Toward Practical Spoken Language Translation

Chengqing Zong

National Laboratory of Pattern Recognition

Institute of Automation

Chinese Academy of Sciences

P. O. Box 2728

Beijing 100080, China

cqzong@nlpr.ia.ac.cn

Mark Seligman

GETA-CLIPS

Université Joseph Fourier

Grenoble, France

and

Spoken Translation, Inc.

mark.seligman@spokentranslation.com

Abstract

This paper argues that the time is now right to field practical Spoken Language Translation (SLT) systems. Several sorts of systems can be built over the next few years if system builders recognize that, at the present state of the art, users must cooperate and compromise with the programs. Further, SLT systems can be arranged on a scale, in terms of the degree of cooperation or compromise they require from users. In general, the broader the intended linguistic or topical coverage of a system, the more user cooperation or compromise it will presently require. The paper briefly discusses the component technologies of SLT systems as they relate to user cooperation and accommodation ("human factors engineering"), with examples from the authors' work. It describes three classes of "cooperative" SLT systems which could be put into practical use during the next few years.

Keywords: spoken language translation, SLT, speech translation, speech-to-speech translation, user cooperation and accommodation, cooperative SLT systems, human factors engineering, machine translation, automatic translation

1 Introduction

Spoken Language Translation (SLT) has made substantial progress since the demonstration of the first experimental speech-to-speech (S2S) translation system, SpeechTrans, in 1989 (Tomabechi *et al.*, 1989; Kitano, 1994). However, the technology still largely remains in the laboratory. As yet, it is being actively put into practice only very rarely.

This paper will urge that the time is now right to field practical speech translation systems – that several sorts of SLT systems can in fact be built over the next few years if system builders recognize that, at the present state of the art, users must cooperate and compromise with the programs.

With this goal in view, we will further suggest that SLT systems can be usefully ordered, or arranged on a scale, in terms of the degree of cooperation and/or compromise they require from users. We will also point out a direct relation between the degree of cooperation a system requires and its intended linguistic coverage or topical coverage. By linguistic coverage, we mean the system's tolerance for freely varied structures and vocabulary, grading from complete freedom when linguistic coverage is maximal to a set of fixed phrases when it is minimal. By topical coverage, we mean the system's relative ability to move from one conversational domain (such as the making of hotel reservations) to another (such as the interchange between doctors and patients): maximal topical coverage would imply complete freedom to change topics, while minimal coverage would mean restriction to a single narrow topic. In general, the broader the intended linguistic or topical coverage of the system, the more user cooperation or compromise it will presently require.

The paper's objectives, in other words, are to call attention to the importance of user accommodation (or "human factors engineering") for bringing speech translation into practical use; to discuss aspects of speech recognition (SR), machine translation (MT), and text-to-speech (TTS) relevant to such accommodation, presenting some examples from our own work; and to usefully categorize speech translation systems which may become practical soon, according to the degree and type of user accommodation which they require.

These objectives are quite modest. In particular, the call for greater attention to human factors in speech translation systems will strike some readers as obvious. We certainly do not present the point as strikingly new or profound. Nevertheless, it is worth making explicitly, if only because it remains insufficiently honored in practice.

"Human factors engineering" is indeed engineering. Thus the topic may seem out of place in a journal normally devoted to more scientific or technical contributions. Nevertheless, we offer this discussion in the conviction that such engineering has become crucial for fielding practical speech translation systems in the near term; and that fielding practical systems in the near term has become crucial for sustaining the necessary momentum for scientific progress.

As a straw man against which user accommodation can be measured in later discussion, we can consider an idealized speech translation system requiring *no* user cooperation or compromise at all – and immediately reject it as over-ambitious at present. This would be a communicator straight out of Star Trek or Hitchhiker's Guide to the Universe, providing effortless and transparent translation. Equipped with such a communicator, you could talk just as you habitually

do when addressing fellow native speakers of your own language: you could freely shift topics; use your full range of vocabulary, idioms, and structures; use extemporaneous language full of fragments, false starts, and hesitations; mumble; converse in noisy environments; and completely ignore the translation program. Meanwhile, your foreign-speaking partner would hear a simultaneously and perfectly translated version of your utterances.

Of course, few SLT researchers would expect this vision to be achieved soon. Even neglecting issues of speech input and output, most researchers in automatic translation have already indefinitely postponed the goal of fully automatic high-quality machine translation (FAHQMT) for any domain or text type (Arnold *et al.*, 1994; Boitet, 1996b). Clearly, if such effortless automatic operation cannot presently be achieved in text-only translation, its achievement in S2S translation is even less likely. In fact, even human interpreters rarely perform as transparently as the Star Trek gadget: they often interrupt dialogs to clarify the intentions of the speech partners (Oviatt and Cohen, 1991).

But while speech translation researchers may be unlikely to harbor unrealistic expectations, naïveté remains a real danger in the user community at large – a problem which we researchers ignore at our own risk, since over-inflated expectation often leads to exaggerated disappointment. Researchers may in fact unintentionally encourage unrealistic expectations through the natural, and usually praiseworthy, tendency to concentrate on interesting and ambitious research problems while postponing work on more mundane practical issues. In several current research systems, for instance, spontaneous speech is expected, in which an input utterances may be fragmentary or ill-formed (Juang, 1998; Furuse *et al.*, 1998); vocabulary is open (Lavie *et al.*, 1999); speech is entered through standard mobile phones instead of microphones (Wahlster, 2000); etc. Observers

outside of the research community could be forgiven for assuming that unrestricted SLT is just around the corner. Our point, of course, is not that SLT researchers should avoid experiments with the most difficult and demanding applications; it is simply that expectations for *practical* SLT systems in the next few years should be grounded in the current technical reality, with the full understanding and cooperation of system users.

In order to achieve more practical SLT in the near term, then, what sorts of cooperation and compromise might users be expected to contribute? Depending on the system design, they might be asked to:

- speak loudly and clearly
- speak in quiet environments
- accept restrictions on the use of audio input equipment, networks, etc.
- correct speech recognition errors, by voice or through other media
- train speaker-dependent acoustic models
- provide only well-formed input
- provide extra information in the input to aid analysis, e.g. word separations or brackets
- resolve lexical or structural ambiguities (by taking the initiative, or by responding to prompts from the system)
- provide missing information, e.g. references for zero pronouns
- tolerate rough or incomplete translations
- spell or type out words that prove hard to recognize
- use richer or more complex interfaces, e.g. including GUIs as opposed to voice-only

We have already made the straightforward observation that, in general, the broader the intended

linguistic and topical coverage of a speech translation system, the more such user cooperation or compromise it will require. In Section 3 below, we will illustrate this relation by describing three classes of "cooperative" SLT systems which can be constructed and put into practical use by a wide range of system designers in the near term. (To be concrete about "the near term," let us take as a signpost the beginning of the 2008 Olympic Games in Beijing. However, this choice should not be taken to restrict the set of languages or situations relevant to our discussion.) We will discuss the SLT system classes in order of increasing linguistic and topical coverage, and thus also in order of increasing need for user cooperation and compromise. First, however, it will be useful to briefly review (in Section 2) relevant aspects of the fundamental technologies – for MT, SR, and TTS.

2 SLT System Components and User Cooperation

During the past decade, and particularly since the establishment of the Consortium for Speech Translation Advanced Research (C-STAR) in 1991, the number of experimental SLT systems has grown impressively. Table 1 summarizes five experimental systems respectively developed by AT&T Labs, Carnegie Mellon University (CMU), ATR-SLT in Japan, the University of Karlsruhe in Germany and its partners, and the National Laboratory of Pattern Recognition of the Chinese Academy of Sciences (CAS-NLPR).

SYSTEM NAME	DEVELOPER	DATES	DOMAINS	LANGUAGES	MT	VOCABULARY
Head Transducers (Alshawi, 1996; Alshawi *et al*, 2000)	AT&T Labs	1996	travel information	English-Chinese, English-Spanish	statistical	1200/1300
JANUS-III (Lavie, 1999; Levin *et al.*, 2000)	CMU	1997-	hotel reservations，flight/train booking，etc.	English-German, English-Japanese, English-Spanish, etc.	multi-engine	open
ATR-MATRIX (Sumita *et al.*, 1999; Sugaya *et al.*, 1999)	ATR-SLT	1998-2001	hotel reservations	Japanese-English、Japanese-German, etc.	pattern-based	2000
Verbmobil (Wahlster, 2000)	Univ. Karlsruhe, DFKI etc.	1993-2000	meeting scheduling	German, English, Japanese	multi-engine	10000/2500
Lodestar (Zong *et al.*, 1999)	CAS-NLPR	1999	hotel reservations, travel information	Chinese-Japanese, Chinese-English	multi-engine	2000

Table 1. Five experimental SLT systems

Several indications of progress are visible in the table: 1) the vocabulary size of the experimental systems has increased to 2000 words or more; 2) the translation engines have developed to the point where several systems can now employ multiple approaches; 3) multiple translation languages are now offered; and 4) the systems are becoming able to operate in multiple domains. Additionally, in most of the experimental SLT systems listed, the input speech is spontaneous and the speech recognizer is speaker-independent.

