**Simultaneous translation of lectures and speeches**

With increasing globalization, communication across language and cul- tural boundaries is becoming an essential requirement of doing business, delivering education, and providing public services. Due to the considerable cost of human trans- lation services, only a small fraction of text documents and an even smaller percentage of spoken encounters, such as international meetings and conferences, are translated, with most resorting to the use of a common language (e.g. English) or not taking place at all. Technology may provide a potentially revolutionary way out if real-time, domain-independent, simultaneous speech translation can be realized. In this paper, we present a simultaneous speech translation system based on statistical recognition and translation technology. We discuss the technology, various system improvements and propose mechanisms for user-friendly delivery of the result. Over extensive component and end-to-end system evaluations and comparisons with human translation perfor- mance, we conclude that machines can already deliver comprehensible simultaneous translation output. Moreover, while machine performance is affected by recognition errors (and thus can be improved), human performance is limited by the cognitive challenge of performing the task in real time.

Keywords: Simultaneous translation, Interpretation, Speech-to-speech translation, Spoken language translation, Machine translation, Speech recognition, Lecture recognition, Lectures, Speeches

Abbreviations

AMI Meeting transcription data from Augmented Multi-party Interaction (see Table 3)

ASR Automatic speech recognition

BN Data from broadcast news corpora (see Table 3)

CHIL Computers in the human interaction loop

cMLLR Constrained maximum likelihood linear regression DG Directorate general

EC European Commission

EM Expectation maximization

EPPS European Parliament plenary sessions GALE Global Autonomous Language Exploitation GWRD Data from Gigaword corpus (see Table 3)

HNSRD Data from UK Parliament debates (see Table 3)

ICSI (Data recorded at) International Computer Science Institute JRTk Janus Recognition Toolkit

MLLR Maximum likelihood linear regression MT Machine translation

MTG Meeting transcription data (see Table 3)

NIST National Institute of Standards and Technology—MT evaluation measure (Doddington 2002) Data recorded at NIST

OOV Out of vocabulary

PROC Data from conference proceedings (see Table 3) RTF Real-time factor

RT-06S NIST 2006 Rich Transcription evaluation

RWTH Rheinisch-Westfälische Technische Hochschule (Aachen) SMNR Lectures and seminars from the CHIL project

SMT Statistical machine translation

SRI Stanford Research Institute

SST Speech-to-speech translation

STR-DUST Speech TRanslation: Domain-Unlimited, Spontaneous and Trainable SWB Data from switchboard transcriptions (see Table 3)

TC-STAR Technologies and Corpora for Speech-to-Speech-Translation TED Translanguage English database

TH Technische Hochschule

TTS Text to speech UKA-*. . .* See Table 3

UW-M See Table 3

VAD Voice activity detection

VTLN Vocal tract length normalization WER Word error rate

1. Introduction

With advancing globalization, effective cross-cultural communication is rapidly becoming an essential part of modern life. Yet, an estimated 4,000 to 6,000 languages still separate the people of the world, and only a small intellectual elite gather by agree- ing to communicate in a common language, usually English. This situation creates a dilemma between a need to integrate and interact and the need to maintain cultural diversity and integrity. It also limits international discourse to an adopted communica- tion medium that does not always reﬂect the culturally nuanced, individual subtleties of the global human experience. To provide access to other languages unimpeded, however, requires translation, but due to the enormous cost of human translation, only a small fraction of text documents are presently translated and only a handful of human spoken encounters are actually interpreted, if they even take place at all as a result of the separation.

All this could change if affordable, real-time, simultaneous speech-to-speech trans- lation (SST) was possible by machine. Machines never suffer fatigue, nor are they lim- ited in principle by memory size or processing speed. They could potentially also open communication channels between uncommon spoken languages, so as to broaden the exchange and interaction between the people of the world.

As a result, automatic SST as a research endeavor has been enjoying increas- ing attention. After humbleﬁrst beginnings in the 1990s, when limited phrase-based andﬁrst domain-dependent speech translation systems were proposed ( Waibel et al. 1991; Morimoto et al. 1993) large-scale domain-unlimited translation systems are now beginning to emerge (Waibel and Fügen 2008). Three recent projects in particular have begun to advance the state of the art from domain-dependent to domain-independent capabilities: the NSF-ITR project STR-DUST, the EC-IP project TC-STAR, and the DARPA project GALE (see Sect. 3 for more details about these projects). They bring within reach the translation of broadcast news and political speeches between several major languages.

