Evaluation Frameworks for Speech Translation Technologies

Marcello Federico

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo - Trento, ITALY.
federico@itc.it

Abstract

This paper reports on activities carried out under the European project PF-STAR and within the CSTAR consortium, which aim at evaluating speech translation technologies. In PF-STAR, speech translation baselines developed by the partners and off-the-shelf commercial systems will be compared systematically on several language pairs and application scenarios. In CSTAR, evaluation campaigns will be organized, on a regular basis, to compare research baselines developed by the members of the consortium. The first evaluation campaign, which will take place in 2003, will focus on written language translation by exploiting a large phrase-book parallel corpus covering several European and Asiatic languages.

1. Introduction

Speech translation is one of the leading-edge technologies enabling multilingual human-human and human-machine interaction. Despite the significant progress made in the last years, performance of text and speech translation technology is still far from being satisfactory.

In the following two sections, I will provide details about the evaluation framework developed in PF-STAR, and give a brief overview about the first CSTAR evaluation campaign.

2. PF-STAR Project

The project PF-STAR intends to contribute to establish future activities in the field of multisensorial and multilingual communication (interface technologies) on firmer bases by providing technological baselines, comparative evaluations, and assessment of prospects of core technologies, which future research and development efforts can build from. To this end, the project focuses on three crucial areas: technologies for speech-to-speech translation, the detection and expressions of emotional states, and core speech technologies for children.

Speech translation technologies are addressed by three partners: ITC-irst (co-ordinator), RWTH Computer Science Department, and Interactive Systems Labs at Universität Karlsruhe (UKA). Objectives will be the comparative evaluation and integration of different speech translation baselines over a range of application scenarios and languages.

2.1. State of the art

Subject of the comparative analysis will be speech and text translation approaches which the partners have been developed during the last years. In the following, a short presentation of each approach is given.

Interlingua based translation. This approach is based on the assumption that the content of task oriented conversations can be well approximated and automatically mapped into a relatively simple compositional formal language, called interlingua. Hence, content based translation from the interlingua to any language can be carried out by developing a suitable natural language generation modules. The pros of the approach can be summarized as follows. Translation can be decoupled into apparently simpler problems, i.e. the analysis and the synthesis steps, and the same interlingua can be used as a pivot for many languages. The cons are in the difficulty to develop an interlingua which should be sufficiently com-
plete and consistent, to cover all possible expressions in a given domain, and which should be, at the same time, easy to generate and interpret automatically. Last but not least is the inevitable loss of information induced by the interlingua representation.

**Statistical direct translation.** This approach tries to directly estimate a statistical model that translates all the single words of the source language sentence and finds out their optimal position in the target language sentence. A family of such models has been developed, starting from five basic models proposed by IBM in the middle 1990s[1]. Pros of the statistical models are in their performance and in the fact that they do not require any knowledge about the domain or task, but a significant sample of translation examples. Cons seem to be the amount of required training data, with respect to the task complexity, and the computational complexity of the best performing translation algorithms.

**Dictionary look-up direct translation.** Recent work on cross-language information retrieval [2] has shown that simple statistical models for dictionary based translation outperform in this task more sophisticated machine translation systems (e.g. Systran). This approach can be seen as a simplification of the IBM approach as no word alignment is taken into account by the model. Indeed the model seems effective in disambiguating translation of content words, by taking into account word co-occurrences in the target language. Pros are that training the model just requires comparable monolingual corpora and a bilingual dictionary, and that performance is expected to be stable across different domains and enough effective to correctly transfer content words across languages. Cons are the intrinsically low readability of the generated translations, which lack of syntax.

Each of the above approaches has some strong points which could be useful to develop better speech translation models, or some new technology for multilingual communication or information access. For instance, the interlingua approach offers interesting possibilities to model the broad topic and arguments conveyed by an utterance. This information could either be exploited to constrain the huge search space of a statistical direct translation model or to produce a rough content description or summary of a conversation. Finally, dictionary based translation could be applied in very complex domains, such as broadcast news, as a back-up translation modality when the full translation fails, e.g. due to bad lexical coverage.