System accuracy appears to have increased over time. While no figures are available for the AT&T system, JANUS I reportedly attained 89.7% translation accuracy with N-best SR hypotheses, while JANUS II attained 82.8% (Kitano, 1994). These accuracy figures are already

impressive (though they must be assessed in light of then-current limitations with respect to input vocabulary, speech style, and sentence types). Evaluations of the more current experimental systems have been widely reported (Akiba *et al.*, 2004).

Despite these indications of progress, the fact remains that speech translation has rarely moved out of the laboratory. To our knowledge, the only system in ongoing field use at present is the Phraselator, a handheld system developed for the US military, and reportedly undergoing field testing in the Middle East (www.phraselator.com). Several additional commercial systems are now in preparation (Transclick (www.transclick.com); SpeechGear (www.speechgear.com); AppTek (www.apptek.com); Spoken Translation, Inc. (www.spokentranslation.com); Sehda (www.sehda.com); NEC (Ikeda *et al.*, 2002)), but no system appears to be available for general use at the time of writing.

In the remainder of Section 2, we will prepare for later description (in Section 3) of several sorts of SLT systems which might move beyond this impasse by emphasizing user cooperation and accommodation. For this purpose, we will briefly examine some issues concerning the major components of SLT systems – SR, MT, and TTS – and their interconnection. For example, with respect to SR, we will discuss the choice between speaker-dependent and speaker-independent recognizers, and will examine the implications of language model choice for enabling user correction of SR errors. Throughout this discussion, focus will remain upon human factors issues. It is not our purpose to attempt an in-depth review of SLT system issues. (See however (Lazzari, 2000a, 2000b.))

2.1 Separation of SLT System Components

Each of the major technologies required for SLT – MT, SR, and TTS – has now reached the point of usability, at least for some applications. However, each technology remains error-prone when used individually. When the techniques are combined in an SLT system, they would ideally be integrated effectively and synergistically so that they help each other, and thus reduce, rather than perpetuate or even compound, each other's errors. Close integration of SR and MT would be especially important in this respect. Such seamless integration, however, has proven to be challenging to achieve. As a result, a division into three separate components is presently maintained in most working experimental systems.

This separation between components is clearly artificial, and can be seen as a temporary tactic to be employed only until enough is learned to implement more integrated approaches. However, the learning process is likely to continue indefinitely. Thus, while close integration of SLT system components remains an important research goal for the middle and long term, it would be unwise to postpone the building of practical systems until it is achieved. Accordingly, programs which aim to produce practical systems in the near term must treat the separation between components, and particularly between SR and MT, as a fact of life.

This de facto segregation has clear implications for human factors engineering in near-term SLT systems, our present concern: it inevitably perpetuates system shortcomings for which, at present, only the user can practically compensate. In particular, the prevailing separation between MT and SR implies that speech input must be transformed into text before the translation component can begin to handle it: for practical purposes, the system cannot currently analyze or translate while

SR is progressing (despite experiments with the processing of incomplete inputs (Matsubara *et al.*, 1999; Furuse *et al.*, 1998)). As a result, any information which might be gained by deeply analyzing input utterances will for now remain unavailable to current SR components. This lack of information can only increase the SR error rate; and when these recognition errors are passed to the MT component, they can only increase its burden in turn, even if robust parsing techniques are applied.

The argument for user cooperation and compromise thus becomes clear. These accommodations are presently the best ways to compensate for the unavoidable errors of state-of-the-art systems. As systems mature and their automatic performance improves – as the error rate of each component falls, and as progress is made in synergistic component integration – the need for user cooperation should gradually lessen. For now, however, SLT systems which aim for fully transparent or effortless operation will remain out of reach.

2.2 Speech Recognition and the User

We now discuss several factors relating to SR from the viewpoint of user cooperation or accommodation in SLT system design. As previewed, no survey the SR field as a whole will be attempted.

Speaker-dependent vs. Speaker-independent SR

One distinction which can be made among SR systems is especially salient for users: audio modeling for SR may be *speaker-dependent* or *speaker-independent* (Foster and Schalk, 1993; Rabiner, 1989; Rabiner and Juang, 1993; Gold and Morgan, 2000).

In preparation for speaker-dependent recognition, a user must read a prepared text so that the system can learn his or her individual pronunciation of each phone to be recognized, in various phonetic contexts. In the early days of commercial SR – around 1998 – creation of such an *audio model* or *voice profile* could take as much as thirty minutes; at the time of writing, some systems can effectively train in less than five minutes.

Since many recent SR systems, especially those in commercial use, strive for maximum user convenience, speaker-*independent* operation is often preferred. In this mode, a new user does not need to create an individual audio model: instead, several generic audio models are prepared in advance, and one of these is selected at the start of a recognition session as most closely matching the current speaker's pronunciation (e.g. in the Mac OS X operating system of Apple Computer, Inc., www.apple.com/macosx/features/speech).

The availability of a personally tuned audio model generally makes speaker-dependent SR more accurate than speaker-independent SR. When the latter technology is nevertheless selected for user convenience, an application with a relatively narrow domain is generally implied. While hundreds or even thousands of words may be recognizable given contextual constraints, speaker-independent applications normally do not attempt to recognize arbitrary running text with a vocabulary of tens of thousands of words.

N-grams vs. Parsing in Language Modeling

A second important distinction affecting user cooperation with SR in an SLT system relates to language modeling: a system's SR component may support *n-gram-based language models*; it

may support *parsing-based language models* expecting full analysis of the speech input; or it may supply both types of models as appropriate (Jurafsky and Martin, 2000). Unlike the distinction between speaker-dependent and speaker-independent SR components, the type of language modeling is not directly visible to system users. However, the modeling strategy does affect the user experience strongly, albeit indirectly.

First, the language modeling strategy partly determines the SLT system's potential coverage of the input language. (When the coverage is limited, users may need to accommodate to the limitations, e.g. by remaining aware of them and staying on topic, or by tolerating warnings when topic limits are breached.) Second and equally important, the choice of language model affects the potential range of user-driven strategies for correcting SR errors.

Where coverage of the input language is concerned: for very large scale recognition of arbitrary running text in an SLT system, *n-gram-based* language models (Chen and Goodman, 1996; Potamianos and Jelinek, 1998) – those which predict word sequences in the current input by consulting statistics concerning word sequences previously observed in a given corpus – offer the only practical choice at present.

Where user correction of SR is concerned: parsing-based language modeling is presently the only practical method for enabling recognition of *voice-driven commands*, as opposed to running text. As an important special case, it thus offers the only way of enabling voice-driven corrections of SR errors *during*, or alternating with, the dictation of text. For example, after inserting some text vocally, users can interrupt text insertion by saying "Scratch that" to delete the most recently pronounced phrase, or "Correct <incorrect segment>" to correct a particular word or phrase by

picking the correct word from a list of candidates. Currently, when parsing-based command recognition is used in commercial SR engines in this way, it is normally used *alongside* n-gram-based recognition for running text. The engine first attempts to recognize the current phrase (pause-delimited group) as a correction command using a grammar of such commands. If that effort fails, however, the phrase is assumed to be text, and n-gram-based language modeling is used for its recognition. This dual-language-model design is now used, for instance, in the commercial dictation systems of ScanSoft/Dragon (www.scansoft.com/ naturallyspeaking), IBM ViaVoice (www.ibm.com/software/voice/viavoice), Philips Speech (www.speech.philips.com), and others. However, the dual design has not yet been widely applied in SLT systems, as far as we are aware.

Voice-drive commands enabled through the use of parsing-based language modeling could be used for other purposes than SR correction. For instance, by using a vocal command like "Translate That," users could run the translation process once SR correction is complete. Similarly, a "Prounounce That" command could be used to execute TTS for a selected text segment; and so on.

Before leaving the topic of language modeling as it is likely to affect the potential user experience of SR in SLT systems, we should mention a further use for parsing-based SR: in addition to its use for enabling command recognition, such modeling is also attractive when the entire spoken corpus can be practically enumerated. In current commercial use, the corpus might be the set of anticipated responses in an Interactive Voice Response (IVR) system for telephony. (A representative system, that of the Amtrak railway, can be reached in the US at (800) 872-7245. A prominent vendor of IVR systems is Nuance Communications, Inc. (www.nuance.com).) In SLT

systems, it might be the set of phrases in a phrasebook. For wide-ranging discussions, by contrast, (corresponding in the text dictation domain to such wide-ranging tasks as the creation of business letters, scientific papers, professional reports, etc.), such full-coverage grammar creation presently remains impractical.

