Bilingual Aligned Corpora for Speech to Speech Translation for Spanish, English and Catalan

David Conejero, Alan Lounds, Carmen García-Mateo(*), Leandro Rodríguez-Liñares(*), Raquel Mochales, Asunción Moreno.

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona, Spain
{david, alraquel, asuncion}@talp.upc.edu, alan.lounds@upc.edu, carmen@gts.tsc.uvigo.es, leandro@uvigo.es
(*) Universidad de Vigo, Spain

Abstract

In the framework of the EU-funded Project LC-STAR, a set of Language Resources (LR) for all the Speech to Speech Translation components (Speech recognition, Machine Translation and Speech Synthesis) was developed. This paper deals with the development of bilingual corpora in Spanish, US English and Catalan. The corpora were obtained from spontaneous dialogues in one of these three languages which were translated to the other two languages. The paper describes the translation methodology, specific problems of translating spontaneous dialogues to be used for MT training, formats and the validation criteria.

1. Introduction

The EU-funded project LC-STAR aims to create lexica and corpora (LR) needed for transferring Speech-to-Speech Translation (SST) components, i.e. flexible vocabulary speech recognition, high quality text-to-speech synthesis and speech-centred translation into the selected languages. LC-STAR concentrates on large lexica with phonetic, prosodic and morphosyntactic content and on bilingual aligned text corpora. Different approaches to speech translation show that it is possible to develop robust speech-to-speech translation systems for small to medium-sized domains using sophisticated speech recognition and machine translation technology. The major problems in the field of speech-to-speech translation are:

- Acquisition of monolingual or bilingual, domain-specific training data.
- Robust behaviour of the MT component for speech recognition errors and spoken language phenomena.
- Development of efficient recognition and translation components, and a high quality text-to-speech synthesis system.

Currently, the most promising approach for speech-to-speech translation systems is to use statistical machine translation (SMT). This system is able to learn from example translations, shows some robustness against speech recognition errors and outperforms other speech-to-speech-translation systems (cf. Verbmobil).

One of the most powerful resources for statistical machine translation are multilingual corpora. For Speech-to-Speech Translation applications, LR need to cope with spontaneous speech. Spoken language does not follow closely-written text patterns. It contains colloquial forms, verb forms of first and second persons, conversational forms and several ungrammaticalities such as false starts, corrections, repetitions, incorrect syntactical structures, etcetera. Transcriptions obtained from an ASR system usually fail to provide accurate punctuation marks (pauses, exclamatory or interrogative sentences) and the translation performance drops dramatically.

In the LC-STAR project bilingual aligned corpora were generated from transcriptions of spontaneous speech, thus providing a source of learning typical problems of spontaneous speech. The data were taken from dialogues in a tourist domain. As is well known, at content level one can distinguish between comparable corpora, parallel corpora and aligned corpora. When alignments are possible, a hierarchical structure would be better in order to preserve and process the data aligned at different levels. The granularity of the alignment may vary from non-aligned or corpus-aligned to document-aligned or section/paragraph/sentence-aligned and even—though not very often—word aligned. In general, the finer the granularity of the alignment, the more the data are appreciated, but more human effort is required. For statistical machine translation purposes, at least a segment alignment is desirable, with each segment consisting of as few sentences as possible.

In the LC-STAR project, three bilingual corpora from spoken dialogue data were created: Spanish/US-English, Catalan/US-English, and Spanish/Catalan. They were aligned at turn level and sentence level.

This paper is organized as follows: Section 2 describes the specifications and the annotation and translation methodology, in Section 3 we discuss the specific problems arising from the specifications and the translation methodology oriented towards MT applications that was applied by professional translators. Section 4 deals with formats and describes the DTD used in this project. Finally, Section 5 shows the validation criteria and validation results carried out by the validation centre of the University of Vigo.

2. Specifications

A set of spontaneous spoken dialogues collected in English, Spanish and Catalan were annotated and translated to the other two languages to create the bilingual aligned corpora.

---

1 EU Project LC-STAR “Lexica and Corpora for Speech-to-Speech Translation Technologies” IST-2001-32216
2.1. Source corpora

The spontaneous dialogues were extracted from Verbmobil and TALP-tourism recordings. The production of these dialogues is described in [1].

