NCHLT II SPEECH RESOURCE DEVELOPMENT

Comparative Assessment

8 April 2014

PROJECT INFORMATION	
PROJECT	NCHLT II Speech Resource Development
MILESTONE	Comparative assessment
BENEFICIARY	MERAKA INSTITUTE, CSIR
AUTHORS	Febe de Wet, Ilana Wilken, Nina Titmus, Karen Calteaux
	HLT Competency Area, CSIR Meraka Institute
PRESENTED TO	Department of Arts and Culture (DAC)
VERSION	0.2
DATE SUBMITTED	08/04/2014

Table of Contents

Introduction

In 2013 the Department of Arts and Culture (DAC) posed speech-to-speech translation in National Parliament as a grand challenge to the human language technology (HLT) community in South Africa. Speech-to-speech translation requires accurate automatic speech recognition (ASR) to convert speech to text, reliable machine translation (MT) to translate the text from one language into another and text-to-speech (TTS) to verbalise the translation naturally and intelligibly. This report aims to give a brief overview of how HLTs are being used to provide support to the language units in the parliaments of other countries.

Speech-To-Speech Translation

Europe

The TC-Star project (Technology and Corpora for Speech to Speech Translation) focusses on building a speech-to-speech translation system that can deal with real life data. The data that was used in the project was collected from the parliamentary speeches held in the European Parliament Plenary Sessions (EPPS) and was used to build an open domain corpus.

The data consists of three different versions, namely the official version of the speeches, which is available on the web page of the European Parliament, the actual transcription of the speeches produced by human transcribers, and the output of an automatic speech recognition (ASR) system. All three versions of the data were used for the evaluation of the translation system.

Twenty official languages are spoken in the European Parliament and simultaneous translations of the original speech are provided by interpreters in all the official languages. About two months after the EPPS, the speeches given by the European Parliament members are made available in their final form after being corrected by the members in that time period. This is known as the Final Text Edition (FTE) and is available in all the official languages.

The FTE format aims to achieve high readability and therefore does not provide a strict word-by-word transcript. Deviations from the original speech include the removal of hesitations, interruptions, etc. Additionally, human transcribers were asked to produce an accurate verbatim transcription of the speeches of the politicians. Both these transcription types were used in the experiments for the project.

The speeches are made available on the EUROPARL website (http://www.europarl.europa.eu/). This website also provides all previous reports since April 1996. The available reports were used for the project and an English-Spanish parallel text corpus was built.

The machine translation was done and WER (word error rate) PER (position-independent word error rate) and BLEU and NIST scores were used as evaluation metrics. In this context, a metric is a measurement. A metric that evaluates machine translation output represents the quality of the output. Any metric must assign quality scores that correlate with human judgement of quality, because human judgement is the benchmark for assessing automatic metrics, as humans are the end-users of any translation output.

The word error rate is a metric that was originally used for measuring the performance of speech recognition systems, but it is also used in the evaluation of machine translation. The metric is based on the calculation of the number of words that differ between a piece of machine translated text and a reference translation. A related metric is the position-independent word error rate. This allows for the re-ordering of words and sequences of words between a translated text and a reference translation.

The central idea behind the BLEU score is that “the closer a machine translation is to a professional human translation, the better it is”. This metric calculates scores for individual segments, generally sentences – then averages these scores over the whole corpus for a final score. The NIST metric is based on the BLEU metric, but with some alterations. Where BLEU simply calculates n-gram precision adding equal weight to each one, NIST also calculates how informative a particular n-gram is.

The translation quality for the FTE corpus has significantly improved since its development and this is consistent in all evaluation metrics. However, for the verbatim transcriptions a slight degradation in the performance of the system can be seen, but this is mainly due to ungrammatical structures in the sentences.

It was also noted that there is a loss in translation performance when the input of the system is the output of the speech recognizer and not the verbatim transcriptions of the speeches. However, the loss in performance is much smaller than the word error rate of the speech recognition system. This shows that the statistical approach to speech-to-speech translation is robust with respect to errors in the speech recognition system.

This project is on-going and difficult because the domain is broad, and the speech is characterised by a large vocabulary and long sentences. However, the results obtained are encouraging. The project has proven that it is possible to directly translate the output of a speech recognition system and that the statistical approach to translation is able to recover from speech recognition errors.