The Need for User Correction of SR: Examples

We have mentioned the enablement of voice-driven correction commands as one important application for parsing-based language modeling in SR systems generally. A few examples will illustrate the need for SR correction by users of SLT systems more specifically.

Input: 是 向 个 里 拉 饭店 吗 ？ (*Is this … Xiang Ge Li La… Hotel?*)

The four underlined characters are originally a hotel name "香格里拉" (Shangri-la), but they have been incorrectly transliterated and separated, due to the absence of this word in the SR lexicon. Until the lexicon is updated, it will be impossible to correctly recognize the input without the user's help.

Input: 有没有 去 黄山 的 问 有 路线？ (*Is there any … ask ... have… route to Huangshan mountain?*)

Here the two characters marked with dots (with the readings "ask … have") have been wrongly recognized in place of the original word "旅游 (tour)", since they have a similar pronunciation. If this result were passed to MT, the translation would be incomprehensible. However, the correct word may well appear among the top SR candidates, and thus can be presented to the user for selection.

Varying Expectations Concerning User Corrections

While discussing user correction of SR results, we can remark on a clear difference of culture or expectation, according to the SR technology in use. Commercial SR applications like the IVR system cited above, incorporating speaker-independent audio modeling and parsing-based language modeling, are most often designed to operate with maximum transparency in a limited domain, and thus without explicit correction of SR results (though users are often prompted to confirm the system's recognition before action is taken, or to pronounce the input again if confirmation was withheld). By contrast, most current dictation applications, using speaker-dependent audio modeling and n-gram-based language modeling for text recognition, are designed for much broader linguistic coverage, and thus do explicitly support error correction, whether by voice (using parsing-based language modeling) or through some other entry mode. Where an SLT system is concerned, the appropriateness of SR correction by users will of course depend on its design criteria, and especially on the desired linguistic coverage. We will return to the last point throughout Section 3.

Correction Issues for Server-based SR

While most current dictation applications support vocal correction of errors, there are important exceptions: the *transcription*, or server-based dictation, systems. (Most dictation software is still *client*-based at present.) Transcription systems (e.g. that of Philips Speech, see link above) allow users, usually doctors or other professionals, to dictate remotely using a telephone. For reasons related to transmission lags and programming intricacies, these systems have until now not supported immediate correction of SR errors. Instead, correction must be made once dictation of the entire document is complete, using client-based software. Progress toward immediate correction for *server*-based dictation may be important for certain types of SLT systems,

especially for wide-ranging discussions.

Other Input Modes to Supplement SR

Even an SR component which does include correction facilities may fail to list the desired output among its recognition candidates. In such cases, or when speaking is for any reason inconvenient, the user can abandon SR – perhaps only temporarily – and input sentences by using a keyboard, handwriting pad, or touch screen. Typed, handwritten, or touched input may be preferred if the environment is noisy; if numbers, names, addresses, or unusual words are to be entered; if privacy is desired; and so on. Thus typed or manually supplied input can serve as a backup for spoken input. In this sense, spoken input becomes optional, and a speech-to-speech translation system can become a *text*-to-speech system in fallback mode. For maximum convenience, it should be possible to freely mix manual and spoken input modes.

Other Ways for Users to Assist SR

As discussed, users can assist SR software by correcting recognition errors. However, they can also cooperate or accommodate in many other ways. Users can help SR by avoiding noisy environments, by using audio equipment properly, and by speaking loudly and clearly. (Speaking "loudly and clearly" means enunciating "announcer style", pronouncing each syllable so that it can be distinctly heard – rather than shouting, which would cause distortion, or unnaturally separating syllables, which can lead to erroneous recognition of syllables as separate words.) They can also help by consciously avoiding rare or unusual words when possible. It remains to be determined what sorts of prompts or training would most effectively elicit such accommodations.

When speaker-independent SR is used, users might additionally help by providing information about their dialects or ways of speaking which would help the system to choose closely tuned audio models. For example, in a future tourist-oriented application (such as the NEC commercial system in preparation (Ikeda *et al.*, 2002)), English speakers might choose between British and American English, indicate their sex or age, and so on.

Interactive Preprocessing for MT

In addition to verifying, in all of the ways just examined, that the text which is sent to the system's MT component is correct, users might further ease the MT task by preprocessing the text in simple ways. For example, when the input language is Chinese, Korean, or another language normally written without explicit separation between words, the user might nevertheless be asked to indicate word boundaries, e.g. using spaces. Such additional cues would normally supplement, and be compared to, any automatic preprocessing available in the system.

2.3 MT and the User

As our focus here is upon the ways in which users can assist the operation of SLT system components by accommodating or cooperating with them, we will not attempt to evaluate the merits of the various MT approaches. Nevertheless, we can cite representative approaches to give an impression of their range and to provide reference points for later discussion.

Mainstream MT methods center around four basic translation techniques: linguistic analysis-based (Boitet and Nédobejkine, 1981); example-based (Sato and Nagao, 1990; Sato, 1991); template-based (Yao and Er-bao, 2003; Zong *et al.*, 2000b); interlingua-based (Hutchins,

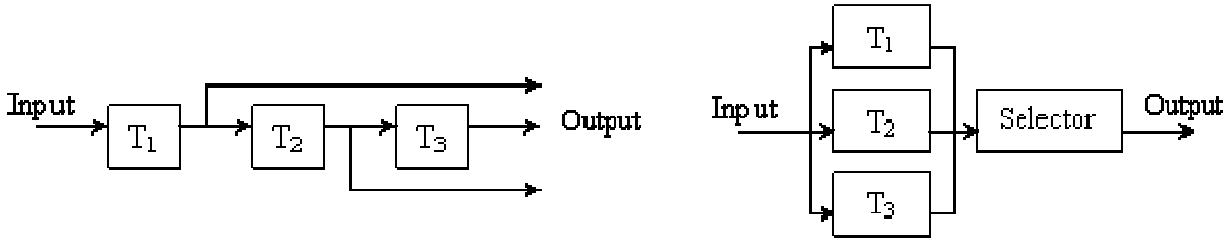
1986); and statistical (Brown *et al.*, 1990, 1993).

A number of studies have attempted to extend mainstream methods. For example,

- Ren (1999) has proposed a "super-function-based" MT method, which focuses tightly upon addressing users' requests, and accordingly translates the input without thorough syntactic and semantic analysis.
- Yamamoto *et al.* (2001) have proposed a paradigm named *Sandglass*. Input utterances from a speech recognizer are first paraphrased, and the paraphrase is then passed on to the transfer stage of translation. Paraphrasing is intended to deal with noisy input from the speech recognizer by providing various expressions of the input.
- Zong *et al.* (2000b) have proposed an MT method based on simple expressions. Keywords in an input utterance are first spotted, and dependency relations among the keywords are analyzed. Then, based on this analysis, the translation module searches for relevant examples in a knowledge base. If a source language example is matched, the corresponding target language expression is produced as the translation result.
- Wakita *et al.* (1997) have proposed a robust translation method which locally extracts only reliable parts of an utterance, i.e., those larger than a threshold size and within a semantic distance threshold from the expected meaning.

Of course, no single method can be expected to replace all others. For example, translating between structurally similar languages, e.g. between Japanese and Korean or between Spanish and French, is likely to require different techniques than translating between structurally disparate languages, like Chinese and English; some MT techniques may apply best to languages structured like Japanese and Korean, while others work for languages like Spanish and French; and so on.

Indeed, rather than attempt to choose a single best translation method, several SLT systems have in recent years combined multiple methods (Lavie *et al.*, 1999; Wahlster, 2000; Zong *et al.*, 2000a). Several combination modes are possible. For example, if T_1, T_2 and T_3 are three translation engines, then theoretically they may be combined in at least the following two ways (Figure 1):



(a) *serial combination mode* (b) *competitive combination mode*

Figure 1. Ways of combining translation engines

In the serial combination mode, the input sentence is first translated by translation engine T_1. If the translation fails, the sentence is sent to engine T_2 to be translated, and so on. Zong *et al.* (2000a; 2002) have studied such failure judgments using template-based, interlingua-based, and statistical MT engines; some findings appear in (Zuo *et al.*, 2004). An example of this serial architecture appears in Section 2.3. In the competitive combination mode, the three translation engines separately translate the input sentence or the components of the input sentence, and the system then selects the best of the three results as the output (Frederking and Nirenberg, 1994; Nomoto, 2003, 2004). Several studies have shown that the overall performance of the competitive arrangement can in fact, as hoped, be better than that of the best participating MT engine (Akiba *et al.*, 2002; Cavar *et al.*, 2000; Hogan and Frederking, 1998).

