FROM MULTILINGUAL MULTIMODAL SPOKEN LANGUAGE ACQUISITION TOWARDS ON LINE ASSISTANCE TO INTERMITTENT HUMAN INTERPRETING: SIM*, A VERSATILE ENVIRONMENT FOR SLP

Georges FAFIOTTE
CLIPS - GETA, IMAG
Université Joseph Fourier (Grenoble I) - CNRS
385, avenue de la Bibliothèque, BP 53
F-38041 GRENOBLE Cedex 9 (France)
georges.fafiotte@imag.fr

ZHAI JianShe
CLIPS - GETA, IMAG
Université Joseph Fourier (Grenoble I) - CNRS
385, avenue de la Bibliothèque, BP 53
F-38041 GRENOBLE Cedex 9 (France)
jianshe.zhai@imag.fr

ABSTRACT

We present and discuss SIM*, a versatile multiplatform simulation environment for Speech MT. Based on a Wizard of Oz scheme, it is firstly intended for supporting and collecting multimodal bilingual spontaneous spoken dialogues through the Internet, in order to later build annotated multimodal multilingual speech corpora, on task-oriented subdomains. Current prototyping investigates a symmetrical paradigm, where the system would provide two distant monolingual speakers, and especially their distant human intermittent interpreter, with automatic lexical and terminological aids derived from Speech MT active functions, through the network. Future developments to be explored are the assistance to corpus annotation and the capture of multimodal and linguistic events.

INTRODUCTION

In recent years Speech Machine Translation (SMT) applied to domain-oriented spontaneous spoken dialogues has improved greatly, namely towards multilingualism and style spontaneity, following coordinated effort such as the C-STAR project (international Consortium for Speech Translation Advanced Research) [1, 4]. In this situation there is an urge for systematically collecting actual multilingual spoken dialogues, for building large-scale annotated speech corpora, for producing specialized or general lexical bases, in order to efficiently elaborate relevant lingware components. We therefore focus our interest on a collection-and-experimentation platform to firstly meet such requirements then to somewhat widen them.

In the context of this research on Spoken Language Processing, we first drew inspiration from our previous experiments on the multimodal Wizard of Oz EMMI platform at ATR-ITL [6, 7], and from a NEIMO Wizard of Oz platform at CLIPS-IMAG [8]. The SIM* project is thus concerned with the stepwise prototyping of a multipurpose environment, for simulating ‘real life speech’ translation, collecting multimodal speech corpora, and progressively enlarging multimodal settings. We also aim at developing both LAN- and Internet-based speech collecting platforms, and at experimenting different architectures in order to enhance genericity towards multilingualism, multimodality, differential resource plug-in, and multipurpose implementation.

After introducing the research background and aims of the simulator, the paper gives an outline of the overall architecture. Then we present the SIM* environment, its current use for bilingual spoken dialogues collection, and ongoing development. Before opening prospects, we propose, derived from the basic simulation and collection platform, to integrate automatic MT-based aids for backing innovative human translation activities, namely an assistance for professional or occasional human interpreters involved in intermittent interpreting through the Internet.

1. CONTEXT, GOALS, PRINCIPLES

1.1 Research background, project context

Within the MIDDIM project (Multimodal Interactive Disambiguation [6]), were designed and ran Japanese-French and Japanese-English speech translation pilot-experiments, on the EMMI Wizard of Oz platform. As well-known [9], in Wizard of Oz environments a concealed human (here an interpreter) meets users requests in place of the software, acting to simulate, remedy or enlarge computer resources, to collect data, and to possibly observe and capture user reactions during experimentation. We also took up generic aspects in NEIMO, a multi-Wizard of Oz monolingual platform intended to observe and analyze multimodal interaction, using Telecoms Multiservicing as a host-application.

Moreover, our motivation is close to the effort of the C-STAR community to bring quality multilingual SMT into realistic domain-restricted spoken dialogues (currently with C-STAR II [1] in six languages —English, French, German, Italian, Japanese and Korean), using a common task-oriented interlingual representation. SIM* here participates in a complementary approach: a multipurpose network-based multimodal multilingual simulation platform. Within the general framework of this research, numerous multilingual spoken dialogues are to be collected, in order to train basic recognition and translation functions for both the C-STAR III and the European NESPOLE! (NEgotiating trough SPOken Language in E-commerce [1]) speech translation demonstration platforms.

1.2 Main goals, linguistic purpose

A multipurpose platform. SIM* is thus intended to,

- simulate speech interpreting of actual dialogues, adding adjustable interpreting quality, possibly from high-standard to still non-perfect linguistic and/or acoustic performance of the Wizard translation, to experiment later on speaker specific subdialogic adaptation, or rejection,
- collect multimodal speech corpora from ‘on the spot’ real spontaneous dialogues between monolingual native speakers,
• add multimodal features to basic C-STAR interaction, to be later experimented on in various situations,
• allow the testing of lingware components through differential plug-in of software and/or Wizard resources,
• enable, in future steps, observation and capture of speakers activity, multimodal and speech-relevant linguistic events, or behavioral factors,
• eventually develop some automated aids to improved annotation of multimodal speech corpora.

