MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-73

Can Multilingual Machine Translation Help Make Medical Record Content More Comprehensible to Patients?

Qing Zeng-Treitler[a], Hyeoneui Kim[b], Graciela Rosemblat[c], Alla Keselman[c]

[a] *Department of Biomedical Informatics, University of Utah, Salt Lake City, UT*
[b] *Division of Biomedical Informatics, Department of Medicine, UC San Diego, San Diego, CA*
[c] *Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD*

Abstract

With the development of electronic personal health records, more patients are gaining access to their own medical records. However, comprehension of medical record content remains difficult for many patients. Because each record is unique, it is also prohibitively costly to employ human translators to solve this problem. In this study, we investigated whether multilingual machine translation could help make medical record content more comprehensible to patients who lack proficiency in the language of the records. We used a popular general-purpose machine translation tool called Babel Fish to translate 213 medical record sentences from English into Spanish, Chinese, Russian and Korean. We evaluated the comprehensibility and accuracy of the translation. The text characteristics of the incorrectly translated sentences were also analyzed. In each language, the majority of the translations were incomprehensible (76% to 92%) and/or incorrect (77% to 89%). The main causes of the translation are vocabulary difficulty and syntactical complexity. A general-purpose machine translation tool like the Babel Fish is not adequate for the translation of medical records; however, a machine translation tool can potentially be improved significantly, if it is trained to target certain narrow domains in medicine.

Keywords:

Comprehension, Translating, Consumer health information, Medical records.

Introduction

Providing patients access to their own health records is key to the new patient-centric health care paradigm, in which patients take charge of their own health by becoming active participants in their health care decisions [1]. In line with this new paradigm, in many countries, including the US, patients have the legal rights to access their own medical records. The recent development of electronic personal health record (PHR) systems holds the promise to significantly improve patients' access to their own records.

While PHR is helping more patients gain access to their records, comprehension remains an obstacle if PHR is to fulfill its full potential to motivate and empower patients to better their health [2]. Medical records typically contain complex information intended for healthcare professionals, not consumers. The challenge posed by the comprehension of health-related material is even greater for patients who are not fluent in the language in which the records are written. According to a 2004 US Census Bureau report, there are 33.5 million foreign-born residents in the US, representing 11.7% of the total population. Undoubtedly, while many foreign-born residents may well be fluent in English, many others cannot speak, read or write in English, or may never reach an acceptable level of proficiency in that language.

With respect to PHR applications in the US, foreign-born patients' lack of English proficiency may well place them in the "hard-to-reach" category, similar to low-literacy English-speaking patients. It has been argued that we particularly need to reach out to such patient population, since low literacy has been associated with poor self-care and poor health outcomes, including increased mortality and hospitalization [3, 4] and, ultimately, increased social and health-care costs. If consumer applications such as PHR do not address the needs of the "hard-to-reach" populations, we run the risk of widening the existing health disparity gap.

One way to minimize this gap for the non-English-speaking segment and to allow them to benefit from the PHR availability is by providing translations into the patients' native languages. Many US hospitals provide on-site translation services for patients (timely point-of-care services). Further, an increasing amount of patient education materials are available in languages such as Spanish. A number of studies have been conducted on the needs, methods, barriers, and benefits to providing human multilingual translation services [5, 6]. However, within the context of medical records, resorting to human translators is not feasible, since each record is unique. Thus, anticipating the need for record-specific, accurate translations, machine translation emerges as the next best option. We believe machine translation is worth exploring in this context. The goal of this pilot study was to test the efficacy of a popular general-purpose freely-available machine translation tools on patient medical records. English was the source language and the target languages were Spanish, Russian, Korean, and Chinese. Even though this study was meant as a proof-of-concept endeavor, our general goal is the development of machine translation tools

specifically devised to render accurate and understandable medical record information.

Background

Machine Translation can be defined as an automated system that takes a given text as input (source language) and produces as output the translation of that text into a target language. The translation process as such is performed using special computational programs, dictionaries, vocabularies, glossaries, and different sets of linguistic rules [7].