User-assisted MT

Human interpreters usually translate interactively. That is, when unable to directly translate an utterance due to ill-formed expressions or other problems, they often ask the speaker for repetition or further explanation (Boitet, 1996a, 1996b). This interactivity is not surprising, given the many difficulties which speakers may introduce: topics may be casually changed; ill-formed expressions may be uttered; and so on.

Thus it is unrealistic to expect an SLT system to outperform human interpreters by producing correct translation results without ever clarifying the speaker's intention, resolving certain lexical or structural ambiguities, etc. Based on these observations, several interactive approaches to SLT translation have been proposed (Boitet, 1996a; Blanchon, 1996; Waibel, 1996; Seligman, 1997a, 1997b, 2000; Ren and Li, 2000).

Given the wide range of MT components and combinations now available, one would expect a correspondingly wide range of techniques and designs for enabling interactive human assistance to them. We will present two examples from our own work, showing possible uses of interactive MT within two very different SLT systems. As preparation, however, we will offer additional observations concerning the locus of initiative in interactive MT.

Human-led vs. System-led Interactive MT

Most of the interactive SLT schemes and demos proposed to date have been *human-led*, in that the user decides where correction or interaction is necessary. However, the system can take a more active role, based upon its own needs – thus initiating *system-led* interactivity.

For example, the system may be designed so that it can recognize an uncertainly or failure, and thus a need for human assistance, with respect to its parsing or other translation processes. On the other hand, the system may determine that certain words in the SR result would make no difference to the translation, so correcting them would be a waste of time. In such cases, the system can actively tell the user what it really does or does not need to know.

The system's active contribution in such cases could remain quite simple. For instance, if there are words in the translation input that the system still considers ambiguous after preliminary analysis, these might be highlighted as requiring interactive lexical disambiguation. (See further the discussion of Class Three SLT systems in Section 3.3.)

Alternatively and more elaborately, to more closely simulate the active participation of a human interpreter, a *dialogue management mechanism* (DMM) might be used to guide the interaction between the system and the user. An example will appear in the next section.

User-assisted MT in SLT (Example 1): Multi-Engine MT with Dialog Management

As forecast, we now present two examples, taken from our own work, of user-assisted MT components in SLT systems.

We first illustrate the possibilities for combining several translation engines in a highly interactive STS system. Zong *et al.* (2002) have proposed such a system based on three translation engines – template-based, interlingua-based, and *slot-based* (Figure 2).

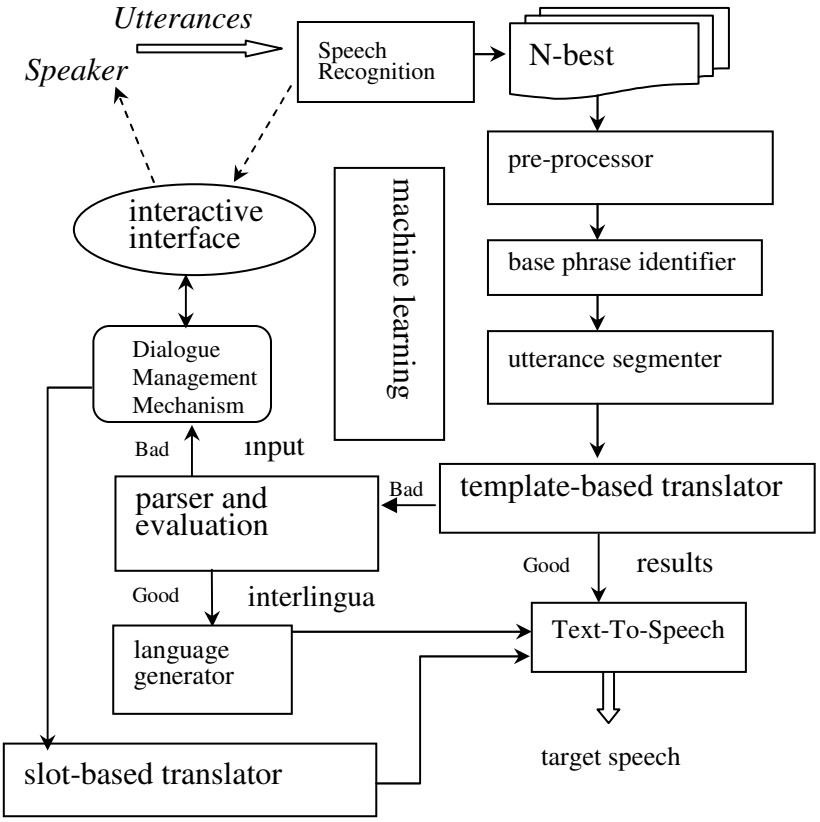


Figure 2. Interactive SLT incorporating a DMM

In this multi-engine interactive system, an input from the SR component is translated through the following four steps.

- First, following SR and selection of the N-best result, the input is pre-processed. Some noise words are recognized; some repeated words are deleted; and any numbers are processed (Zong *et al.*, 2000a). Then the base phrases (BP) in the input – mainly noun and verb phrases – are identified. In addition, if the input is a long utterance containing several simple

sentences or certain fixed expressions, it may be segmented (Liu and Zong, 2003).

- Second, each segment of the input is passed to a template-based translator. If the input segment matches a translation template, the translation result is immediately generated and sent to the TTS synthesizer directly. Otherwise, the input segment will be passed to an interlingua-based translator.

- Third, in the interlingua-based translator, the input is parsed and the parsing results are evaluated (Xie *et al.*, 2002). If the evaluation score is greater than a specified threshold value, the parsing results will be mapped into the interlingua, and the translation result will be generated by the interlingua-based target language generator.

- Otherwise, in the step of greatest interest for the present discussion of user-assisted MT, the system performs a process called *slot-based translation*. In order to disambiguate the current input segment, a dialog management mechanism (DMM) asks the user questions regarding it. According to the answers, the system fills certain slot values, and then uses these to generate the translation result.

Because this SLT scheme stresses cooperative interaction between the user and the translation system, Zong's group calls it *mutually assisting SLT*. Within this paradigm, the central research issue is the development of an effective DMM and corresponding slot-based translator.

User-assisted MT in SLT (Example 2): Lexical Disambiguation in Transfer-based MT

Our next example of user-assisted MT in an SLT system focuses upon interactive resolution of lexical ambiguities, presently in the context of transfer-based MT engines.

The second author's group has collected cues concerning the meanings of words and expressions

– synonyms, definitions, examples, etc. – from various electronic sources (WordNet, (www.cogsci.princeton.edu/~wn), online dictionaries and thesauri, etc.) and has aligned the cues with the word senses of a commercial engine for English<>Spanish MT. (Alignment with a separate commercial lexicon for English<>German is in progress, and mapping is planned to a third lexicon for English<>Japanese.) The cues can thus indicate to a user the currently analyzed sense of an ambiguous word or expression following a preliminary translation. If that sense proves to be inappropriate, the same cues can be used to select the right sense. Finally, translation can be redone, using the newly selected sense as a constraint. Further, since the translation engine has been modified in this way to enable lexical constraint during translation, word senses can also be exploited to constrain *back*-translation, thus avoiding the introduction of new ambiguities. Retro-translations can aid users to judge the faithfulness of a preliminary or revised translation. A usage sequence will help to make the procedure clear.

As we join the sequence, the user has already vocally entered, and if necessary corrected, the text "Send the file immediately." The text is thus ready for translation, triggered by the spoken (or clicked) *Translate* command. The results, including a preliminary translation, appear in Figure 3.

