the sensitive technical and regulated nature of medical documentation and to guarantee a high translation quality.

In the 1980s, when machine translation become more popular and computational power increased, some few initiatives regarding machine translators were carried out in order to ease communication between health professionals and patients who do not share the same language or culture and are brought to interact. TRANSOFT [3], for example, was a public-domain medical document translator, in which translations were obtained by implicit word parity across languages, disambiguation based on context in the source language, and word rearrangement from source into target language word order. However, a report presented in [4] analyzed the feasibility of German-to-English TRANSOFT in a medical domain. Among other issues, it concluded that not all the potentially ambiguous terms were resolvable from the immediate context. Thus, although machine translation systems may provide a rapid and inexpensive means of obtaining draft translation, a further post-edition is required. As stated in [1], post-edition is feasible and better than human translation from scratch. However, because of the sensitive of the topic we are dealing with, the main objective of medical translation should focus on providing translation outputs free of any possible error and ambiguity.

B. *Universal Doctor tool*

In particular, the UniversalDoctor project was born from the need for initiatives that facilitate intercultural communication in our increasingly global world, so that access to health care could be universal. More explicitly speaking, the UniversalDoctor project has as main objective facilitate communication between health professionals and patients who have difficulty in communicating in the language used by the health professionals themselves.

Within the framework of UniversalDoctor project, the *UniversalDoctor Speaker ®:Family Medicine in 9 languages* is a computer application that allows primary care physicians to carry out a diagnostic patient interview and prescribe treatment in nine different languages. Any primary care physician will be able to formulate hundreds of questions in English, French, German, Portuguese, Russian, Romanian, Moroccan Arabic, Mandarin Chinese and Urdu.

III. METHODS: STATISTICAL MACHINE TRANSLATION

Machine translation refers to the automatic translation of text or speech in one specific source language into another target language. The main goal of statistical machine translation (SMT) is to translate a given string in the source text into a string in the target language. Among all possible target strings, the system chooses the string with the highest probability. Whereas these probabilities can be estimated by thinking about how each individual word is translated, modern SMT is based in the intuition that a better way to compute these probabilities is by considering the behavior of phrases (sequences of words). In addition to the translation model, SMT systems use a language model, which is usually formulated as a probability distribution over strings that attempts to reflect how likely a string occurs inside a language [5].

Building an SMT system requires written and high computational resources. The written resources needed are parallel corpora between a source and a target language at the sentence level. Generally speaking, less than 20,000 sentences produce unreadable translations even in closed domains. Given these parallel corpora, it is quite easy to obtain a translation system, since some applications (such as Moses [6]) integrate all necessary tools to build an SMT and are available as open source. SMT has been proved to be one of the best options in machine translation showing a greater robustness than other methods for the translation of spontaneous speech. However, SMT quality depends largely on the language pair of the specific semantic domain being translated, and it still have several significant challenges to pursue.

IV. MATERIALS

The current section reports the details of the experiments and the evaluations performed. First, the data provided by UniversalDoctor project is described. Second, some details on the SMT system being used are reported. Finally, automatic and human evaluations over the translated language pairs are presented.

TABLE I. CORPUS STATISTICS OF THE MEDICAL TEST SET

Language	Statistics			
	Words	*Vocabulary*	*Maximum sentence size*	*Average sent. size*
French	5091	1210	21	5.01
Portuguese	4372	1153	17	4.31
Spanish	5113	1113	17	5.04
English	5128	1053	18	5.06
German	4657	1205	18	4.59
Russian	4458	1386	19	4.04
Basque	4147	1232	15	4.09

A. *Data*

Experiments are performed over seven language pairs using a test corpus provided by the UniversalDoctor project. It is a medical corpus consisting of one thousand source sentences and their corresponding manual references, and the size of the test set meets the standards of international machine translation evaluations such as IWSLT[1], where the test set contains around 500 sentences and 3.9k words.

The directions of translations were always from English as source language, being the target languages: Basque, French, German, Portuguese, Russian and Spanish, grouped in four different linguistic families: Romance (French, Portuguese and Spanish), Germanic (German, English), Slavic (Russian) and a language Isolate (Basque). Table 1shows the statistics of the corpus: number of words, vocabulary, maximum sentence size and averaged sentence size for each language.

B. *Machine Translation system: Google Translate*

The Google Translate application is the machine translation system used in the current work, since it has been proven state-of-the-art machine translation system in relevant international evaluation campaigns (e.g. NIST, WMT) and it is freely available in the web.

Google Research Group has developed its statistical translation system for a large quantity of language pairs. For experiments, we use the Google API Translation for all pairs of languages.

V. EVALUATION RESULTS

In order to evaluate the translations of the seven language pairs involved in the current experiments, two kinds of evaluations are used: automatic and human (or manual). These are the most traditional measures when evaluating a machine translation system. They are both described in the following sections, together with the obtained evaluation results.