To create the bilingual aligned corpora, some of these dialogues were selected. The criteria for selection were:
- Verbmobil corpus: only public material [2] was used.
- TALP-tourism recordings: the scenarios detailed in Table 1 were selected.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Dialogues</th>
<th>Turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Hotel</td>
<td>50</td>
<td>3,205</td>
</tr>
<tr>
<td>2 Asking about Services</td>
<td>36</td>
<td>1,543</td>
</tr>
<tr>
<td>3 Asking about Leisure Activities</td>
<td>19</td>
<td>718</td>
</tr>
<tr>
<td>4 Demanding a Service</td>
<td>61</td>
<td>1,734</td>
</tr>
<tr>
<td>2 Travel agency</td>
<td>78</td>
<td>5,775</td>
</tr>
<tr>
<td>1 Booking Tickets</td>
<td>9</td>
<td>590</td>
</tr>
<tr>
<td>2 Booking Hotel Rooms</td>
<td>25</td>
<td>2,297</td>
</tr>
<tr>
<td>3 Booking Trips (tickets, hotel + extras...)</td>
<td>25</td>
<td>2,297</td>
</tr>
</tbody>
</table>

Table 1: TALP-tourism recording scenarios, number of dialogues, and number of turns

The transcription of spoken dialogues [3] is orthographic, and it includes tags that represent audible acoustic events (speech and non-speech) that are present in the recorded dialogue. These tags were annotated because they were later to be needed during the translation task, in the sense that they could carry either some meaning or discourse information.

The selected corpora main features are summarized in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Verbmobil</th>
<th>TALP-tourism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogues</td>
<td>509</td>
<td>208</td>
</tr>
<tr>
<td>Turns</td>
<td>4,720</td>
<td>10,511</td>
</tr>
<tr>
<td>Sentences</td>
<td>10,763</td>
<td>23,333</td>
</tr>
<tr>
<td>Words</td>
<td>105,537</td>
<td>252,583</td>
</tr>
</tbody>
</table>

Table 2: Selected dialogues.

In order to check the distribution of words per dialogue in each corpus, an automatic tool was devised. The criteria applied by this tool are conservative, i.e. in case of doubt, words are not counted. These criteria are:
- The number of words per dialogue is obtained from the source language segments.
- Proper nouns and punctuation marks are not counted.
- Unintelligible words and truncated waveforms are not counted.

One histogram for each source language was built using the number of words per dialogue. These histograms are included in Figures 1 and 2 for Spanish and English respectively. In these figures it can be observed that when the source language is English (Verbmobil database), the dialogues are somewhat shorter than when the source language is Spanish (TALP tourism database). These differences are caused by language specific factors: Spanish and Catalan speakers tend to use longer and more complex sentences than English speakers. The figures for Catalan figures are quite similar to those for Spanish.

2.2. Spontaneous oral translation for machine learning

In some experiments within the LC-STAR project [4] it has been proven that automatic translation of languages that are similar in word order and morphology is more accurate than that of languages that differ in these aspects. It is reasonable to assume that translation resources should try to keep the translation as literal as possible.

As a result of the previous experiments, a translation style was designed for the translation process. The main goal was to keep the translations as literal as possible to the source text, while preserving their correctness. This can be summarized in the following golden rule:

“The target sentence must be as literal as possible to the source sentence in the sense of word-to-word translation and word order as long as it is a proper sentence likely to be ever uttered by a competent target language speaker”

For some specific translation issues the following general criteria were applied:
- Ambiguities: select the most plausible meaning considering the phrase within the tourist domain.
• Proper nouns: translate if they are commonly used or known in the Target Language (TL).
• Punctuation marks: simplify their use in the TL.
• Dates, hours and numbers: translate by concepts, not by words.
• Abbreviations: translate in their full orthographic form unless they are spoken in their abbreviated form and this form is meaningful in the TL.
• Acronyms: depends on their use in the TL.
• Foreign words: depends on their use in the TL.
• Spellings: the semantic meaning must remain in the TL.
• Neologisms and slang: translate to the TL if possible.
• Abbreviations: translate in their full orthographic form unless they are spoken in their abbreviated form and this form is meaningful in the TL.
• Acronyms: depends on their use in the TL.
• Dates, hours and numbers: translate by concepts, not by words.
• Punctuation marks: simplify their use in the TL.
• Proper nouns: translate if they are commonly used or known in the Target Language (TL).
• Dates, hours and numbers: translate by concepts, not by words.
• Abbreviations: translate in their full orthographic form unless they are spoken in their abbreviated form and this form is meaningful in the TL.
• Acronyms: depends on their use in the TL.
• Foreign words: depends on their use in the TL.
• Spellings: the semantic meaning must remain in the TL.
• Neologisms and slang: translate to the TL if possible.

As explained above, there are some spontaneous speech artefacts and noises annotated in the source corpora. The following rules were applied:
• False starts, verbal deletions, non-speech acoustic events and prosodic annotation are deleted in the TL.
• Word fragments, unintelligible words or truncated waveforms are translated if the meaning of the word is understandable. If not, they are marked.