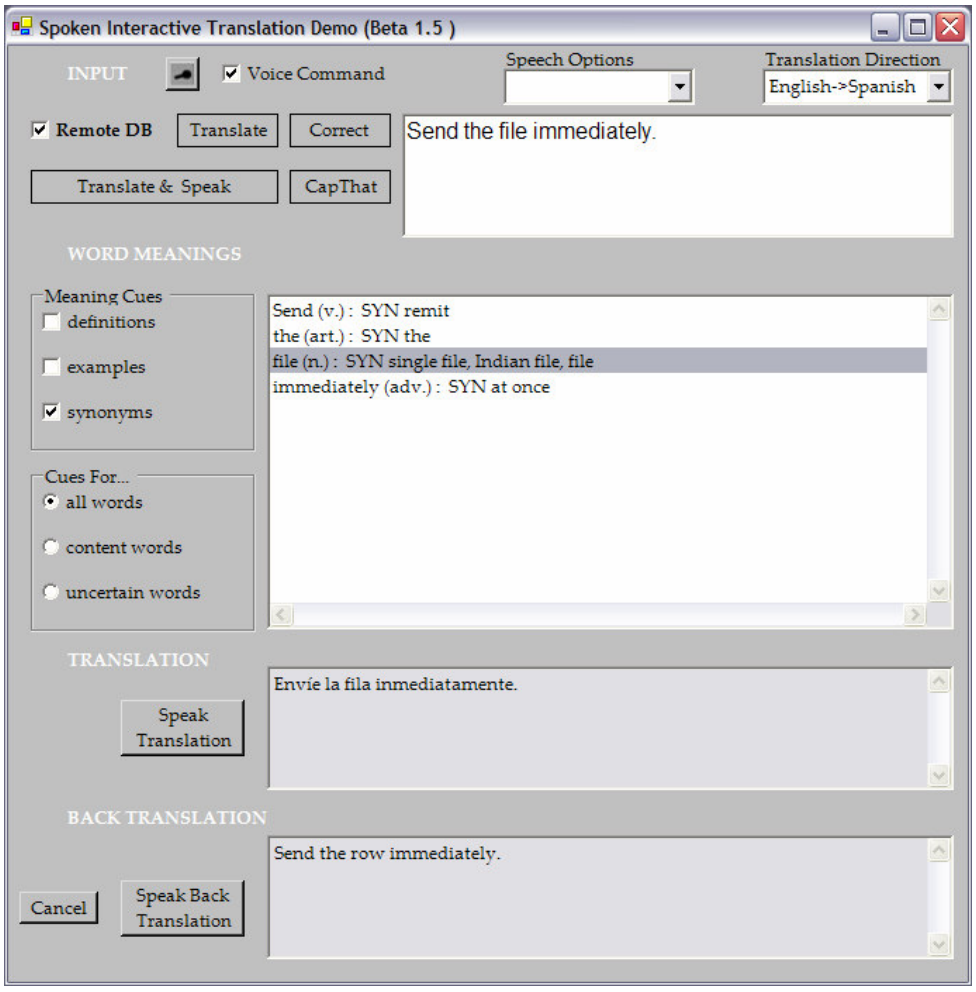


Figure 3. Results of a preliminary translation.

Also provided are (1) a Back Translation (the translated sentence re-translated back into the original language, specially constrained so that the meaning accurately represents the lexical meanings of the forward translation) and (2) an array of Meaning Cues in the Word Meanings window, indicating the word meanings used in the preliminary translation. The user can exploit all of these cues in combination to verify that the system has interpreted and translated the input as intended.

In this example, synonyms are used as Meaning Cues, but definitions, examples, and associated words can also be shown. Here the Back Translation ("Send the row immediately") indicates that the system has understood "file" as meaning "row". Presumably, this is not what the user intended. By clicking on "file" in the Word Meanings list, he or she can bring up a list of alternative word meanings (Figure 4).

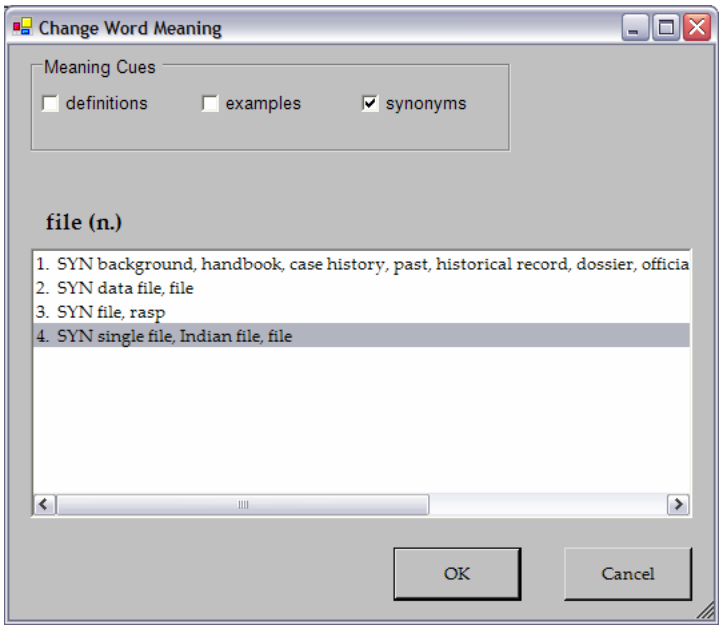


Figure 4. The user can select alternative meanings for ambiguous words.

When a new word meaning has been chosen from this list, e.g. the "background, … case history, …dossier, …" meaning, the system updates the display in all windows to reflect that change. In this example, the updated Spanish translation becomes "Enivíe el archivo inmediatamente." The corresponding back translation is now "Send the file immediately," and the Word Meanings list is updated to indicate the newly selected meaning.

When the user is satisfied that the intended meaning has been correctly understood and translated by the system, its *Send* command can be used to transmit the translation to the foreign-language

speaker via instant messaging, chat, or on-screen display for face-to-face interaction. At the same time, synthesized speech can be generated and transmitted, thus completing the speech-to-speech cycle.

2.4 Text-to-speech Methods

For our discussion of human factors engineering for SLT, we need only distinguish two basic methods of TTS or speech synthesis, *formant-based* and *concatenative*. In the first method, speech is synthesized from scratch; in the second, it is built up by sequencing together, or concatenating, short segments of recorded speech. The concatenative method presently gives more lifelike results, but requires more storage space; and space issues affect the user experience because they limit the choice of computational platform: while stand-alone handheld computers might be the platform of choice for many SLT systems, at present they often constrain storage and memory to the point that concatenative TTS becomes impractical.

Interactive Text-to-speech

No SLT systems yet allow users to control aspects of the vocal expression during TTS (or just prior to it), but the possibility exists. One could, for instance, attempt to supply emotional prosody according to cues supplied by the user. For example, a parenthesized typed or spoken abbreviation like LOL ("laughing out loud"), often seen in English-language chat, might be rendered with a synthesized laugh. Synthesized voice might also attempt to express emotions signaled via so-called emoticons, like the ubiquitous smiley face (☺), or by menu selections, etc.

3 Three Classes of "Cooperative" SLT Systems

Having discussed some aspects of SR, MT, and TTS relevant to human factors engineering in the construction of SLT systems, we now present three classes of such systems which we believe could be practically built by a wide range of researchers and developers in the next few years – in fact, by our artificial deadline, the start of the 2008 Olympics. The three classes require increasing amounts of user cooperation or compromise, but in return offer increasing degrees of linguistic coverage (freedom of lexical and structural choice) and topical coverage (freedom to switch domains). We certainly do not suggest that our discussion will cover the full range of possible system configurations, but we do believe that these classes can helpfully categorize likely developments, particularly with respect to their limitations and risks.

3.1 Class One: Voice-Driven Phrasebook Translation

linguistic coverage: narrow

topical coverage: narrow

cooperation required: low

It is possible to collect useful phrases – for ordering at restaurants, for checking into hotels, for responding to questions at customs inspections, and so on – and to store them in a suitable database along with prepared translations. A speech translation system can then directly search this phrase translation database.

A system in this class (see for instance (Zong *et al.*, 2000b), or the Phraselator project (www.phraselator.com)) would allow little or no freedom of expressive choice. Instead, users

would choose among the stored phrases. To gain slightly more freedom, they might also select *templates*, as in many existing (and profitable) travel phrasebooks in the style popularized by Berlitz and others. (A template contains at least one variable element with a specified range of values, e.g. "I'd like a bottle of [beer, wine, soda], please." Or "I'd like a bottle of [BEVERAGE], please.")

While a phrase-oriented system would probably offer a range of standard travel-oriented situational topics, the system would be useless outside of those situations (though extensions to new situations, or additions to existing ones, would always be possible).

From the user's viewpoint, such a system would have several important advantages over a classical printed phrase book: (1) there would be no need to carry a book: instead, the system would be accessed using a PDA or telephone; (2) selection of phrases or templates would be made by voice, possibly combined with other input modes, rather than by leafing through the pages or using an index; and (3) the translation output would be pronounced by a native, rather than simply being printed and pointed out to target-language readers (or bravely spoken by the user with the aid of imperfect pronunciation tips, as in "PAR-lay voo ahn-GLAY?").

From the viewpoint of system architecture, memory-based translation approaches can be used as preliminary or first-pass elements of a more ambitious translation system. If they provide usable results, no further processing will be required. If not, the input can be passed to a full-service (and more computationally expensive) system capable of translating arbitrary text.