A network-based platform, with an Intranet as a start up. We want the system to naturally support real-life bilingual dialogues, possibly between far distant speakers. Hence communication and control engines should be Internet-supported, albeit being first implemented on local networks. We are currently collecting as well through network-driven audio-conferencing tools. However on alternative LAN-based settings, in order to slightly speed up the exchange rate during collecting sessions, speech can of course be sent locally through wired audio lines, between close rooms. Speech transfer should be network-driven indeed for distant conversing.

Linguistic processes to be conducted on the simulator. We started using the SIM*/1 platform mainly for collecting task-oriented bilingual corpora of raw spontaneous speech, between monolingual native not very distant yet speakers, presently on Chinese-French dialogues. Based on the shell of this stepwise project, we intend to study an assistance to corpus annotation, and how to enrich primary transcriptions or raw automatic annotations. Another coming process in data acquisition will be the instant capture of multimodal events to be resynchronized with translated utterances.

1.3 A generic architecture

Multilingual genericity on the simulator roots in the architecture itself, around one or several Wizard of Oz interpreters (cf. Figure 1). Functional genericity is required as well, that is to provide simple reconfiguration of basic functions (e.g., recognition, translation, synthesis). On request we then could choose, for any function, either human-driven (Wizard of Oz) or software processing, and possibly test alternative lingware functional components. Lastly the environment should be an open multiplatform, operating on heterogeneous configurations (PC, Mac, Unix).

2. THE SIM*/1 PLATFORM: FUNCTIONAL AND TECHNICAL VIEWS, CURRENT USE AND DEVELOPMENT

2.1 Salient functionalities

On the first platform. We favoured a local network implementation, with two speakers located in possibly distant rooms in the laboratory. The current SIM*/1 environment carries out (cf. Figure 2):

- a bilingual Wizard of Oz interpreting resource, for simulating future Speech MT of multilingual dialogues,
- a communication engine based on a client-server scheme, in charge for speech turns sequencing and dialogue regulation using a preemptive push-to-talk,
- the creation, in the collect-oriented mode, of bilingual session bases with speech and textual events, with dating and identification attributes for every utterance (for each speaker, for the interpreter).

On the current implementation both Communication Server and Wizard Interpreter processes may run on a same station.

![Figure 2: Current version of the SIM* prototype, as being validated for collecting Chinese-French dialogues.](image2)

On a SIM*/2 variant. Multimodal features are introduced: interactive marking on a shared local whiteboard, proper nouns spelling and text editing, user-driven file transfer, on line display of Web information (to complement or enliven dialogues), and some video-conferencing (cf. Figure 3).

Automatic speech recognition will be integrated first on the Interpreter workstation, displaying recognized utterances for confirmation or spoken correction. A trained Wizard Interpreter is expected to sound clean and clear. Speech recognition for speakers will then be completed, after enough domain-oriented data may be bootstrapped in the process.

Recent development has explored collecting-oriented variants (for monolingual dialogues, or with bilingual speakers while by-passing the translation process), or 3-speaker situations (e.g., 1 inquiring Client and 2 answering Agents) paving the way for the study of specific aspects in the translation of multiple dialogues.
2.2 Technical features

SIM*/1 currently runs on different platforms (PC, Apple Macintosh, some versions on Unix stations). Basically TCL-TK resources for multiplatform development and MBone or video-conferencing applications are used to implement Client-Server servicing and data recording, as well as speakers and interpreter interfaces, including local whiteboard and some video-conferencing facility. Speech files are in standard PCM 16kHz-16bit-mono format.

2.3 First corpus collection

The current SIM*/1 platform was initially tested in collecting monolingual dialogues in the context of the C-STAR II project, then to gather raw speech corpora in French-English task-oriented small dialogues (travel and hotel information and reservation, additional tourism information for business trips), close to C-STAR II scenarios.

It is presently used for collecting bilingual speech corpora in realistic spontaneous task-oriented translated French-Chinese dialogues between monolingual speakers, in the same domain (hotel reservation, then travel information), in cooperation with the Natural Language Processing Research group (Academy of Sciences of China, Beijing).

Up to now a first course of some 4 hours of raw spoken dialogues was recorded locally, on the SIM*/1 platform with a Wizard of Oz translation through our intranet. No annotation resource is available yet. Internet-based SIM*/2 version for distant collecting is under development. During a session, sound and identification descriptor files are recorded locally on speakers and interpreter stations, to be centralized afterwards towards a session file base.

2.4 Towards assistance to transcription and annotation

Lastly, on a SIM*/3 version, we intend to develop aids (based on generic multilevel dependency parsers) to produce some kind of quality transcripts or enhanced annotated corpora, that is correcting primary automatic transcripts of multimodal multilingual dialogues, and supplementing them in a way at semantic and/or pragmatic levels.

