Speech translation in pedagogical environment using additional sources of knowledge

J. Tomás (1), A. Canovas (1), J. Lloret (1), M. García Pineda (1), J.L. Abad (2)

(1) Universidad Politcnica de Valencia,
Inst. de Inv. Gestión Integrada de Zonas Costeras, 46730 Gandia, Spain.
Hewlett-Packard,
(2) Av. Graells 501, 08174 Sant Cugat del Valles, Spain.
{jtomas,alcasol,jlloret,migarpi}@upv.es, abad@hp.com

Abstract

A key aspect in the development of statistical translators is the synergic combination of different sources of knowledge. This work describes the effect and implications that would have adding additional other-than-voice information in a voice translation system. In the model discussed the additional information serves as the bases for the log-linear combination of several statistical models. A prototype that implements a real-time speech translation system from Spanish to English that is adapted to specific teaching-related environments is presented. In the scenario of analysis a teacher as speaker giving an educational class could use a real time translation system with foreign students. The teacher could add slides or class notes as additional reference to the voice translation system. Should notes be already translated into the destination language the system could have even more accuracy. We present the theoretical framework of the problem, summarize the overall architecture of the system, show how the system is enhanced with capabilities related to capturing the additional information; and finally present the initial performance results.

1. Introduction

The development of automatic real-time translation systems from voice signals constitutes a long term objective. However, recent advances in the field of statistical translation increase the possibility of an actual widespread usage in the near future (Loof et al., 2007; Casacuberta et al., 2004).

An ever-more increasing number of foreign students, interchange programs and alike in Europe (reaching a 15% in Gandia Polytechnic University School) is requiring more effort to provide tools and means that help the integration in the learning process while the new language skills are getting developed. A tool like the one presented in this article could increase the learning rate provided by the spoken classes.

We hence provide a prototype that demonstrates the viability of the real-time speech translation in the pedagogical student-teacher environment. Given the fact that current status-of-the-art products and techniques in the area of automatic real-time translation are far from perfect, we enhance the results by providing beforehand material about the elements of translation, e.g., specific vocabulary, texts, etc. With this purpose our speech translation system is fed with slides and class notes previous to the operation of the system. Often these sources or information are already translated and handed over to the student. We will make use of the offline translation as input to the system as well.

In this paper we explain also how to adapt an existing real-time speech translation system to incorporate the usage of the above mentioned additional information, and how this impacts positively in the accuracy results of the translation.

2. Prototype Description

The prototype tested for our model implements a real-time speech translation system to support a Spanish classroom with English-speaking students.

The teacher provides the slides in a MS PowerPoint format making sure that the notes area for each slide contains an explanation for the slide itself. The closest is the best to the speech that would be used when describing the actual slide.

Previous running the class the teacher must load the PowerPoint file in the system. At this point the system gets adapted to reflect the text within the slides, increasing the probability of the corresponding words to appear at the time when the slide is presented, and all along the duration of the class.

When the teacher is speaking, the system recognizes the sentence, translates it, and is able to display the subtitle translation as caption to the slide (see an example in Figure 1). In this way, an English-speaking student is able to relate the content, with the Spanish representation through the displayed translation.

3. Statistical Spoken Language Translation

The goal of Statistical Spoken Language Translation (Brown et al., 1994; Amengual et al., 2000) is to translate a given acoustic observation vector $x_1^T = x_1...x_T$ into a target sentence $t_1^I = t_1...t_I$. The methodology used (Brown et al., 1993; Tomás et al., 2005) is based on the definition of a function $Pr(t_1^I|x_1^T)$ that returns the probability that t_1^I is a translation of a given acoustic observation. We can introduce a hidden variable that represent the source sentence, $s_1^J = s_1...s_J$. Then, we can write:

Figure 1: Example as seen through demonstration The last two lines are superimposed to the projection of the slide. In blue appears what the teacher said; while black is the corresponding translation.

$$\begin{aligned} \hat{t_1^{\hat{I}}} = \underset{t_1^I}{\text{argmax}}\, Pr(t_1^I|x_1^T) = \\ \underset{t_1^I}{\text{argmax}} \sum_{s_1^J} Pr(s_1^J, t_1^I|x_1^T) = \\ \underset{t_1^I}{\text{argmax}} \sum_{s_1^J} Pr(s_1^J|x_1^T) Pr(t_1^I|s_1^J) \simeq \\ \underset{t_1^I, s_1^J}{\text{argmax}}\, Pr(s_1^J|x_1^T) Pr(t_1^I|s_1^J) \end{aligned} \quad (1)$$

Following the log-linear approach (Och and Ney, 2002)(Tomas et al., 2007), $Pr(t_1^I|s_1^J)$ can be expressed as a combination of a series of feature functions, $h_m(t_1^I, s_1^J)$, that are calibrated by scaling factors, λ_m:

$$Pr(t_1^I|s_1^J) = \sum_{m=1}^{M} \lambda_m h_m(t_1^I, s_1^J) \quad (2)$$

This framework allows us a simple integration of several models in the translation system. Moreover, scaling factors allow us to adjust the relative importance of each model. For this objective, Och and Ney propose a minimum error rate criterion (Och and Ney, 2002).