Interface Design for Voice-Driven Phrase Translation

Voice-driven phrase translation systems would be unlikely to perform well if users were simply invited to pronounce any phrase they would like to have translated. Instead, most systems of this sort would include voice-driven dialogs (designed with VoiceXML or similar authoring systems) which would guide users toward the phrases or templates a system could handle, gradually converging on the desired phrase the manner of a decision tree. For instance, a preliminary voice prompt would present a range of topics to be selected (by voice). Once the main topic had been chosen, subtopics might be successively presented using voice prompts. Finally, three or four phrases or templates might be presented, from which a final selection could be made vocally. The prepared translations and their pronunciations would then be produced. (For added flexibility, or to cope with noisy environments, the input and output could also be graphically displayed on phones with even limited GUI capabilities.)

Since the system could retain information about the last translated phrase, a response from the target-language speaker could be easily enabled for very simple dialogs in which the range of probable response would be largely predictable. The telephone could thus be passed back and forth between speakers. (For example, the user might ask, "What beverages do you have?" The expected response would of course be a list of one or more beverages.)

Technology for Voice-driven Phrase Translation

The required technology for voice-driven phrase translation is all in place and commercially available.

- For SR, the speaker-independent, parsing-based technology now widely used for IVR

systems would be appropriate. Fixed phrases could appear in the current grammar as flat strings, while templates could include sub-grammars representing the variables.

- For translation, a range of memory-based approaches could be employed: at minimum, simple databases associating phrases with their prepared translations; at maximum, sophisticated example-based or generalizing approaches. At a middle level of sophistication, template-based approaches, in which simple finite-state grammars for translation were synchronized with simple SR grammars, would seem especially promising (since development of grammars for SR and MT could proceed in tandem).

Thus the project becomes mainly an engineering exercise, with emphasis on interface design. Accordingly, we judge the degree of risk in preparing a phrase-translation system in time for the Beijing Olympics in 2008 to be quite low.

3.2 Class Two: Robust Speech Translation within Very Narrow Domains

linguistic coverage: broad

topical coverage: narrow

cooperation required: medium

A system in this class allows users to choose expressive words and structures quite freely, but in compensation handles only a sharply restricted range of topics, e.g. hotel or event reservation. Users can be made aware of those constraints, and could cooperate by remaining tightly within a pre-selected topic. If inputs are misunderstood (as indicated through a written or synthesized

voice response), users could cooperate by repeating or rephrasing. Their reward would be that, instead of having to work down through voice-driven decision trees as in a phrase translation system, they could express their needs relatively naturally, even to the extent of using hesitation syllables or fragments. For instance, any of the following utterances might be used to begin a reservation dialog:

Uh, could I reserve a double room for next Tuesday, please?

I need to, um, I need a double room please. That's for next Tuesday.

Hello, I'm calling about reserving a room. I'd be arriving next week on Tuesday.

Such relatively broad linguistic coverage can be achieved through a variety of robust analysis techniques: the recognized text is treated as a source for pattern-driven information extraction, from which programs attempt to extract only crucial elements for translation, ignoring less-crucial segments. Thus all of the above utterances might be recognized as instances of some specialized speech act, e.g. ROOM-RESERVATION-REQUEST, by recognizing that "reserve" and "room" occur within them, along with such speech act markers as "could I", "I need", "I'm calling about", etc. Given this particular speech act, the system can attempt to extract such specific information as ROOM-TYPE (here, a double room); ARRIVAL-DATE (here, Tuesday of the following week), etc. Narrowly restricted responses, e.g. from a reservation agent, can be treated in a comparable way.

Example of the Need for Robust Parsing

An example will help to demonstrate the usefulness of such robust parsing. The following sentence is a single utterance, drawn from an ongoing conversation.

Input: 喔，那个…… 这样吧，就给我预订一个单人间吧，对，单人间，一个。

(*Oh, that … well, please reserve a single room for me, sure, a single room, one.*)

In this input, there are numerous relatively vacuous words, such as 喔 (Oh), ，那个 (that), 这样吧 (well), and so on. Obviously, if all of these words are translated, the result will be verbose and fragmentary. Actually, however, only three keywords in the input are really crucial: 预订(reserve), 一个(one), and 单人间(single room). The prepositional phrase '给我 (for me)' need not be translated.

Advantages of Class Two Systems

Class two systems have several advantages from the system builder's viewpoint.

Such systems have been widely studied: most experimental speech translation systems have in fact been of this type (as are all of the systems tabulated in Table 1, above). Thus a good deal of practical experience in building them has been gained, and could be quickly applied in the effort to field practical systems in the near term. In fact, the CMU-Karlsruhe group, with ten years' experience in constructing speech translation systems of the Class Two type (Lavie *et al.*, 1999), has already obtained grants from the Chinese government to build such systems during the time period which concerns us.

Another major advantage is that, since systems of this sort remain narrowly restricted with respect to topic, it has proven practical to incorporate within them technology, for both SR and translation, which can be tightly tuned or optimized for particular applications.

With respect to SR, as for Class One systems, it is possible to exploit standard speaker-independent technology; and because the range of input utterances is sufficiently limited, one can construct grammars covering significant segments of the relevant domain.

Concerning translation technology, the narrowness of the selected domain makes attractive the use of interlingua-based translation approaches, those using pragmatic and semantic representations applicable for several languages. Since in narrow domains the expected vocabulary, syntax, topics, and dialogue acts are limited, it is practical to define limited interlingua representations for them, and to implement effective parsers (Xie *et al.*, 2002; Xie, 2004). The most widely used interlingua within the current speech translation community is the interchange format, or IF, used by members of the C-STAR consortium. Other interlinguas, for example the UNL representation propounded by UNU (www.unl.ru), could also be tried. Whichever representation is chosen, one well-known advantage of this translation style is gained: multiple translation directions can be quickly built. Of course, styles of MT other than interlingua-based are also available for Class Two systems.

Challenges of Class Two Systems

But of course, Class Two systems face several challenges as well.

To date, most systems stressing robust recognition with narrow topical coverage have not

required users to correct SR results before sending them to analysis programs. Perhaps it has been felt that users were already compromising enough in terms of topical coverage, and thus should not be required to expend this additional editing effort as well. In any case, if this approach continues to be followed, a certain amount of noise must be expected in the input to translation.

At the same time, robust parsing remains an area of forefront research rather than a mature technology, and can itself be expected to provide noisy or imperfect results.

Thus the degree of risk in this approach must be considered at least medium: practical systems of this sort can probably be built by 2008 given sufficient resources, but considerably more research and development time will be needed than for phrase translation systems. Further, since errors are inevitable, user frustration is likely to be somewhat greater. Designers of Class Two systems must therefore hope that this degree of frustration (along with the stringent restriction as to topic) will be tolerated in exchange for greater freedom of expression, i.e. linguistic coverage.

3.3 Class Three: Highly Interactive Speech Translation with Broad Linguistic and Topical Coverage

linguistic coverage: broad

topical coverage: broad

cooperation required: extensive

A speech translation system in this third class allows users to choose expressive structures quite freely, and allows free movement from topic to topic. However, to maintain sufficient output

quality while permitting such freedom, the system requires that users monitor, and when necessary correct, the progress of both SR and translation. In effect, at the present state of the art, users of a Class Three system must pay for linguistic and topical freedom by spending considerable time and effort to help the system.

Where SR is concerned, since very broad-coverage SR is required, dictation technology appears to be the only practical choice. Thus it appears that each user will need to train an individual acoustic model, a process which would take several minutes. However, automatic or user-aided selection among several default models can also be evaluated. The underlying language models for running text must employ n-grams rather than parsing grammars. Users must also expect to correct recognition results for each utterance (by using voice commands or through other input modes) before passing corrected text to the translation stage. While considerable linguistic freedom is allowed, spontaneous speech features are not: roughly any standard grammar and vocabulary can be used, but hesitation syllables or stutters may be misrecognized as words, repetitions and false starts will appear in the visible text output, etc.

Where translation is concerned, users must provide standard, well-formed text for input: fragments or false starts will degrade quality. They must also be prepared to help the system to resolve ambiguities, e.g. by indicating the desired meaning of an ambiguous input word, or the desired attachment of an input prepositional phrase. For some translation directions, users will also need to supply information missing in the input, e.g. the referents of zero pronouns when translating from Japanese.

There are two main technical requirements for the translation engine: (1) it must provide very

broad coverage with sufficient output quality; and (2) it must allow interruption of the translation process, so that users can interactively provide lexical and structural constraints. In practice, these attributes are currently found mostly in mature transfer-based or interlingua-based systems. (The second author's group has worked to date with such MT engines from Word Magic (www.wordmagicsoft.com), Lingenio (www.lingenio.com), and Sharp (Honyaku Kore Ippon; see e.g. ourworld.compuserve.com/homepages/WJHutchins/Compendium-8.pdf).) However, some statistical systems may also prove suitable (since some lexical and phrasal constraints can be enforced by heavily weighting the desired translation choice).