### 2.2. Work-plan

Three types of technological baselines will be evaluated by the project partners: (i) interlingua based translation systems, developed at ITC-irst and UKA; (ii) direct translation baselines mainly based on statistical models, developed at ITC-irst, RWTH, and UKA; (iii) off the shelf commercial systems, e.g. Systran.

As concerns application scenarios, two experimental frameworks have been identified, which partners of the project have been investigated in the recent years, and for which suitable baselines are either available or can be developed in reasonable time. The two application scenarios are:

- human-to-human translation in the tourism/traveling domain,
- cross-language spoken document retrieval (CLSDR) in the news domain.

The first scenario has been investigated in the German national project Verbmobil[3], by the CSTAR consortium, and the EU project Nespole![4]. The CLSDR scenario has been recently developed by ITC-irst under the Cross-Language Evaluation Forum[5] (CLEF) umbrella. Relevant facts about both scenarios are summarized in Table 1.

As concerns evaluation criteria, the project partners will make use of objective measures, based on automatic alignment metrics - e.g. the BLEU score [4]-, subjective evaluation measures, based on different grading criteria, and task completion metrics, i.e. precision/recall in CLSDR.

### 2.3. Constraints

In order to define an evaluation plan for PF-STAR, constraints posed by the considered translation approaches have been taken into account. Constraints are mainly related to the suitability of the approaches to different evaluation criteria and to their portability across domains and languages.

- The interlingua approach followed by the partners, henceforth indicated by IF (interchange format), aims at transferring elements of meaning relevant

<table>
<thead>
<tr>
<th>Human-to-human interaction (HHI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Domain:</strong> Tourism/traveling</td>
</tr>
<tr>
<td><strong>Languages:</strong> English, Italian, German, French, Chinese, and Japanese</td>
</tr>
<tr>
<td><strong>Benchmarks:</strong> Verbmobil, CSTAR, and Nespole!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cross-language spoken document retrieval (CLSDR)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Domain:</strong> Broadcast news</td>
</tr>
<tr>
<td><strong>Languages:</strong> English, Italian, German, French, Spanish</td>
</tr>
<tr>
<td><strong>Benchmarks:</strong> CLEF evaluation campaigns</td>
</tr>
</tbody>
</table>

Table 1: Application Scenarios in PF-STAR

---

3http://verbmobil.dfki.de/verbmobil/
4http://nespole.itc.it
5http://clef.iei.pi.cnr.it:2002/
to the communication rather than producing a literal translation. The past experience suggests that this approach can only be applied on limited domains and requires significant human effort to define an appropriate interlingua formalism and to annotate training data. On the opposite, statistical translation models can be trained on larger domains, they need little human effort, but typically require large data samples to be properly estimated, e.g. a large parallel corpus.

- For the sake of a comparative evaluation, IF-based translation seems not well suited to automatic scoring. As its goal is to convey elements of meaning rather than producing a complete translation of the input, it can be difficult to define appropriate target translations for objective evaluation.

Given the above constraints and the available language resources, a check list for comparative experiments has been designed, which is reported in Table 2. Briefly, a comparison between IF, direct, and commercial translation systems will be performed on the tourism/traveling domain, in three languages (English, German, and Italian), by using subjective and automatic evaluation metrics. As a benchmark, data collected during the EU project Nespole! will be used. Moreover, a comparison between direct and commercial translation systems will be carried out on the broadcast news domain, with four translation directions: French to English, German to English, Italian to English, and Spanish to English. Evaluation in this domain will be carried out within the cross-language spoken document retrieval framework. The adopted benchmark will be the test-suite developed by ITC-irst for the CLSDR track at CLEF 2003.

### 2.4. Specifications

Specifications of the experiments in terms of training data, testing data, evaluation protocol, primary and secondary conditions are summarized in Table 3. Primary conditions directly address the main objective of the comparison, while secondary conditions aim at gathering additional information about strong and weak points of each approach. Even if not stated explicitly, these evaluations also aim at comparing different settings and implementations of baselines based on the same approaches, which may significantly differ in terms of statistical models, search algorithms, etc.