Machine Translation

India

In India, MANTRA (MAchiNe assisted TRAnslation tool) is used to translate English text into Hindi. This is done in specified domains, including gazette notifications, office orders, office memorandums, and circulars in the Government of India.

The translation was implemented by building a Tree Adjoining Grammar (TAG) parser that could parse English, Hindi, Gujarati, Sanskrit and German. However, translation in the Indian context was a more pressing concern. An English-Hindi language pair in the domain of Official Language, as used in Central Government Departments, was chosen. A prototype translation system was decided upon, built and progressively refined.

The translation strategy adopted during the machine translation (MT) process is lexical tree-to-lexical tree and not the more commonly used word-to-word or rule-to-rule approaches, because the languages in the English-Hindi language pair belong to two different language families and are hence dissimilar in structure and style. Such differences are not provided for by standard translation strategies. MANTRA therefore uses the Lexicalized Tree Adjoining Grammar (LTAG) formalism to represent the English as well as the Hindi grammar.

The accuracy of the translations has been measured at over 93% within the specified domain. Translation capabilities are also being expanded to translate Hindi texts into English in the domains of personnel administration, agriculture, banking and transportation.

A way to understand the tree adjoining grammar is to compare it to a context-free grammar. The Context-Free Grammar (CFG) is a typical way of describing the grammar of a language. For example, the sentence “the dog ate” can be described as a sentence which consists of a noun phrase and a verb phrase. The description of this sentence is independent of the context.

The TAG on the other hand, is a more powerful representation because it captures context through trees. These contextual trees are better suited to translate between two language families with different grammatical structures than standard “flat” approaches. In terms of data annotation and linguistics it is far more complex to accomplish this task.

The Salient Features of MANTRA are:

- Format retention (font sizes, alignment, styles)
- Input into MANTRA can be through:
    - .rtf or .htm files
    - speech recognition programs
    - optical character recognition package
- Pre-processing tools
    - phrase marker
    - proper noun, dates and other domain specific identifiers

- spell and grammar checker

- Uses Tree Adjoining Grammar (TAG) for parsing and generation
- Custom modifications to the Earley's style bottom-up parsing algorithm to speed up the parse
- User-friendly selection tool for multiple outputs
- Online word addition and grammar creation, updation facility.

Automatic Transcription

Australia

In all the parliaments of Australia, reporters make use of HLT when transcribing parliamentary sessions, but this was chosen by the reporters themselves and the parliaments are yet to officially implement HLT. Different strategies are followed in the provincial parliaments of Australia.

Australian Parliament
In the national parliament, sessions are recorded and the transcriptions are prepared by editors who use voice recognition software, typing or a stenotype machine to input the text.

Western Australian Parliament
Reporters record sessions with a Stenograph machine by dictating their shorthand using ASR and transcribing directly from audio using ASR.

New South Wales Parliament
Reporters dictate to computers using ASR.

Canada

In the Canadian Parliament, administration personnel record debates and decisions and oversee the televising and transcription of proceedings. The morning after the parliamentary session, all transcriptions are available in both English and French. It is assumed that some form of technology is used to achieve this, but no official information could be found.

The Government of Canada has also introduced the Language Portal of Canada (http://www.noslangues-ourlanguages.gc.ca/index-eng.php).

This portal aims to:

- Disseminate and promote language resources developed in Canada
- Share and highlight Canadian expertise in the area of language
- Help Canadians communicate in both official languages
- Support and promote bilingualism.

The site is available in both English and French, but these sites are not completely identical. However, both sites have the same objective of promoting the use of Canada’s official languages, English and French. The portal provides access to language quizzes, dictionaries and writing guides. It also provides the opportunity to learn or teach English as a first and second language and learn or teach French as a second language.

Czech Republic

Since November 2008, Czech Television (the public service broadcaster in the Czech Republic) in cooperation with the University of Bohemia in Pilsen, provides live captions for the Czech Parliament broadcasts. They currently use an alternative approach to live captioning by using speech recognition directly to transcribe the spoken speech in parliament without re-speaking. Live audio is sent from Prague to Pilsen five seconds ahead of broadcasting over an ISDN telephone line. The captions produced by the ASR system are sent back immediately after speech recognition. The captioning system is currently being evaluated and members of the public can view the live captions on a teletext page. The acoustic and language models used by the captioning system were trained on approximately 100 hours of Czech Parliament broadcasts and the stenographic records of past meetings of parliament.