Class Three Speech Translation Sessions

In general, users of a speech translation system intended for broad topical coverage will want to hear the speech partner's original speech as well as the synthesized pronunciation of the translation output. They may also want to read the translated output (and perhaps the corrected text of the input as well). Thus a system of this sort could take advantage of existing software infrastructure for chat, instant messaging, etc.

A typical interactive session might proceed as follows. Users will

- dictate text in the input language (e.g. English "Hello");
- optionally examine the dictated text and correct any dictation errors ("Hello", not "Hollow");
- obtain a preliminary text translation of the corrected text into the desired output language (e.g. German "Guten Tag");
- perhaps obtain a re-translation of the preliminary translation back into the input language (e.g. English "Good Afternoon") – a paraphrase of the input, giving some indication of the fidelity and comprehensibility of the preliminary translation;

- use interactive facilities to resolve lexical and structural ambiguities (Was that "bank" as in money, or "bank" as in river?);
- give the command, by voice or other media, to authorize transmission of the corrected translation via a shared screen, chat, instant messaging, etc. The partner can then view text representations in the messaging display and hear synthesized spoken versions of the output (e.g. German pronunciation of "Guten Tag").

Prior Demonstrations of Class Three Speech Translation Systems

Class Three spoken language translation systems have been demonstrated successfully, though with some significant limitations. The demonstrations of "quick and dirty" speech translation presented by (Seligman, 2000) in cooperation with CompuServe, Inc. did permit voice-driven correction of dictation results prior to translation and transmission via chat. Further, commercial translation technology with broad linguistic and topical coverage was in fact used (in server-based mode). Consequently, the demos did provide speech translation with very broad linguistic and topical coverage for English<>French dialogues while maintaining usable output quality (as judged subjectively by the demo audience; see the cited paper for examples and discussion).

Challenges of Class Three Speech Translation Systems

In the demos just described, SR and TTS remained client-based. For maximum flexibility of use, server-based implementation may sometimes be preferable. Further, no interactive correction of MT (as opposed to SR) was yet enabled. The second author's group is now pursuing research and development addressing both of these limitations.

Where SR is concerned, while commercial systems for server-based dictation have recently become available (e.g. for Philips Speech; see link above), current systems do not yet support immediate interactive feedback and (voice-driven or multimedia) correction. Users must instead dictate blindly, normally by telephone, and make corrections later using client-based programs. However, when remote dictation systems with real-time correction are desirable (again, to provide maximally flexible use, to and from anywhere), engineers will need to eliminate disruptive lags due to transmission inefficiencies. They will also need to implement complex client-server interactions to maintain synchronization as dictation results are obtained by servers and displayed by clients.

Where MT challenges for Class Three systems are concerned, one approach to interactive lexical disambiguation has already been outlined (in Section 2.3). Work on interactive resolution of structural ambiguities, e.g. in the mode of (Boitet and Blanchon, 1995), remains for the future.

A further challenge for highly interactive speech translation systems relates to the burden imposed by the interaction itself. Will users accept the inconvenience of monitoring and correcting dictation and translation? For socializing, they may choose to pass up such oversight entirely and to tolerate the resulting errors. However, for serious communication, we anticipate that the degree of tolerance for interaction may rise with the importance of quality results. Research is needed to confirm this expectation and explore ergonomic issues. We plan to undertake such usability testing during 2005.

In view of these challenges, the degree of risk for a Class Three speech translation system must be considered medium to high. Nevertheless, we believe that concerted development efforts could

bring several highly interactive, broad-coverage speech translation directions into practical use by 2008.

4 Conclusion

We have suggested that several sorts of practical SLT systems can be built over the next few years – specifically, in time for the 2008 Olympics in Beijing – if system designers make adequate provision for user cooperation and compromise. It was further suggested that SLT systems can be arranged on a scale, in terms of the degree of user cooperation and/or compromise they require. Generally, the broader the intended linguistic or topical coverage of the system, the more such cooperation or compromise will presently be needed. After briefly discussing the component technologies of SLT from the viewpoint of human factors engineering, with examples from our own work on two quite different systems, we have described three classes of "cooperative" SLT systems that we believe can be put into practical use by 2008.

Overall – to repeat the disclaimer with which we began – we are simply making explicit and amplifying a point which we do not expect to be controversial: that the SLT field has reached a stage at which basic and applied research should be distinguished. On the basic research side, many issues clearly remain with respect to system components – SR, MT, TTS – and their integration. While steady progress can be expected in each of these research areas, no researcher would expect the goal of fully automatic, high-quality, broad-coverage SLT to be reached anytime soon – and this expectation should be clearly communicated to the user community at large. However, on the side of applied research and engineering, it should now indeed be possible to develop a variety of useful SLT applications. While near-term SLT applications will of course

be imperfect and fall far short of meeting all needs, they can nevertheless provide useful aids for human communication. And communication – after all, the original purpose of SLT – is inherently a cooperative process. Conversational partners have always cooperated to make themselves understood. In practical SLT systems, humans and computers should likewise cooperate, collaborate, and mutually accommodate.

References

Akiba, Yasuhiro, Taro Watanabe, and Eiichiro Sumita. 2002. "Using Language and Translation Models to Select the Best among Outputs from Multiple MT Systems." In *Proceedings of COLING'2002*. pp. 8-14.

Akiba, Yasuhiro, Marcello Federico, Noriko, Kando *et al*. 2004. "Overview of the IWSLT04 Evaluation Campaign." In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Sept. 30 - Oct. 1, 2004. Kyoto, Japan. pp. 1-12.

Alshawi, H. 1996. "Head Automata for Speech Translation." In *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, PA., pp. 2360-2364.

Alshawi H., S. Bangalore, and S. Douglas. 2000. "Head-Transducer Models for Speech Translation and Their Automatic Acquisition from Bilingual Data." *Machine Translation*, 15, 105-124.

Arnold, D.J., Lorna Balkan, Siety Meijer, R. Lee, Humphreys and Louisa Sadler. 1994. *Machine Translation: an Introductory Guide*. Blackwells-NCC, London, 1994, ISBN: 1855542-17x. (See also www.essex.ac.uk/linguistics/clmt/MTbook/PostScript)

Blanchon, Hervé. 1996. "A Customizable Interactive Disambiguation Methodology and Two Implementations to Disambiguate French and English Input." In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*, Col de Porte, France.

Boitet, Christian. 1996a. "Dialogue-based Machine Translation for Monolinguals and Future

Self-explaining Documents." In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*, Col de Porte, France.

Boitet, Christian. 1996b. "Machine-aided Human Translation." In *Survey of the state of the art in human language technology*. Managing editors: Giovanni Battista Varile and Antonio Zampolli. Editorial board: Ronald A.Cole (editor in chief), Joseph Mariani, Hans Uszkoreit, Annie Zaenen, Victor Zue. Cambridge University Press, Cambridge. (Also at cslu.cse.ogi.edu/ HLTsurvey/ch8node6.html)

Boitet, Christian and H. Blanchon. 1995. "Multilingual Dialogue-Based MT for Monolingual Authors: the LIDIA Project and a First Mockup." *Machine Translation*. Vol. 9(2), pp. 99-132.

Boitet, Christian and N. Nédobejkine. 1981. "Recent Developments in Russian-French Machine Translation at Grenoble." *Linguistics*, 19, 199-271.

Brown, Peter F., J. Cocke, Stephen A. Della Pietra, *et al.* 1990. "A Statistical Approach to Machine Translation." In *Computational Linguistics*, 16(2), pp. 79-85.

Brown, Peter F., Stephen A. Della Pietra, *et al.* 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Computational Linguistics*, 19(2), pp. 263-309.

Cavar, Damir and Uwe Kussner, and Dan Tidhar. 2000. "From Off-line Evaluation to On-line Selection." In *Verbmobil: Foundations of Speech-to-Speech Translation (Wolfgang Wahlster, ed.)*. Springer-Verlag, Berlin/Heidelberg, pp. 597-610.

Chen, S. F., and J. Goodman. 1996. "An Empirical Study of Smoothing Techniques for Language Modeling." In *Proceedings of the 34th Annual Meeting of ACL-96*, pp. 310-318.

Foster, Peter and Thomas Schalk. 1993. *Speech Recognition.* Telecom Library, Inc., New York.

Frederking, Robert, and Sergei Nirenburg. 1994. "Three Heads Are Better Than One." In *Proceedings of the 4th Conference on ANLP*, Stuttgart, Germany, pp. 95-100.

Furuse, O., Satsuo Yamada, and Kazuhide Yamamoto. 1998. "Splitting Long or Ill-formed Input for Robust Spoken-language Translation." In P*roceedings of COLING-ACL,* Montreal, Canada, volume I,

pp. 421-427.