As systems developed within these three projects focus on the offline processing of huge data material, translation quality instead of speed was of primary importance. As we will show in this paper, however, even simultaneous domain-unlimited translation is possible thereby opening up the range of applications to interactive lectures, seminars, speeches, and meetings. To do so, we have combined advances in speech translation technology from these projects with advances in human interface technology devel- oped under the European Commission (EC) integrated project CHIL (Computers in the Human Interaction Loop) (Waibel et al. 2004; Waibel and Stiefelhagen 2009).

While the work within CHIL focused on the delivery and interface aspect of a simultaneous translation service and the integration into the Smart Room, core rec- ognition and translation components were advanced under TC-STAR. In the present paper, we describe how these technologies were extended to deliver real-time simul- taneous speech translation of spontaneous lectures. We will discuss how the added requirements found in technical university lectures and seminars, such as real-time performance, online sentence-level segmentation, special terms for lecture vocabular- ies, and disﬂuencies, can be addressed. In human user experimentation,ﬁnally, we will attempt to establish an assessment of whether simultaneous lecture translation is possible and if current output can lead to usable results. We will conclude that, while human interpreters are achieving a higher translation quality, automatic sys- tems can already provide usable information for people who would otherwise be unable to understand the source-language lecture. Moreover, we will see that present system performance trails human performance by less than expected as human trans- lators suffer from cognitive limitations in face of the real-time performance require- ment.

Differences between interpreting and translating

Although the terms “translation” and “interpreting” are used interchangeably in every- day speech, they vary greatly in meaning. Both refer to the transfer of meaning between two languages. However,*translation*refers to the transfer of meaning from text to text, often with time and access to resources such as dictionaries, glossaries, etc. On the other hand*interpreting*consists of facilitating oral or sign language communication, either simultaneously or consecutively, between two or more speakers who are not speaking the same language.1 The profession expects interpreters to be more than 80% accurate and translations, by contrast, over 99% accurate.2

Simultaneous interpreting is sometimes incorrectly referred to as “simultaneous translation” and the interpreter as the “translator”. However, in computer science the term “machine translation” (MT) is commonly used for systems translating text or speech from one language to another. The reason for this is that in the past the main focus of MT was the translation of text and only recently is SST attracting a wider interest. Therefore, throughout this paper we use the term*translation*as in “SST”, “simultaneous translation” or “simultaneous speech translation” to mean the auto- matic*interpretation*of spoken language.

Outline

The paper is structured as follows. In Sect. 2, challenges and advantages of human inter- pretation and automatic simultaneous translation are compared. Section 4 then explains aﬁrst baseline system for simultaneous translation, which is then further adapted to achieve a lower latency and better translation quality. The necessary techniques used for speaker and topic adaptation are explained in Sect. 5. Section 6 copes with the problem ofﬁnding a chunking of the speech recognizer’s output which achieves an optimal translation performance. Therefore, different chunking strategies are com- pared in an empirical study. The overall simultaneous translation system, its architec- ture and informationﬂow is presented in Sect. 7, and Sect. 8 discusses the latency and real-time issues of such a system and possible solutions. Section 9 concentrates on the automatic and human end-to-end evaluation of the simultaneous lecture translator. Finally, Sect. 10 concludes the paper.

1. Human interpretation versus automatic simultaneous translation

Anyone speaking at least two different languages knows that translation and especially simultaneous interpreting are very challenging tasks. A human translator has to cope not only with the special nature of different languages like terminology and compound words, idioms, dialect terms or neologisms, unexplained acronyms or abbreviations and proper names, but also with stylistic differences and differences in the use of punctuation between two languages. And, translation or interpretation is not a word- by-word rendition of what was said or written in a source language; instead the meaning and intention of a given sentence has to be re-expressed in a natural andﬂuent way in another language.