Although the concept of machine translation (MT) has been around since the 30's and 40's, it gained popularity only in the 60's and 70's, when it was touted as the perfect solution for text translation, capable of rendering translated text of human translation quality [8]. However, MT lost much of its appeal when it became evident that those were unrealistic expectations. With the expansion and application of natural language processing (NLP) techniques and stochastic methods to MT in the mid to late 80's, there was a renewed interest given that these techniques produced superior results. Further, despite its potential for numerous errors and disjointed narrative, people became aware that raw, unedited MT output could be used to obtain the gist of a document. More recently, the widespread use of MT in the Internet opened the door to other novel uses and applications, such as cross-language information retrieval [9] and multilingual communication [10].

While it is still true that post-editing is unavoidable for first-rate translation quality, the use of controlled language and vocabularies can go a long way towards reducing the need for intense revisions, especially when restricted to specific domains or to particular types of documents, as in the case of meteorological reports. This is also the setting for electronic health records: they are contextually restricted to the medical field and the narrative can be quite formulaic. An MT system specifically tailored to medical texts could potentially be used on English electronic health records to make them understandable to speakers of other languages.

One such system, engineered by the Pan American Health Organization [11], has several micro-glossaries with domain-specific vocabularies, two of which are geared to the medical field: *Super Medical* and *Patient Information*, with consumer health vocabulary. The fact that the Pan American Health Organization Machine Translation System (PAHOMTS®) is limited to three languages (English/Spanish/Portuguese), however, made it unsuitable for our project, as we were also evaluating the application of MT to other languages not included in this system, namely, Russian, Chinese, and Korean.

Another system, the open source medical speech translation system (MedSLT), has been described in the literature. The prototype of MedSLT translates spoken questions from English into French, Japanese and Finnish in three medical subdomains (headache, chest pain and abdominal pain) using a vocabulary of about 250-400 words per sub-domain [12]. Since the goal of this project was not to evaluate how well or how appropriately different MT systems perform, but rather, whether the use of unedited MT output represents a viable option for translating electronic health records, it was important that the same software be used for all languages involved. This consideration, along with the wide availability offered by the Internet translation tools, pointed to Altavista for this experiment.

Methods

Materials and Reviewers

We retrieved 11 publicly available sample medical records from two websites as testing materials:

- MedLEE demo site (http://zellig.cpmc.columbia.edu/medlee/demo/)
- MT (Medical Transcript) resources (http://www.mt-resources.com/index.html)

They include discharge summaries, surgical notes, admission notes, and radiology reports. For a translation tool, we selected the freely available Babel Fish by AltaVista (http://babelfish.altavista.com/), as it is one of the most easily accessible and widely known online translating tools.

A total of five reviewers participated in the study. All reviewers were proficient in English and a native speaker of either Chinese, Korean, Spanish, or Russian. All were medical informatics researchers with graduate school level of education or higher.

Procedure

Identifying the testing variables

In order to identify operational variables that measure the quality of the translation, we first translated one record using Babel Fish from English into each of the four different languages mentioned in the introduction. In order to review the quality of the translated text, we used two testing variables which had also been identified as important evaluation criteria for MT by other studies [13, 14]: understandability and correctness of the translated sentence. Regardless of the language in which they are written, medical records are difficult to understand because they are fairly technical documents that require a certain level of expertise to understand. Moreover, they often contain abbreviations and grammatically incorrect phrases and expressions. In order to control for such intrinsic confounding factors we added a third variable: understandability of the original sentence.

Establishing the evaluation rules

We divided each text into its component sentences, to assess the translation quality of each sentence via a 3-point Likert scales. We considered each sentence to be a self-contained chunk of information. In order to promote consistency in using the scale and set uniform parameters, an initial detailed instruction sheet with examples were developed and distributed to the reviewers. First, four reviewers coded 39 sentences collected from 2 records. The coding results were shared and discussed in a group meeting, and the rules and instructions were further refined. For example, when the translator failed to translate certain terms the reviewers were instructed to replace the untranslated terms with blank spaces

and determine the understandability and correctness as if the untranslated terms were missing.

Coding the translation

The specific coding steps were similar to those described in (13). They were as follows:

1) Rate the understandability of the translated sentence first without seeing the original sentence in the source language (English) and without consulting any dictionary.
2) Rate the understandability of the original sentence in English without consulting any dictionary.
3) Rate the correctness of the translated sentence in terms of accuracy, by comparing the translated sentence to the original sentence. Dictionaries may be consulted in this step.

Reviewers then rated 213 sentences collected from 8 records, following the finalized evaluation rules and instructions.