Gold, Ben and Nelson Morgan. 2000. *Speech and Audio Signal Processing.* John Wiley & Sons Inc.

Hogan, Christopher, and Robert E. Frederking. 1998. "An Evaluation of the Multi-Engine MT Architecture." In *Proceedings of AMTA '98*, pp. 113-123.

Hutchins, John. 1986. *Machine Translation: Past, Present, Future*. Ellis Horwood, Ltd., England.

Ikeda, Takahiro, Shinichi Ando, Kenji Satoh, Akitoshi Okumura, and Takao Watanabe. 2002. "Automatic Interpretation System Integrating Free-style Sentence Translation and Parallel Text Based Translation." In *Proceedings of the Workshop on Speech-to-speech Translation: Algorithms and Systems*, Philadelphia, July 2002, pp. 85-92.

Juang, B. H. 1998. "The Past, Present, and Future of Speech Processing." *IEEE Signal Processing Magazine*, pp. 24-48.

Jurafsky, Daniel and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ, USA.

Kitano, Hiroaki. 1994. *Speech-to-speech Translation: A Massively Parallel Memory-Based Approach.* Kluwer Academic Publishers, Boston.

Lavie, A., Lori Levin *et al.* 1999. "The JANUS-III Translation System: Speech-to- Speech Translation in Multiple Domains." In *Proceedings of the C-STAR II Workshop*, Schwetzingen, Germany.

Lazzari, Gianni. 2000a. "Spoken Translation: Challenges and Opportunities." In *Proceedings of ICSLP*, Beijing, China, pp. 430-435.

Lazzari, Gianni. 2000b. "Speaker-Language Identification and Speech Translation." In *Multilingual Information Management: Current Levels and Future Abilities* (E. Hovy , N. Ide., R. Frederking, *et. al.* eds.), pp. 143-166.

Levin L., A. Lavie, M. Woszczyna, D. Gates, M. Gavaldá., D. Koll, and A. Waibel. 2000. "The Janus-III Translation System: Speech-to-Speech Translation in Multiple Domains." *Machine Translation,* 15, pp. 3-25.

Liu, Ding and Chengqing Zong. 2003. "Utterance Segmentation Using Combined Approach Based on Bi-directional N-gram and Maximum Entropy." In *Proceedings the Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Matsubara, Shigeki, Hiroshi Ogawa, Katsuhiko Toyama, and Yasuyoshi Inagaki. 1999. "Incremental Spoken Language Translation Based on a Normal-Form Conversion of CFG." In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS)*. Beijing, China, pp. 515-518.

Nomoto, Tadashi. 2003. "Predictive Models of Performance in Multi-Engine Machine Translation." In *Proceedings of Machine Translation Summit IX* (IAMT). New Orleans, September 2003.

Nomoto, Tadashi. 2004. "Multi-Engine Machine Translation with Voted Language Model." In *Proceedings of ACL*. Barcelona, Spain, July 21-26, 2004, pp. 494-501.

Oviatt, S. and P. Cohen. 1991. "Discourse Structure and Performance Efficiency in Interactive and Non-Interactive Spoken Modalities." *Computer Speech and Language,* 5(4), pp. 297-326.

Potamianos, Gerasimas and Frederick Jelinek. 1998. "A Study of N-Gram and Decision Tree Letter Languages Modeling." *Speech Communication*, 24(3), pp. 171-192.

Rabiner, Lawrence R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." In *Proceedings of IEEE 77*, 7(2), February, 1989, pp. 257-286.

Rabiner, Lawrence and Bing-Hwang Juang. 1993. *Fundamentals of Speech Recognition.* Prentice-Hall, Englewood Cliffs, NJ.

Ren, F., and Shigang Li. 2000. "Dialogue Machine Translation Based upon Parallel Translation Engines and Face Image Processing." *Journal of INFORMATION,* 37(4), pp. 521-531.

Ren, F. 1999. "Super-function Based Machine Translation." *Communications of Chinese and Oriental*

Languages Information Processing Society (COLIPS), 9(1), pp. 83-100.

Sato, S. 1991. *Example-based Machine Translation*, PhD dissertation, Kyoto University.

Sato, S. and M. Nagao. 1990. "Towards Memory-based Translation." In *Proceedings of COLING-90*, vol. 3, pp. 247-252.

Seligman, Mark. 1997a. "Interactive Real-time Translation via the Internet." In *Working Notes, Natural Language Processing for the World Wide Web* (AAAI-97 Spring Symposium), pp. 142-148.

Seligman, Mark. 1997b. "Six Issues in Speech Translation." In *ACL-97 Workshop on Spoken Language Translation*. Madrid, Spain, pp. 83-89.

Seligman, Mark. 2000. "Nine Issues in Speech Translation." *Machine Translation,* 15, pp. 149-185.

Sugaya, Fumiaki, Toshiyuki Takezawa, Akio Yokoo, and Seiichi Yamamoto. 1999. "End-To-End Evaluation in ATR-MATRIX: Speech Translation System Between English and Japanese." In *Proceeding of EuroSpeech'99*. Vol.6, pp. 2431-2434.

Sumita, Eiichiro, Setsuo Yamada, and Kazuhide Yamamoto. 1999. "Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach." In *Proceedings of MT Summit VII*, Sept. 1999, pp. 229-235.

Tomabechi, H., H. Saito, and M. Tomita. 1989. "SpeechTrans: An experimental real-time speech-to-speech translation." In *Proceedings of the 1989 Spring Symposium of the American Association of Artificial Intelligence*, 1989.

Wahlster, Wolfgang. 2000. "Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System." In *Verbmobil: Foundations of Speech-to-Speech Translation (Wolfgang Wahlster, ed.)*. Springer-Verlag, Berlin/Heidelberg, pp. 3-21.

Waibel, Alex. 1996. "Interactive Translation of Conversational Speech." *Computer*, 29(7), pp. 41-48, July, 1996.

Wakita, Y., Jun Kawai, and Hitoshi Iida. 1997. "Correct Parts Extraction from Speech Recognition Results Using Semantic Distance Calculation, and Its Application to Speech Translation." In *Proceedings of a Workshop Sponsored by the ACL and by the European Network in Language and Speech (ELSNET)*, Madrid, Spain, 1997, pp. 24-29.

Xie, Guodong, Chengqing Zong, and Bo Xu. 2002. "Chinese Spoken Language Analyzing Based on Combination of Statistical and Rule Methods." In *Proceedings of International Conference on Spoken Language Processing (ICSLP-2002)*, Colorado, USA, pp. 613-616.

Xie, Guodong. 2004. *Research on Statistical Approach to Spoken Language Parsing* (in Chinese). Dissertation for Ph.D. Degree. Institute of Automation, Chinese Academy of Sciences, 2004.

Yao, Tian-shun, and Yang Er-bao. 2003. "A Method of Similar Template Based C-E Machine Translation." In *Proceedings of CJNLP-2*, Peking University, Beijing, 2003.

Yamamoto, K., Satoshi Shirai, Masashi Sakamoto, and Yujie Zhang. 2001. "Sandglass: Twin Paraphrasing Spoken Language Translation." In *Proceedings of the 19th International Conference on Computer Processing of Oriental Languages (ICCPOL- 2001)*, pp. 154-159.

Zong, Chengqing, Taiyi Huang and Bo Xu. 1999. "Technical Analysis on Automatic Spoken Language Translation Systems" (in Chinese). *Journal of Chinese Information Processing*, 13(2), pp. 55-65.

Zong, Chengqing, Yumi Wakita, Bo Xu, Kenji Matsui and Zhenbiao Chen. 2000a. "Japanese-to-Chinese Spoken Language Translation Based on the Simple Expression." In *Proceedings of International Conference on Spoken Language Processing (ICSLP-2000)*, Beijing, China, pp. 418-421.

Zong, Chengqing, Taiyi Huang, and Bo Xu. 2000b. "An Improved Template-Based Approach to Spoken Language Translation." In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, October 16-20, 2000, Beijing, pp. 440-443.

Zong, Chengqing, Bo Xu, and Taiyi Huang. 2002. "Interactive Chinese-to-English Speech Translation Based on Dialogue Management." In *Proceedings of ACL '2002 Workshop on Speech-to-speech*

Translation: Algorithms and Systems, Philadelphia, Pennsylvania, USA, pp. 61-68.

Zuo, Yuncun, Yu Zhou, and Chengqing Zong. 2004. "Multi-Engine Based Chinese-to-English Translation System." In *Procedings of the International Workshop on Spoken Language Translation*, Sept. 30 – Oct. 1, 2004, pp. 73-77.