INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

A Robust Medical Speech-to-Speech/Speech-to-Sign Phraselator

Farhia Ahmed[3], *Pierrette Bouillon*[1], *Chelle Destefano*[4], *Johanna Gerlach*[1], *Sonia Halimi*[1], *Angela Hooper*[5], *Manny Rayner*[1], *Hervé Spechbach*[2], *Irene Strasly*[1], *Nikos Tsourakis*[1]

[1]University of Geneva, FTI/TIM, Switzerland
[2]Hôpitaux Universitaires de Genève
[3]Geneva Society for the Deaf, Geneva, Switzerland
[4]Gypsysnail Arts, South Australia
[5]NABS Interpreting Services, South Australia

{Pierrette.Bouillon, Emmanuel.Rayner}@unige.ch,
{Herve.Spechbach, Irene.Strasly, Nikolaos.Tsourakis}@unige.ch

Abstract

We present BabelDr, a web-enabled spoken-input phraselator for medical domains, which has been developed at Geneva University in a collaboration between a human language technology group and a group at the University hospital. The current production version of the system translates French into Arabic, using exclusively rule-based methods, and has performed credibly in simulated triaging tests with standardised patients. We also present an experimental version which combines large-vocabulary recognition with the main rule-based recogniser; offline tests on unseen data suggest that the new architecture adds robustness while more than halving the 2-best semantic error rate. The experimental version translates from spoken English into spoken French and also two sign languages.

Index Terms: medical applications, speech translation, robustness, sign language

1. Motivation and background

With the increasing popularity of Google Translate (GT) and other large-vocabulary speech translators, it may seem that phrasal translators are no longer relevant. Medical practitioners are however inclined to disagree. A recent study [1] suggests that GT mistranslates typical medical questions at least 30% of the time; our own experiments support these findings. In practice, doctors are much more comfortable with simple menu-driven fixed-phrase translators, where all the translations have been prechecked by translation experts and can be assumed reliable, and there are now a number of such systems available. The most well-known is probably MediBabble (`http://medibabble.com/`), which won the "Best App for Healthcare Professionals" award at the 2011 Medical App awards. A similar app is Canopy (`https://learn.canopyapps.com/translator`).

Apps like MediBabble are reliable and useful, but they are also slow and frustrating. Although doctors prioritise reliability highest, they also value speed and ease of use [2]. It is natural to seek a compromise between the two competing paradigms. This immediately leads to the idea of building a phrasal speech translator that can map a large set of possible spoken inputs into a fixed set of semantic concepts, which, as with MediBabble, will be rendered into the target languages using predefined translations crafted by experts. The system can be made reliable by showing a backtranslation of the semantic concept to the doctor after the recognition stage, with a translation only being produced if the doctor approves the backtranslation.

The app we will demonstrate, BabelDr, is the result of a collaboration between the Geneva University Faculty of Translation and Interpreting and the A&E group at the Hôpitaux Universitaires de Genève (HUG), Geneva's largest hospital, whose goal is to produce a system of this general type. BabelDr (`http://babeldr.unige.ch/`; [3]) has been built using Regulus Lite, a platform for rapid construction of web-enabled spoken language applications that has been under development at Geneva University since 2014 [4]. It supports translation of medical examination questions from French into several languages, prioritising coverage relevant to Arabic- and Tigrinya-speaking migrants presenting at HUG's A&E and migrant health faculties. Coverage is divided up into several domains, by type of symptom (chest pain, headache, etc); each domain has a semantic coverage of on the order of 2,000–2,500 sentence types. The project has now reached the point where the app is being tested in scenarios where doctors use it to carry out diagnostic dialogues with standardised patients, with fairly positive results [5].

Since the language-pair covered by the current BabelDr prototype will not be accessible to the majority of the participants at Interspeech, we will also demo an experimental version of the system. This contains substantial new functionality and uses English as the source language and French as the target. The presentation will focus on two specific aspects:

Robustness Speech and language processing in the version of BabelDr described in [5] is entirely rule-based. The new experimental version, in contrast, integrates a large-vocabulary recogniser, which both increases robustness and more than halves the semantic error rate.

Translation into sign language As well as French, the experimental version also supports translation into Swiss French sign language (LSF-CH) and Australian sign language (Auslan).

In the remainder of this note, we describe the above issues in a little more detail.

2. Robustness

The current BabelDr prototype, which uses a pure rule-based architecture, has a semantic error rate of a little over 30% when tested by doctors in moderately realistic simulated triaging settings; in other words, about one sentence in three fails to produce a correct backtranslation. This is good enough that doctors in practice seem to arrive at a correct triaging decision most of

the time, but it is still uncomfortably high[1]. Analysis of the rejected sentences reveals that about a third of them fail because the doctor is trying to express a concept which is not in the semantic coverage; in the remaining cases, the problem is defective speech understanding. This is unsurprising in a grammar-based system: "all grammars leak".

The experimental version of the system uses a hybrid approach to speech understanding which combines the current grammar-based processing route with robust processing based on a large-vocabulary recogniser. The large-vocabulary recogniser's language model is created by interpolating a general language model with that of the grammar-based recogniser. The robust language processor attempts to find a closest match between the large-vocabulary recogniser hypothesis (call it h_{large}) and the grammar. As the companion paper explains, we explored several different approaches and were surprised to find that we got best performance from a very simple one which uses tf-idf indexing [6] and dynamic programming. The method is split into two phases. In the first phase, the candidate semantic interpretations and their associated sets of grammar rules are viewed as a collection of documents, each represented as a plain bag of words, and a short-list of rule-sets is found which maximize the td-idf score of the rule-set with respect to h_{large}. In the second phase, the structure of the rules is taken into account. Each rule-set on the short-list is matched against h_{large} using dynamic programming, and the scores obtained are used to reorder the short-list.