### 3. CSTAR Consortium

The CSTAR consortium has been active over the last decade with the purpose of investigating robust systems for multilingual speech communication. Currently, the consortium includes the following organizations: ATR (Japan), CLIPS (France), CMU (Usa), ITC-irst (Italy), NLPR (China), and UKA (Germany).

In a recent internal workshop, the consortium decided to organize, on a regular basis, speech translation evaluation campaigns and workshops, mainly focusing on speech translation research and evaluation. Moreover, activities within CSTAR will address the development of a large multilingual parallel corpus to be shared among the partners and used for common evaluations.

The first evaluation campaign and workshop will be held in June 2003 and September 2003, respectively. Both events will be restricted to CSTAR members only, and the evaluation will be limited to written texts. In particular, training and testing data will be based on the multilingual Basic Travel Expression Corpus (BTEC) developed at ATR [3] and extended by each partner to its respective language. It is foreseen that future evaluation campaigns will be open to external participants, too.

### 3.1. Specifications

The first evaluation campaign will concentrate on assessing text translation algorithms on the tourism domain. Translation directions will be from Chinese, Italian, Japanese, and Korean into English, for the primary condition, and vice versa, for the secondary condition. Specifications of this evaluation are summarized in Table 4.

Training data will consist of a fixed amount of English sentences provided with translations into the respective source language. Participants will be allowed to use

---

### Table 2: PF-STAR comparative evaluations.

<table>
<thead>
<tr>
<th>Baselines</th>
<th>HHI</th>
<th>CLSDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interlingua</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Direct</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Products</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Languages</th>
<th>HHI</th>
<th>CLSDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>French</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>German</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Italian</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Spanish</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metrics</th>
<th>HHI</th>
<th>CLSDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjective</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Alignment</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Task</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Benchmarks</th>
<th>HHI</th>
<th>CLSDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nespole!</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CLEF</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

---

6http://munst.itc.it/clef-sdr.html

7http://cstar.atr.co.jp/cstar-corpus/index.htm
HHI
Training data: parallel corpora, bilingual word lists
Test data: 4 dialogues from Nespole!
Languages: Italian-English, German-English
Primary cond.: end-to-end translation
Secondary cond.: (a) transcript-to-IF translation, (b) English as an interlingua
Evaluation: subjective on primary cond. automatic on s.c. (a) automatic on s.c. (b)

CLSDR
Training data: SDR TREC data, bilingual dictionaries, monolingual corpora
Test data: 100 queries (for the moment written)
Languages: from French, German, Italian, Spanish to English
Primary cond.: end-to-end
Secondary cond.: end-to-end using English as an interlingua
Evaluation: mean-average precision

Table 3: Specifications of the PF-STAR evaluation plan.

CSTAR
Training data: 50K-150K parallel sentences
Test data: 500 written sentences
Languages: Chinese, Korean, Italian, Japanese, English
Primary cond.: from X to English
Secondary cond.: from English to X
Evaluation: autom. & subj. on primary cond. autom. on secondary cond.

Table 4: Specifications of the CSTAR evaluation plan.

3.2. Evaluation Protocol

In the first evaluation campaign, automatic scoring will be carried out with the NIST/BLEU software. In particular, a central server will be set-up which will permit participants to remotely score the output of their system. For each translation direction, multiple human translations will be used as references.

Subjective evaluation on the primary condition will be distributed across the participant sites. English native speakers will evaluate the output of each systems against one gold-standard reference. Evaluation will follow guidelines similar to those applied by LDC in the NIST MT evaluation campaigns.

While automatic evaluation will be applied to all submitted runs, subjective evaluation will be applied to only one run per participant, namely the first run submitted under the primary condition.

4. Acknowledgements

This work has been partially funded by the EU project PF-STAR (IST-2001-37599). The authors would like to thank E. Pianta (ITC-irst), M. Paul (ATR, Japan) and A. Tribble (CMU, Usa) for their useful comments and suggestions.

5. References