Because the quality of captions produced by re-speakers is generally higher, live captioning involving trained captioners using speech recognition is being developed. Developing a language model for the Czech language is one of the main challenges that has to be addressed. Czech and other Slavic languages have a high degree of inflection and a large number of prefixes and suffixes. Therefore, the vocabulary for a Czech speech recognizer must be 8-10 times larger than that for an English recognizer. The current version of the speech recognizer uses a vocabulary of more than 200 000 words, but the speech recogniser is able to operate in real-time.

Denmark

Automatic transcription in the Danish parliament started in 2006 with the Koninklijke Philips Electronics company providing its SpeechMagic speech recognition system to the Danish government. The technology has since evolved to a solution provided by a vendor of Nuance’s Danish ASR systems.

Isle of Man

In April 2008, the Parliament of the Isle of Man became the first parliament in the Commonwealth to enter the instant transcription age. The Isle of Man’s official languages are English and Manx Gaelic, and with clearly articulated English speech, recognition accuracy of 95% is achieved. The accuracy drops slightly for accented speech.

The speech recognition software relies on individual voice profiles. These are harvested by asking each of the 35 members of parliament to record a five-minute pre-prepared passage. The resulting profile is then improved over a period of time by using the corrected audio from each session to adapt its acoustic and language elements. In other words, a member's speech patterns and language usage is learned.

Context is an important element in speech recognition software. The parliamentary context is captured in a dictionary of Manx parliamentary phrases and expressions. In addition, material from Order and Question Papers are added prior to debates taking place. This approach helps to create the framework of words within which the speech recognition engine can successfully navigate.

Editors modify the text produced by the automatic system and verify the transcriptions against the audio during parliamentary sessions. In 2009, when the article was written, electronic publication within one or two days was achieved.

Japan

One of the first semi-automatic transcription systems to be deployed in parliament is the one used in the Japanese Parliament (Diet). The system was launched for evaluation in March 2010 and has been in official operation since April 2011. The ASR system that was developed to support transcription was required to meet the following criteria:

- High recognition accuracy

Plenary sessions can be transcribed very accurately and the average recognition rate achieved on this type of speech is around 90%. Committee meetings are much more difficult to transcribe automatically as these meetings are interactive, spontaneous and often heated.

- Fast turn-around

Parliamentary reporters are assigned speech for transcription in 5-minute segments. ASR should be performed almost in real-time so that parliamentary reporters can start working on the transcriptions as soon as possible, often while a particular session is still in progress.

- Compliance to the standard transcript guidelines of the Diet

The system enforces compliance with the transcription guidelines by using only parliamentary meeting records to build the lexicon and language model.

In order to achieve high recognition performance, acoustic and language models had to be customised to parliamentary speech. A large amount of data of parliamentary meetings was used for this purpose. The data consisted of an archive of official meeting records in text (around 15 million words per year) and an archive of meeting speech (more or less 1 200 hours per year). However, there were substantial differences between the official meeting records and the speech data due to the editing process followed by the parliamentary reporters. The differences arise because of the difference between the spoken and written forms of Japanese, disfluency phenomena like fillers and repairs, redundancy such as discourse markers, and grammatical corrections.

Extensive research was conducted on the differences between the official meeting records and verbatim transcripts. It was found that the majority of the differences can be modelled computationally by statistical machine translation (SMT). With the statistical model of the difference, it could be predicted what was uttered from the official records. By applying the SMT model to a huge selection of past parliamentary meeting records, a precise language model was generated. Moreover, by matching the audio data with the model predicted for every speaker turn, the actual utterances could be reconstructed. This approach was used to implement lightly supervised training of the acoustic models very effectively. In this manner, a huge speech archive that was not faithfully transcribed could be used to develop precise models of spontaneous speech in parliament. These models will also evolve in time, reflecting the change of members of parliament and topics discussed.

The ASR system was deployed as a core part of the new transcription system in the Japanese House of Representatives. Speech is captured by stand microphones in meeting rooms for plenary sessions and committee meetings. Separate channels are used for interpellators and ministers. Channel selection and speaker segmentation modules were also incorporated.