[1] http://iwslt2010.fbk.eu/

TABLE II. BLEU RESULTS USING GOOGLE TRANSLATE (FROM ENGLISH)

Language (from English)	BLEU (%)
	Google Translate
French	24.30
Portuguese	19.51
Spanish	26.34
German	16.61
Russian	12.90
Basque	7.16

A. *Automatic evaluation*

Automatic evaluation is performed by using BLEU, one of the most popular measures in the field of machine translation with the corpus detokenized and case sensitive [7]. BLEU, acronym for BiLingual Evaluation Understudy, consists essentially of an N-gram corpus-level measure and it is always referred to as a given N-gram order (BLEU*n*, being *n* usually 4).

Table II shows the translation quality results in terms of BLEU for Google Translate, when translating from English into any other language. It can be seen that when translating from English into the Romance languages, a better translation is provided than when translating into any other language family (Germanic, Slavic or Isolated). The worst translation is obtained when translating from English to an isolated language (Basque). These results could be expected a priori, given the complexity of the Basque language with respect to other language families, as shown and explained in previous translation works such as [5].

B. *Human evaluation*

The human evaluation was performed by a single evaluator for each language direction of translation. In all cases, the evaluators were required to be fluent in English, native in the non English language and professional doctor translators.

The human evaluation performed in the current work grades translations as *acceptable* versus *non acceptable*. The following criterion was applied: a sentence provides an acceptable translation only if *the source message is preserved and any possible misunderstanding exists*. This human evaluation accepts only sentences with a 100% of correctness in terms of preserving the message semantically. This limited evaluation is of interest in the medicine field, given that it would be the criterion followed by some hospital to introduce an MT system in their routines.

The evaluation criterion was transmitted to human evaluators, and then they were presented a framework with two sentences at a time. One sentence was the original English sentence and the other was the Google translation. Then, they had two options: *accepting* the translation or *not accepting.*

TABLE III. HUMAN EVALUATION RESULTS

Language (from English)	Perfect understandable outputs (%)
	Google Translate
French	86.8
Portuguese	85.0
Spanish	84.8
German	78.1
Russian	47.8
Basque	27.2

The results obtained in the human evaluation are shown in Table III. As it can be seen, the medical machine translation quality depends highly on the pair of languages being translated. Translating from English into any Romance or Germanic language provides very good results, reaching in most cases an 80% of acceptance in the human evaluation. However, the performance changes totally when translating from English into a Slavic language (Russian) or a language isolate (Basque).

Most of the common problems in SMT include unknown words, incorrect word order, word disagreement when translating into Romance languages, and wrong declination when translating into Basque (a language isolate). A state-of-the-art SMT in the medical domain does still not provide quality translations in terms of communication. Machine translations may lead to misinterpretations and misdiagnoses. Therefore, we may conclude that, in the medical domain, other means of communication such as validated applications or professional mediators are still required.

VI. CONCLUSIONS

This paper reported an analysis of the state-of-the-art machine translation in the particular domain of medicine.

This work reports experiments on a popular machine translation application such as Google translator in order to evaluate if medical texts can be translated using these tools. However, an automatic and manual evaluation reported on 7 pairs of languages concludes that the performance of these systems is still not good enough in such a domain where 100% of accuracy is required.

Machine translation systems may be used as a complementary tool, but post-edition or human revision is required in order to guarantee communication among doctor and patient.

ACKNOWLEDGMENT

This work has been partially funded by the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-29951).

REFERENCES

[1] K. Kirchhoo, A. M. Turner, A. Axelrod, F. Saavedra, “Application of statistical machine translation to public health information: a feasibility study,” J. Am. Med. Inform. Assoc., vol. 18, pp. 473–478, 2011.

[2] K. D. Mandl, I. S. Kohane, “Electronic patient-physician communication: problems and promise,” Annals of Internal Medicine, vol. 129, issue 6, pp. 495—500, 1998.

[3] G. W. Moore, I. Wakai, Y. Satomura, W. Giere, “TRANSOFT: Medical translation expert system,” Artificial Intelligence in Medicine, vol. 1, issue 4, pp. 149--157.

[4] G. W. Moore, U. N. Riede, R. A. Polacsek, R. E. Miller, G. M. Hutchnis, “Automated translation of German to English medical text,” The American Journal of Medicine, vol. 81, issue 1, pp. 103—111, 1986.

[5] S. F. Chen, J. T. Goodman, “An empirical study of smoothing techniques for language modeling,” Technical report 1998. Harvard University.

[6] P. Kohen, H. Hoang, A. Birch, et al., “MOSES: Open source toolkit for statistical machine translation,” Proceed. of the 45th Annual Meeting of the ACL. Prague, Czech Republic, 2007.

[7] K. Papineni, S. Roukos, T. Ward, et al., “BLEU: A method for automatic evaluation of machine translation,” Proceed. of the 40th Annual Meeting of the ACL. Philadelphia, PA, 2002.