Reliability of the translation evaluation

In order to test the reliability of the human evaluation results, we trained a second native Spanish speaker (the fifth reviewer) in the use of the same evaluation rules just described, and then assessed the level of agreement between the two reviewers. After practicing with the evaluation rules on 15 sentences, the second Spanish reviewer independently rated the quality of the Spanish translation of 65 sentences which were randomly selected from the 213 sentences. We observed the percentage of agreement in each variable. In addition, the level of agreement in each variable was tested with a McNemar test. In the McNemar test, the categories of "partial" and "no" were merged as one.

Analysis of the sentence characteristics

In order to identify factors that affect the accuracy and comprehensibility of the translation, we investigated the characteristics of the original sentences. We extracted three kinds of text features using an in-house built Natural Language Processing (NLP) tool: HITEx (Health Information Text Extraction) [15]. We also measured the overall readability of each sentence using the readability analysis tool called HIReA (Health Information Readability Analyzer) which assesses the text readability based on three types of text characteristics [16].

Results

Translation quality

Reviewers found the translation to be understandable only between 11.27% and 31.46% of the time (Table 1). In other words, for each language tested, reviewers found that the vast majority of translations were either incomprehensible or partially comprehensible. In contrast, the majority (65.73% to 85.73%) of the original English sentences were deemed comprehensible by all reviewers.

When examining the correctness of the translations, we found that only a small percentage (7.98% to 11.74%) of the Chinese, Russian, and Korean translations were deemed correct by the coders. Spanish translations did comparatively better, with 33.80% deemed correct. Nevertheless, for all languages involved, the majority of the translations we not totally correct. Due to the nature of medical records, which contain critical information, the lack of accuracy in translations is very problematic.

Table 1 - Understandability of the original and translated sentences and correctness of the translated sentences

		Spanish	Chinese	Russian	Korean
Translation understand-able?	Yes	31.46%	11.27%	14.55%	19.25%
	Part.	43.19%	26.29%	20.19%	28.64%
	No	25.35%	62.44%	65.26%	52.11%
Original Sent. Understand-able?	Yes	88.73%	66.67%	80.28%	65.73%
	Part.	10.33%	25.35%	13.62%	26.76%
	No	0.94%	7.98%	6.10%	7.51%
Translation Correct?	Yes	33.80%	7.98%	11.74%	9.39%
	Part.	44.60%	7.51%	10.33%	24.88%
	No	21.60%	84.51%	77.93%	65.73%

Reliability of the translation evaluation

When comparing the results of the two Spanish coders on the 65 test sentences, we found that they agreed with each other 71.31% to 80.00% of the time on the three parameters. This agreement rate is acceptable considering that the two reviewers spoke somewhat different Spanish dialects, as they came from two different Spanish-speaking countries. When applying the McNemar's test, statistically significant differences ($p<=0.01$) were found in their judgment of comprehensibility and correctness of each sentence, but not in the comprehensibility of the original sentences.

Analysis of the sentence characteristics

The correlation analysis (Table 2) shows that the sentence length is a significant feature that negatively affects the understandability of the original sentence and the translation quality. In other words, longer sentences were less likely to be understood or yield a correct translation. Both vocabulary features (vocabulary familiarity score and out of dictionary word ratio) were significantly correlated with the understandability and correctness variables. The use of familiar terms showed positive correlations whereas the use of out-of dictionary terms showed negative correlations. No part-of-speech categories consistently and significantly correlated with understandability and correctness. The readability score (how difficult a sentence is) was positively correlated with understandability and correctness.

Table 2 – Correlation analysis: sentence characteristics and translation quality

	Sentence length	Vocab. familiarity score	Out of dictionary word ratio	Readability score
Understand-ability of original sentence	-0.2722*	0.2949*	-0.4267*	0.3554*
Correctness of translation	-0.4393*	0.1201†	-0.2502*	0.2664*
Understand-ability of translated sentence	-0.3625*	0.1996*	-0.2702*	0.2588*

* *correlation coefficients are significant at 95% significance level,*

† *correlation coefficient is significant at 90% significance level.*

The mean values of text features that showed significant correlations with translation quality and the understandability of the original sentences are presented in Table 3. "Yes" answers received a weight of 2, "no" answers a weight of 0, and "partial" a weight of 1. The average original and translation understandability and translation correctness were categorized into two groups: incomprehensible or incorrect (=<1) and comprehensible or correct (>1).