3. SIM*/4: AN ASSISTANCE TO NET-BASED HUMAN INTERPRETERS

From the modeling of the primary platform, parallel development towards a SIM*/4 translingual human communication platform is currently under way. We explore a symmetrical paradigm, where the system behaves as a linguistic assistant through the net: it may provide distant speakers, but especially a human interpreter who intermittently operates between them through the Internet, with some automatic lexical or terminological aids derived from SMT active functions.

3.1 Situation

In place of a hidden Wizard, we picture here the interpreter as a ‘warm body’, to be possibly assisted by MT resources. Given the same basic communication engine, the situation is as follows: this professional or occasional interpreter, working home or from an office, will translate in intermittent sessions through the net. A more precise potential situation could be as follows: this interpreter is called on through joint request by two speakers (distant from each other or nearby, or by a video-conference group) who otherwise manage to understand each other, possibly in a common language. They wish to temporarily ask the interpreter for assistance, during sensitive or critical parts of their dialogue. On demand, their previous conversation will have been monitored and processed for a while by some domain-focusing watchful assistant module. Before intervening the interpreter, if willing to, may then consult instant lexical support or a dialog story, automatically derived by active MT resources from the follow-up of speakers’ dialog under way.

3.2 Aided interpreting on the Internet

Terminological backing. This SMT-driven assistant will for instance forward to the interpreter a short dialog history, with specific or difficult terms most often used by the speakers (through theme-tracking), with their translations if requested, and some detection of the dialogue type or phase.

Asynchronous aids. The consulting of asynchronous or semi-synchronous aids by such supporting human interpreters should be for them fully optional and strongly regulated by their pre-specified preferences, in order to avoid unbearable perceptive or cognitive load.

Lexical support to speakers. Besides, some multilingual terminological assistance can be similarly proposed to speakers as well, based on available corpora, with the same concern for very flexible and dynamic settings of this backing.

4. PROSPECTS

The SIM*/1 primary platform progressively meets some main early requirements. First use gives evidence of its potential versatility. It needs be enhanced according to each different scope of interest: more technical assessment for feasibility, performance on Internet-based real-size real-time distant dialogues, on data acquisition, on converser
observation with different hypotheses on multimodal interface.

Further work towards SIM* coming versions includes, from short- to long-range developing:

- Long-distance dialogues: development will soon bring remote connections over the net in order to collect speech in very realistic situations. Transition to Internet-based dialogues easily derives from the current implementation. Up to now, some drawbacks in distant live speech transmission on the net (or possible uncertain reliability through uncostly regular connection), and the high cost of ISDN connections, may still require for a while parallel telephone transmission of the dialogues —even if locally recorded speech segment files may be transmitted in a row, at a slower pace. Hopefully the situation here is evolving rapidly.
- A more generic communication engine, inspired by the ComSwitch architecture (particularly appropriate for strongly distributed development), and therefore highly compatible with related platforms (C-STAR III, NESPOLE).
- Extending plug-in facilities, with an efficient start-up configuration resource.
- Extending multimodality, and providing some more acute synchronization between oral translation and multimodal events (pointing, marking, spelling…).
- Investigating both interpreter, or speaker, instant or asynchronous multimodal linguistic aids.
- Integrating specialized ‘observer Wizard(s)’; for capturing multimodal events, prosodic or speech-relevant linguistic events (in order also to trace particular features relevant to multilingual dialogues, e.g. anaphorae in multipprinter situations).
- Experimenting differentially with MT automatic processing and with Wizard handling, on various dialogue sub-domains.
- Producing instant on-line transcription and annotation means (possibly involving extra dedicated Wizard stations).
- Expanding multilingualism: due to the architecture, generalization to the simulation of multilingual interpreting would follow quite easily, beyond logistic and session protocol aspects to be settled.

CONCLUSION

As the recent development of multilingual Speech Machine Translation technology triggers a need for methodically building large specialized or general annotated speech corpora, we believe the Internet to be a stimulating and possibly efficient vehicle for collecting spontaneous multilingual oral dialogues. We presented here the first achievements and work in progress on SIM*, an open multi-usage and net-based simulation and collection environment, currently experimentally used for bilingual speech acquisition.

There is a growing interest as well in observing and modeling speakers behaviour or expectation, in the situation of multimodal multilingual machine-aided dialogues over large networks, whereas multimodal portable teleservice devices currently appear. Wizard of Oz technology proves to be efficient for such observatory investigation, in order to progress along the line of ‘human-aided’ MT. Conversely, as we may envision new involvement for professional or occasional human interpreters while operating through the Internet, we set out to devise automatic MT-based aids for backing such human interpreting activities.

We expect this open environment to eventually contribute to explore novel types of Internet- or LAN-based multilingual communication applications.

ACKNOWLEDGEMENTS

 Reported work is backed by UIF (University Grenoble 1) and INPG, and granted in part by Rhône-Alpes Region (ERIM, Emergence 99’) and by LIAMA (Beijing-Paris). The authors are indebted to Laurent AUBLET-CUVELIER from MULTICOM at CLIPS for contributing to early specification and experimentation, and thankful to many members of the GETA/GEOD teams for their valuable and willing experimenting on the simulator prototypes.

REFERENCES