Trials and evaluations of the system have been conducted since the system was deployed in March 2010. The accuracy defined by the character correctness compared against the official record is 89.3% for 60 meetings done in 2010. When limited to plenary sessions, it is over 95%. All meetings were transcribed with at least 85% accuracy. It takes about 2.5 minutes to process a 5-minute segment assigned to each parliamentary reporter. The system can also automatically annotate and remove fillers, but automation of other edits is still under investigation. Furthermore, the semi-automated update of the acoustic and language models using the data throughout the trial operations accomplished an additional gain in accuracy of 0.7% absolute.

The semi-automatic transcription system has been in official operation from April 2011 and now handles all plenary sessions and committee meetings. The speaker independent ASR system generates an initial draft, which is corrected by parliamentary reporters. The average character correctness measured for 118 meetings held in 2011 is 89.8%, and the additional insertion errors, excluding fillers, account for 8.2%. It translates that, roughly speaking, the system's recognition error rate is around 10%, and disfluencies and colloquial expressions to be deleted or corrected also account for 10%. Parliamentary reporters still play an important role in the transcription process, but the semi-automatic transcription system is much more streamlined than the conventional shorthand scheme that was used before.

Manual Transcription

United Kingdom

The UK parliament does not use HLT in parliament. The reporters manually record and transcribe parliamentary sessions. They are aware that HLTs exist, but they argue that at the moment, no technology exists that could replace what the reporters currently do.

Conclusion

This survey seems to indicate that ASR is the most widely used HLT in parliamentary applications. ASR works particularly well in monolingual or bilingual countries. ASR is mostly deployed as part of a semi-automatic solution with human editing still being an important component of the transcription process. Operational ASR technology is mostly provided by commercial companies. No examples of operational ASR in a multilingual context could be found.

In a few instances, MT or a combination of MT and ASR is used. However, operational applications of MT are limited to text translation, no speech-to-speech translation systems have been deployed yet. The Indian case study also illustrates that standard approaches to MT are not suitable for languages that differ substantially in terms of structure and style.

Speech-to-speech translation in National Parliament is an enormous challenge. The number of languages spoken in South Africa as well as the differences between languages from different families contribute to the complexity of the problem. Developing the HLTs that would be required to support speech-to-speech translation in South Africa would therefore require sustained and adequate funding as well as careful planning and implementation.

The development of HLT relies on the availability of appropriate resources. The first steps towards achieving the grand challenge of speech-to-speech translation should therefore be in the domain of resource development. The speech, text and translation data that are generated in National Parliament can be converted to resources if the necessary measures to harvest and process the data are put into place. The subsequent development of ASR and MT can then be informed by the availability of the relevant resources.

Resources[1]

Europe

http://www-i6.informatik.rwth-aachen.de/PostScript/InterneArbeiten/Vilar_MTSummit05.pdf

http://tcstar.org/

India

http://pune.cdac.in/html/about/success/mantra.aspx

http://pune.cdac.in/html/aai/mantra.aspx

Australia

http://www.aph.gov.au/about_parliament/senate/about_the_senate

http://www.parliament.wa.gov.au/webcms/webcms.nsf/content/hansard

http://www.parliament.nsw.gov.au/Prod/parlment/publications.nsf/key/ParliamentHouse,HansardintheParliamentofNSW/$File/History+Bulletin+7.pdf

Canada

http://publications.gc.ca/collections/Collection/YL2-12-2002E.pdf

Czech Republic

http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2009-04/tv-captioning/

Denmark

http://www.sail-labs.com/news-events/press-releases.html?tx_clpresse_pi1%5BshowUid%5D=121

http://www.dictus.dk/

Isle of Man

http://www.tynwald.org.im/business/hansard/Documents/voice-recognition.pdf

Japan

http://www.ar.media.kyoto-u.ac.jp/EN/bib/intl/KAW-ICASSP09.pdf

http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5684907&tag=1

http://www.ar.media.kyoto-u.ac.jp/diet/intl/KAW-IAAI12.pdf

United Kingdom

http://www.parliament.uk/get-involved/outreach-and-training/resources-for-universities/open-lectures/the-history-workings-and-future-challenges-of-hansard/

[1] All links cited were last accessed 8 April 2014