4. Architecture

Our system architecture is based in two modules:
The **speech recognition module** (SRM) gets audio input stream from a microphone and obtains an N-best output text. Each hypothesis in the N-best list is scored according to the equation $Pr(s_1^J|x_1^T)$.
Although there are several open source speech recognition systems, like Sphinx or HTK available, we have used the standard system provided by the MS Windows Vista OS, as it happens to be the only one incorporating acoustic models for Spanish that are available to non-restricted tasks. The communication with this engine is based on the SAPI interface (Hao Shi, 2006).
In addition this engine has a few interesting capabilities that make it well-suited for a real-time application like ours, for example customizing its functionality for a specific speaker and task which allows us to work with multiple output hypothesis simultaneously.
The **machine translation module** (MTM) is based on previous work described in (Tomás et al., 2006). Basically, in order to estimate $Pr(t_1^I|s_1^J)$ a log-linear combination of several statistical models is used. In our application two models are needed, one for the translation itself and one for the target language selected. A new important feature is introduced in this work, the output score of the speech recognition module.
Therefore the machine translation module integrates the following knowledge:

- Translation model: based on monotone phrase-base models. Phrase-base models divide the sentence in segments each composed of a series of words. The translation probabilities now relate a sequence of words in a source sentence with another sequence of words in the target sentence. The simplest and farter formulation with such models is based on monotone models (Tomás et al., 2006). However to operate in real time the speed of the translation search is a critical factor. Thus, we have to select a monotone phrase-based model.

- Target language model: is comprised of two sub-models, a conventional trigram model: $p(t_i|t_{i-2}^{i-1})$ and a five-gram class model: $p(T_i|T_{i-4}^{i-1})$.

- Speech recognition score: that is simply the output of the speech recognition module.

4.1. Adaptation

If available we make use of additional information, as text, closely related to the task we go to translate. This text can be written in the source language, in the target language or in both.
The **SRM adaptation** is performed via the SAPI interface (Hao Shi, 2006). SAPI adaptation uses specific calls to the SAPI interface. Specifically, we extract each word from source adaptation data and use it with the SAPI call AddVocabulary to extend the SRM vocabulary.

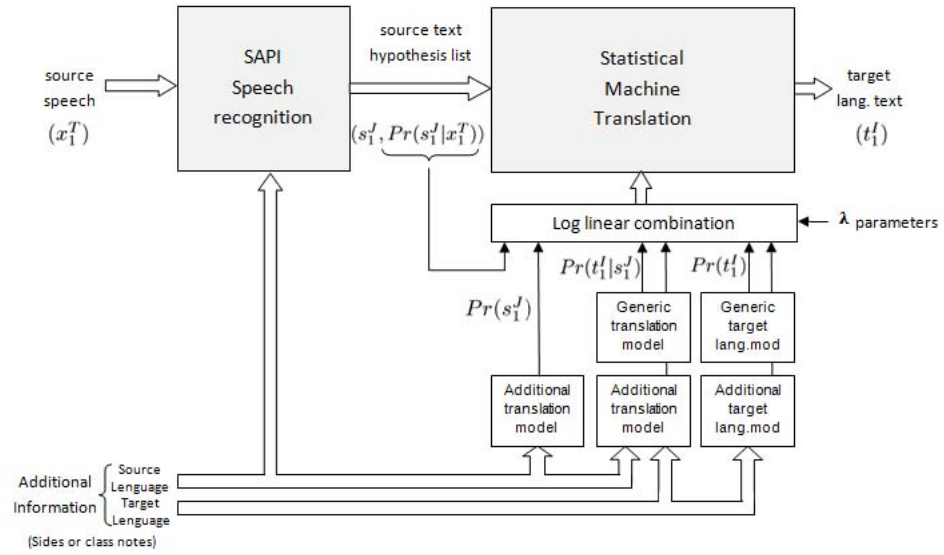

Figure 2: System architecture.

The **MTM adaptation** is hence performed as follows: using the source language text we first train an additional source language model, and using the target language text we train a second additional target language model. Finally, using both source and target text we train an additional third translation model.
These three new models are then incorporated to the system using the log-linear framework. In this framework each model needs a scaling factor parameter that is estimated by using a minimum error rate criterion (Och and Ney, 2002). A development corpus is needed for this propose.

5. System Evaluation

The system introduced in this work was assessed through a series of experiments stressing the system in different situations and using three different speakers. Experiments have run in a scenario that reproduces the normal conditions of a university class.
In the scenario, a teacher provided a 20 minutes class supported with projected slides and class notes which where beforehand translated into Spanish and English. The class was recorded in an empty room without students for the sake of comparing output results with the same background noise conditions. Sentences from the recording in Spanish were then segmented, transcribed and translated into English.

	spontaneous speech	speaker adaptation	genere
speaker 1	Yes	no	male
speaker 2	No	yes	male
speaker 3	No	no	female

Table 1: Qualities of tree speakers.