Table 3 - Mean values of text features that had significant correlations with translation quality and vocabulary familiarity score

		Number of words per sentence	Vocabulary familiarity score	Out of dictionary word ratio	Readability score
Original	Incompr.	17.6920	0.6784	0.2427	-0.6450
	Compreh.	12.9700	0.6737	0.0904	-0.4740
Transl.	Incorrect	14.4670	0.6774	0.1024	-0.5010
	Correct	8.6136	0.6612	0.0894	-0.4180
Transl.	Incompr.	15.2540	0.6684	0.1087	-0.5260
	Compreh.	9.2670	0.6854	0.0817	-0.4010

As suggested by the text feature analysis, the main cause of incomprehensible and incorrect translations appears to be the technical domain-related medical vocabulary on one hand, and irregular or complex syntax used by the original English sentences on the other. Longer sentences tend to have more complex syntax and a higher chance of containing difficult words. To a lesser extent, the vocabulary and syntax also made the original English sentences fully incomprehensible or partially comprehensible at times.

Upon closer examination, ambiguous and out-of-dictionary terms are the two main vocabulary problems for MT. Take the example of the sentence "Patient has had small to moderate amounts of serous drainage at site". The word "moderate" has several dictionary definitions and different parts of speech. In the original sentence, it is used as an adjective with the meaning "of medium or average quantity or extent." However, it could also be used as a verb to mean "to lessen the violence, severity, or extremeness of," which apparently was how the word was wrongly interpreted by the MT system when converting the sentence to Chinese, Korean, and Spanish. The system that we used is a general purpose, general-vocabulary application. Thus, some medical terms were missing from the system's dictionary and were not translated. This rendered some sentences not understandable. We rated all incomprehensible translations as incorrect.

Discussion

Although there is an increasing need for automated translation of medical content, especially in the context of personal health record applications, there has been no reported study on the efficacy of applying existing MT technologies to patient medical records. Our pilot study evaluated the quality of a popular general purpose translator by applying it to a novel use, that is, to the translation of 213 sentences from patient medical reports, from English into four different languages.

This study found that the translation results are quite frequently incomprehensible and inaccurate, for all four languages tested. While the original English sentences were not easy to read, the majority of them were deemed totally comprehensible by each of the coders. However, the reverse was true for the comprehensibility of the translations. In addition, in terms of accuracy well below 50% of the translations were deemed (totally) correct in each language. The results are not equivalent or uniform in the languages tested. The machine translation system performed noticeably better in the English to Spanish direction than in other language translations. One possible explanation for this may well lie in the fact that English and Spanish are much more similar (word order, inflections, etc.) than English and Chinese, Korean or Russian.

Our findings suggest that off-the-shelf and general purpose machine translation systems in their present state are unlikely to be of real help to non-English speaking-patients in understanding their medical records. First and foremost, the incorrect translations could seriously misinform patients and lead to more serious safety problem. Because of liability issues, it is hard to imagine any PHR application would incorporate such a tool unless the accuracy of the translation improves dramatically. The comprehensibility of the translated sentences is also too low for practical use.

We do believe, however, that machine translation could be dramatically improved to the point in which it could be useful and helpful for the task at hand, starting with the possibility of including medical vocabularies or glossaries if one so desired. It should be pointed out that by our observation the majority of the incorrect translations appear to be associated with medical terminology. However, the irregular or complex

grammar structure of medical reports is also a source of errors. Machine translation is a type of natural language processing application and with regards to medicine and medical texts, it has been more successful when applied to more narrowly defined domains (e.g. radiology). If a machine translator is trained for a very narrow domain in medicine (e.g. medication instruction), and is equipped with a comprehensive medical vocabulary, it would become much more reliable, as the vocabulary would be more controlled, and so will the grammar. When dealing with a specific and relatively small domain, the syntactical variance that needs to be addressed is reduced and can be more easily tackled.

