Comparison of the quality of two speech translators in emergency settings: A case study with standardized Arabic speaking patients with abdominal pain

SPECHBACH, Hervé, et al.

Abstract

In the context of the current European refugee crisis, at the Geneva University Hospitals, 52% of the patients are foreigners and 10% don't speak French at all. In 2015, the languages which caused most problems were Tigrinya, Arabic and Farsi. Several researchers pointed out the serious problems for quality, security and equity of medical care in such a situation. BabelDr is a joint project of Geneva University's Faculty of Translation and Interpreting and Geneva University Hospitals. BabelDr application is a flexible speech-enabled phrase-book. The linguistic coverage is organised into domains, centered around body parts (abdomen, chest, head, kidney/back). Each of the four domains has a semantic coverage consisting of a pre-specified set of utterance-types, but users can use a wide variety of surface forms when speaking to the system. Each utterance-type is associated with a canonical source language version, which is rendered into the target languages by suitably qualified translation experts. We compare BabelDr with Google Translate. French speaking doctors were asked to use both systems to diagnose Arabic speaking [...]
In the context of the current European refugee crisis, at the Geneva University Hospitals (HUG), 52% of the patients are foreigners and 10% don't speak French at all. In 2015, the languages which caused most problems were Tigrinya, Arabic and Farsi. Several researchers pointed to serious problems for quality, security and equitability of medical care in such situations.

BabelDr (http://babeldr.unige.ch/) is a joint project of Geneva University's Faculty of Translation and Interpreting (FTI) and Geneva University Hospitals. The BabelDr application is a flexible speech-enabled phrase-book which specifically addresses the problem of lack of qualified interpreters in hospitals in languages spoken by refugees. The linguistic coverage is organised into domains, centered around body parts (abdomen, chest, head, kidney/back). Each of the four domains has a limited semantic coverage consisting of 2000-2500 canonical sentences (which include medical questions and instructions), but users can use a wide variety of surface forms when speaking to the system. The translation is not automatic; the canonical forms are translated into the target languages by translation experts from the FTI, which guarantees the quality of translation. At runtime, the system matches the spoken doctor’s utterance to a canonical sentence and echoes it back to the source-language user, only producing a translation if the source-language user approves. The current version is unidirectional: the patient is expected to respond non-verbally, e.g. by nodding or pointing with their fingers.

In this paper, we present an experiment where we compared BabelDr with the statistical MT system Google Translate (GT) for the anamnesis in emergency settings. French speaking doctors were asked to use both systems to diagnose Arabic speaking patients with abdominal pain, based on two scenarios. For each scenario (appendicitis and cholecystitis), a patient was standardized by the HUG. Participants were four medical students and six doctors from HUG, who each performed two diagnoses, one with BabelDr and one with GT. All participants received a short introduction to both systems one week before the test and were given 30 minutes to practice. One of the doctors was dismissed because her level in French was too low.

We analysed the user's interactions with both systems, the quality of translation, the participant's ability to reach a diagnosis with the two systems as well as user satisfaction. The translation quality was evaluated in terms of adequacy and comprehensibility by three Arabic advanced translation students. Adequacy was judged on a four point scale (nonsense/mistranslation/ambiguous/correct) and comprehensibility on a four point scale (incomprehensible/syntax errors/non idiomatic/fluent). Evaluation was carried out in context and taking into account the sex of the patient (male or female). For the BabelDr translations, 93% of doctor's interactions sent to translation were correct and 94% fluent at the majority judgements. For GT, we respectively obtained 38% and 38%. Inter-annotator agreement for both evaluations was moderate (Light's Kappa for adequacy: 0.483; for comprehensibility: 0.44), according to Landis & Koch.

In this study, we found that the quality of translation has direct impact on the quality of diagnosis and satisfaction. With Google Translate 5/9 doctors found the correct diagnosis, against 8/9 with BabelDr. The satisfaction of doctors was also higher with BabelDr than with Google Translate: doctors were more confident in the translation to the target language with BabelDr than Google Translate (1/9 negative opinion with BabelDr vs 8/9 in GT). They also think they could integrate BabelDr in their everyday practice in the emergency room, contrary to GT (1/9 negative opinion with BabelDr vs 5/9 with GT).

These results tend to show that BabelDr is a promising tool for the task and that GT translations are insufficiently adequate, accurate and comprehensible for emergency settings.