* 1. Translating and interpreting in the EC

The majority of professional full-time conference interpreters work for international organizations like the United Nations, the European Union, or the African Union, whereas the world’s largest employer of translators and interpreters is currently the EC with its two Directorate Generals for Translation and Interpretation.3

The Directorate General (DG) for Interpretation is the EC’s interpreting service and conference organizer and provides interpreters for about 50–60 meetings per day in Brussels and elsewhere. The language arrangements for these meetings vary consider- ably, from consecutive interpreting between two languages, for which one interpreter is required, to simultaneous for 23 or more languages, which requires at least 69 inter- preters. At present, the Council of the European Union accounts for around 46% of the interpreting services provided, followed by the Commission itself with around 40%. There are more than 500 staff interpreters, accompanied by 2,700 accredited freelance interpreters.4

In 2006, the European Parliament spent aboute300 million, or 30% of its budget,

on the interpretation and translation of the parliamentary speeches and EU documents. In total, approximatelye1.1 billion are spent per year for the translating and interpret- ing services within the European Union, which is around 1% of the total EU budget .

* 1. Challenges in human interpretation

According to Al-Khanji et al. (2000), researchers in theﬁeld of psychology, linguistics and interpretation like Henderson (1982), Hendricks (1971) and Seleskovitch (1978) seem to agree that simultaneous interpreting is a highly demanding cognitive task involving a basic psycholinguistic process. These processes require the interpreter to monitor, store and retrieve the input of the source language continuously in order to produce the oral rendition of this input in the target language. It is clear that this type of difﬁcult linguistic and cognitive operation will force even professional interpreters to resort to elaborate lexical or synthetic search strategies.

Fatigue and stress negatively affect the interpreter, leading to a decrease in interpret- ing quality. In a study by Moser-Mercer et al. (1998) in which professional speakers were told to work until they could no longer provide acceptable quality it was shown that (a) during theﬁrst 20 minutes the frequency of errors rose steadily, (b) the inter- preters however appeared to be unaware of this decline in quality, (c) at 60 minutes, all subjects combined committed a total of 32.5 meaning errors, and (d) in the cate- gory of nonsense the number of errors almost doubled after 30 minutes on the task. In experiments with students and professional interpreters, Al-Khanji et al. (2000) found that the most frequent*compensatory strategies*, which are also used in such situations, are—in the order of occurrence—skipping, approximation,ﬁltering, comprehension omission, and substitution. Therefore, in Vidal (1997), the conclusion was drawn that interpreters should work in teams of two or more and have to be exchanged every 30 minutes. Otherwise, the accuracy and completeness of simultaneous interpreters decreases precipitously, falling off by about 10% everyﬁve minutes after holding a satisfactory plateau for half an hour.

* + 1. *Fluency and the ear–voice span*

Since the audience is able to evaluate the simultaneously interpreted discourse only by its form, theﬂuency of an interpretation is of utmost importance. According to a study by Kopczynski (1994, pp. 87–100),ﬂuency and style was third on a list of priorities of elements rated by speakers and attendees that contribute to quality after content and terminology. Following the overview in Yagi (2000), an interpretation should be as natural and as authentic as possible, which means that artiﬁcial pauses in the middle of a sentence, hesitations, and false starts should be avoided (Jones 1998), and tempo and intensity of the speaker’s voice should be imitated (Kopczynski 1994).

Another point to mention is the time span between a source-language chunk and its target-language chunk, which is often referred to as “ear–voice span”, “delay”, or “lag”. Following the summary in Yagi (2000), the ear–voice span is variable in dura- tion depending on some source- and target-language variables, like speech delivery rate, information density, redundancy, word order, syntactic characteristics, and so on. Nevertheless, the average ear–voice span for certain language combinations has been measured by many researches, but the range varies largely from two to six seconds (Barik 1969; Lederer 1978), depending on the speaking rate. Short delays are usually preferred for several reasons. For example, the audience is irritated when the delay is too large and is soon asking whether there is a problem with the interpretation. Another reason is that a short delay facilitates the indirect communication between the audience and the speaker but also between people listening to the interpreter and to the speaker. Therefore, interpreters tend to increase their speaking rate when the speaker hasﬁnished. In English, a latency of about two to six seconds is equivalent to a delay of about 4 to 12 words, since the average speaking rate is about 120 words per minute (Ramabhadran et al. 2003; Yuan et al. 2006).