The sentences obtained where divided into two parts: the test corpus (made of 240 sentences) and the development corpus (with 120 sentences). Sentences from the test corpus were also recorded later by two additional speakers. Table 1 represents the different quality features for each speaker.

	Speech Recognition (WER)
speaker 1	23.6
speaker 2	17.5
speaker 3	27.0

Table 2: Speech recognition performance for tree test speakers.

The generic models of MTM were initially trained by the Europarl corpus, (Koehn, 2005) and the slides and class notes have been used to train the specific other models of MTM. The developed corpus was used to estimate the lambda parameters based on a simple minimum error rate criteria.

	Speech Recognition (WER)	Machine Translation (WER)	(BLEU)
base line	17.5	54.2	34.8
+ SAPI adaptation	16.5	53.8	35.1
+ source slides	15.3	53.3	35.6
+ target slides	15.4	42.1	45.7
+ source class notes	9.7	40.1	48.4
+ target class notes	9.7	35.0	56.4

Table 3: Performance achieved with different adaptation sources for speaker 2.

In Table 2, speech recognition performance is compared for

the tree test speakers. It is evident that the speaker adaptation capabilities are crucial to obtain good speech recognition rates.

In Table 3, different adaptation mechanisms have been compared. As base line there was no adaptation used. In the next experiments we have added additional information cumulatively (+). The SAPI adaptation mechanism uses specific calls to the SAPI interface. Specifically, we extract each word from the source slides and the class notes to extend the SRM vocabulary. To evaluate the MTM adaptation we have considered two sources of knowledge: using just the slides or the slides with class notes; and in each case with the combination of just using the source language, and using both source and target language. Signficant improvement is observed when class notes are introduced, because in this experiment class notes are very similar to teacher speach.

6. Conclusions

A real-time speech translation system specific to pedagogical environments has been presented. The main innovation of this work is the way in which additional sources of knowledge are used to improve the accuracy of the system, while remaining practical.

The experiments demonstrate how this usage of additional information source is really improving significantly the overall results by a 12%. Especially if we can make use of class notes in both languages previous the real-time operation. In this case the accuracy rate increases by a 35%. Providing translations is obviously an extra effort for teachers, but is often worth doing it when the number of foreign students in the class is high.

Training the system with other textual sources of information that are also related to the class could also help, even if they are not the exact notes to the slides, but provided the texts refer to the same concepts developed, for example reference books in the material. In this case translations of these texts are often available in different languages, and training the system with such pre-existing material could also help.

As future work we will improve the system by using models of confusion networks as interfaces between the automatic speech recognition and machine translation modules.

Acknowledgments

This work has been partially supported by the Generalitat Valenciana and the Universidad Politecnica de Valencia.

7. References

J.C. Amengual, J.M. Benedí, F. Casacuberta, M.A. Castao, A. Castellanos, V.M. Jiménez, D. Llorens, A. Marzal, M. Pastor, F. Prat, E. Vidal, and J.M. Vilar. 2000. The EuTrans-I speech translation system. *Machine Translation*, 1.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

P.F. Brown, S.F. Chen, V.J. Della Pietra, S.A. Della Pietra, A.S. Keller, and R.L. Mercer. 1994. Automatic speech recognition in machine translation. *Computer Speech and Language*, 8(1):177–187, April.

F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. Garca-Varea, D. Llorens, C. Martnez, S. Molau, F. Nevado, M. Pastor, D. Picó, A. Sanchis, and C. Tillmann. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 18:25–47, January.

Alexander Maier Hao Shi. 2006. Speech-enabled windows application using microsoft sapi. *International Journal of Computer Science and Network Security*, 6(9):33–37.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*, Phuket, Thailand, September 12 to 16.

Jonas Loof, Christian Gollan, Stefan Hahn, Georg Heigold, Bjorn Hoffmeister, Christian Plahl, David Rybach, Ralf Schluter, and Hermann Ney. 2007. The rwth 2007 tc-star evaluation system for european english and spanish. In *Interspeech*, pages 2145–2148, Antwerp, Belgium, August.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, July.

J. Tomás, J. Lloret, and F. Casacuberta. 2005. Phrase-based alignment models for statistical machine translation. In *Pattern Recognition and Image Analysis*, volume 3523 of *Lecture Notes in Computer Science*, pages 605–613. Springer-Verlag.

J. Tomás, J.M. Vilar, and F. Casacuberta. 2006. The ITI statistical machine translation system. In *Proceedings of the TC-Star Speech to Speech Translation Workshop*, pages 49–55, Barcelona, Spain, June 19-21.

J. Tomas, J. Lloret, and F. Casacuberta. 2007. Phrase-based statistical machine translation using approximate matching. In *3rd Iberian Conference on Pattern Recognition and Image Analysis*. Girona (Spain), June.