The robust method turns out to be considerably better than the grammar-based one, reducing the speech understanding error by about 30% relative. What surprised us most is how effectively the grammar-based and robust methods combine to form a hybrid system. For reasons we still do not understand, errors in the two methods turn out to be very weakly correlated; this produces a startling reduction in the 2-best speech understanding error, which is more than halved. The experimental version of the system exploits this by returning multiple speech understanding hypotheses; the number of hypotheses returned (n) is controlled by the user, with a default of 2. Our expectation, based on previous systems of this kind that we have developed [7], is that novice users will set n high, using the multiple speech understanding hypotheses to get information about the system's coverage, while experienced users, who already know the coverage well, will prefer a low value of n.

3. Translation into sign language

As noted, the experimental version of BabelDr supports output in the form of sign language. We summarise results from [8].

There has been surprisingly little work on true speech-to-sign-language systems. Many claimed systems build on the incorrect idea that utterances in sign language are formed from corresponding utterances in oral/aural languages by a process of substituting manual signs for words; in fact, the syntactic structures of sign languages are completely different from those of oral/aural languages, so this approach cannot work. True speech-to-sign systems are difficult to build and quite rare, the best one probably still being TESSA [9], which in 2002 was able to translate English speech into British Sign Language in a post office counter service domain. The strategy used by TESSA and similar systems is to map source language grammatical structures into target (sign language) grammatical structures, and generate the output using a signing avatar. Although we have experimented with this idea ourselves [10], the quality of avatar-generated signing still appears insufficient for a safety-critical application like medicine [11]. We consequently reverted to the much simpler idea of recording a signed language video for each semantic form, constructing a web tool to manage the recording process efficiently.

Although the recording process was straightforward, and latency is satisfactory when using a high-speed broadband connection, we found many more problems than we had expected in the actual translation process, particularly in LSF-CH; this seems to us to throw an interesting light on the question of how expressive sign languages actually are. We isolated as many as 50 concepts in the medical domain, some of them intuitively quite simple ("HIV positive", "cocaine", "groin", etc), which it turned out were not straightforward to render into sign language and required various workarounds. We are currently in the process of organising empirical tests with Deaf subjects to investigate the adequacy of our solutions.

[1] Note that although BabelDr and Google Translate both have semantic error rates of ~30%, there is a crucial difference: the BabelDr user sees an incorrect backtranslation and can discard the utterance without translating, but the Google Translate user has in general no way to know that the system has mistranslated.

4. References

[1] S. Patil and P. Davies, "Use of Google Translate in medical communication: evaluation of accuracy," *BMJ*, vol. 349, p. g7392, 2014.

[2] N. Tsourakis and P. Estrella, "Evaluating the quality of mobile medical speech translators based on ISO/IEC 9126 series: definition, weighted quality model and metrics," *International Journal of Reliable and Quality E-Healthcare (IJRQEH)*, vol. 2, no. 2, pp. 1–20, 2013.

[3] P. Bouillon and H. Spechbach, "BabelDr: A web platform for rapid construction of phrasebook-style medical speech translation applications," in *Proceedings of EAMT 2016*, Vilnius, Latvia, 2016.

[4] M. Rayner, P. Bouillon, S. Ebling, I. Strasly, and N. Tsourakis, "A framework for rapid development of limited-domain speech-to-sign phrasal translators," in *Proceedings of FETLT 2015*, Seville, Spain, 2015.

[5] P. Bouillon, J. Gerlach, H. Spechbach, N. Tsourakis, and S. Halimi, "BabelDr vs Google Translate: a user study at Geneva University Hospitals (HUG)," in *Proceedings of the 20th Conference of the European Association for Machine Translation (EAMT)*, Prague, Czech Republic, 2017.

[6] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[7] M. Starlander, P. Bouillon, N. Chatzichrisafis, M. Santaholma, M. Rayner, B. Hockey, H. Isahara, K. Kanzaki, and Y. Nakao, "Practising controlled language through a help system integrated into the medical speech translation system (MedSLT)," in *Proceedings of MT Summit X*, Phuket, Thailand, 2005.

[8] F. Ahmed, P. Bouillon, C. Destefano, J. Gerlach, A. Hooper, M. Rayner, I. Strasly, N. Tsourakis, and C. Weiss, "Rapid construction of a medical speech to sign translator," in *Proceedings of FETLT 2016*, Seville, Spain, in press.

[9] S. Cox, M. Lincoln, J. Tryggvason, M. Nakisa, M. Wells, M. Tutt, and S. Abbott, "Tessa, a system to aid communication with deaf people," in *Proceedings of the fifth international ACM conference on Assistive technologies*. ACM, 2002, pp. 205–212.

[10] M. Rayner, P. Bouillon, J. Gerlach, I. Strasly, and N. Tsourakis, "An open web platform for rule-based speech-to-sign translation," in *Proceedings of ACL 2016*, Berlin, Germany, 2016.

[11] M. Kipp, A. Heloir, and Q. Nguyen, "Sign language avatars: Animation and comprehensibility," in *International Workshop on Intelligent Virtual Agents*. Springer, 2011, pp. 113–126.