One of the limitations of this pilot study is the relatively small sample size and number of coders. The coders of this study are bilingual, have an educational level and exposure to medical terminology well above those of the average US population, and presumably higher than those of the average non-English speaking population in the US. It is to be expected that an average non-English speaking patient would find even fewer of the translated sentences understandable. Another limitation is that the Babel Fish translator we used may not be the best machine translator. As mentioned in the background section, the PAHOMTS system would most likely produce higher quality translations and improve the results for Spanish, but it could only be used for much fewer language options. Nevertheless, based on our experience with the PAHOMTS and other machine translation tools, comprehension and accuracy are commons issues for all machine tools.

In order not to leave one of the most in-need populations – the people with limited or no English proficiency behind – in the development of personal health records, we intend to further explore the use of MT technology. As a start, we would focus on one or two narrow and relatively simple domain areas and employ strategies (e.g. translate the translation back to the original language) for quality assurance purposes.

Acknowledgments

This work is supported by the National Institute of Health (NIH) grant R01 LM07222 and by the Intramural Research Program of the NIH, National Library of Medicine /Lister Hill National Center for Biomedical Communications.

References

[1] Framework for Strategic Action in The Decade of Health Information Technology: Delivering Consumer-Centric and Information-Rich Health Care: U.S. Department of Health and Human Services2004 July 21.

[2] Patel VL, Arocha JF, Kushniruk AW. Patients' and physicians' understanding of health and biomedical concepts: relationship to the design of EMR systems. J Biomed Inform. 2002 Feb;35(1):8-16.

[3] Berkman ND, Dewalt DA, Pignone MP, Sheridan SL, Lohr KN, Lux L, et al. Literacy and health outcomes. Evidence report/technology assessment (Summary). 2004 Jan(87):1-8.

[4] Dewalt DA, Pignone MP. The role of literacy in health and health care. American family physician. 2005 Aug 1;72(3):387-8.

[5] Hornberger J, Gibson CJ, Wood W, Dequeldre C, Corso I, Palla B, et al. Eliminating language barriers for non-English-speaking patients. Med Care. 1996 Aug;34(8):845-56.

[6] Bradshaw M, Tomany-Korman S, Flores G. Language barriers to prescriptions for patients with limited English proficiency: a survey of pharmacies. Pediatrics. 2007 Aug;120(2):e225-35.

[7] Slocum J. A survey of machine translation: its history, current status, and future prospects. Comput Ling. 1985;11(1):1-17.

[8] Willée G, Schröder B, Schmitz H. Machine translation today and tomorrow. . Computerlinguistik: was geht, was kommt? Computational linguistics: achievements and perspectives. 2002:159-62.

[9] Oard DW, editor. A comparative study of query and document translation or cross-language information retrieval. Third Conference of the Association for Machine Translation in the Americas (AMTA); 1998 Oct 28-31.

[10] Abdus Salam M, editor. Machine translation and multilingual communication on the internet. . MT2000: machine translation and multilingual applications in the new millennium; 2000 Nov 20-22.

[11] Keselman A, Tse T, Crowell J, Browne A, Ngo L, Zeng Q. Assessing Consumer Health Vocabulary Familiarity: An Exploratory Study. Journal of Medical Internet Research. 2007;9(1):e5.

[12] Starlander M, Bouillon P, Rayner M, Chatzichrisafis N, Hockey BA, Isahara H, et al. Breaking the language barrier: machine assisted diagnosis using the medical speech translator. Stud Health Technol Inform. 2005;116:811-6.

[13] Chatzichrisafis N, Bouillon P, Rayner M, Santaholma M, Starlander M, editors. Evaluating task performance for a unidirectional controlled language medical speech translation system. International Workshop on Medical Speech Translation; 2006 June 9 New York.

[14] Nyberg EH, Mitamura T, Carbonell JG, editors. Evaluation metrics for knowledge-based machine translation. International Conference on Computational Linguistics; 1994 Aug 5-9; Kyoto, Japan.

[15] Zeng-Treitler Q, Goryachev S, Weiss S, Sordo M, Murphy S. Extracting principal disgnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak. 2006;6:30.

[16] Kim H, Goryachev S, Rosemblat G, Keselman A, Browne A, Zeng-Treitler Q. Beyond surface characteristics: a new health text-specific readability measurement. Proc AMIA. 2007 In press 2007.

Address for correspondence

Qing Zeng-Treitler
Department of Biomedical Informatics, University of Utah
26 South 2000 East, Room 5775 HSEB
Salt Lake City, UT 84112
q.t.zeng@utah.edu