3. Translation
The most important problem encountered in the translation was that the nature of the project made it necessary to apply criteria that went against standard translation practice. Translators normally aim to convey the meaning of the original in a way that is as natural as possible, so that the resulting text reads as if it had been written originally in the target language. In the creation of the present corpus, however, the aim was to create a match that was as close as possible between the original and the translation in order to facilitate the machine translation process. The orders to translators were thus to conserve the word order as far as possible, providing that the resulting sentence was a statement that conveyed the meaning and could possibly be uttered by a native speaker in a similar context. In other words, the best result was not necessarily the most natural translation. Needless to say, this was a tall order, particularly when translators have spent years training to do exactly the opposite.

An added complication was the sheer size of the corpus and the fact that the translators were working to a very tight deadline. This made it necessary to form a large team of freelance translators who were briefed at several meetings on how they should approach the work. In addition, because it was impossible to define all the criteria beforehand, the queries that emerged in the early stages were sent weekly to the project managers and the solutions were distributed to the whole team.

The translators had to deal with questions such as what to do with “ums” and “ers”, incoherent sentences, missing words, ungrammatical expressions (some of the dialogues involved non-native speakers, but even native speakers often use ungrammatical expressions), puns involving proper nouns and even offensive language. Finally, the rambling nature of spoken language meant that it was hard to resist the temptation to “tidy up” the text rather than to respect the word order.

4. Format
An XML-based mark-up language was chosen to represent the linguistic information in a formal, unambiguous and easy-to-read manner. To deal with XML structures, a DTD was defined.

It was intended to have one file for each original-target language pair. Therefore there are 6 different files (Spanish-Catalan, Spanish-English, Catalan-Spanish, Catalan-English, English-Catalan and English-Spanish) following the structure defined by the DTD.

4.1. Description
The DTD defines a type of document called a Bilingual Corpus. The Bilingual Corpus is divided into dialogs. A dialogue is a set of spoken turns between two speakers. It contains:
• Dialogue number (nsec)
• Dialogue scenario or subdomain (SCN)
• Turns
A turn is a sentence or a set of sentences spoken by one speaker without interruption by the other speaker. At DTD level each turn is defined by the original sentences, their translation and the alignment between the source and the target sentences. It contains:
• Turn number (nsegm).
• Speaker (SPKID).
• Turn orthographic transcription (Content).
• Source language segments.
• Target language segments.
• Alignment.

The source language segment refers to a sentence of the source data. It contains:
• The position of the language segment in the turn.
• The orthographic transcription.

The target language segment refers to a sentence of the target language data. It contains:
• The position of the language segment in the translation of the turn.
• The translation of the source sentence in the target language.

Alignment refers to the alignment of each source language segment and each target language segment. It contains:
• The position of the source language segment(s) (origin)
• The position of the corresponding target sentence(s) (translation)

In case of non-alignment of a given source (or translation) sentence, value 0 is used for the translation (or source) respectively.

4.2. XML example
Table 3 is an example of the xml final format of a dialog of only one sentence and its translation.

```xml
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE BILINGUAL_CORPORA SYSTEM "bil_corp.dtd">
<BILINGUAL_CORPORA description="LC-STAR bilingual corpora" domain="touristic" date="14/01/2005">
  <DOC sourcel="EN-US" targetl="ES-ES">
    <SEC nsec="001" SCN="11">
      <SGM nsegm="001" SPKID="0001">
        <SLSGM position="1">yesterday I found ~John_Andretti's~ car in $L $A</SLSGM>
        <SLSGM position="2">he was [sta] very happy</SLSGM>
      </SGM>
    </SEC>
  </DOC>
</BILINGUAL_CORPORA>
```
The total number of validated sentences was 370, and the trilingual (English, Catalan and Spanish) expert.

The selected fragments of the corpus were validated by a bilingual (English and Spanish) expert.

5.2. Results of the validation process

The imposed validation criterion [1] was that misalignment errors should be less than 5% and that the addition of naturalness, syntactic and semantic errors should be below 5% (4.32% was achieved). Thus, the corpus was considered valid.

Some additional remarks can be made about the translations and the validation process in order to illustrate that naturalness errors are difficult to define:

- In the process of validation, sometimes a choice has to be made between naturalness and closure to the structure of the source language. For example, to translate the Catalan “quina sort!” into English, “a real stroke of luck!” seemed to be more natural here, although “how lucky!” is still correct, and closer to the source structure.

- In some cases, lack of naturalness was perceived in the context of the dialogue. Without the context (i.e. taken in isolation) the utterances cease to be unnatural. To take an example, “let’s start” is a perfectly natural phrase in English and an accurate translation of the Spanish “comencemos”. However, “here we go” is preferred on the grounds of the context: the speaker is about to dictate a credit card number, and in this context “let’s start” seems unnaturally brisk.

5. Conclusions

Within the scope of the LC-STAR project 6 bilingual aligned corpora were produced. These corpora are disseminated via ELRA.

To represent the linguistic information within the corpora, an XML-based mark-up language was chosen because it is portable, has a hierarchical structure, and is easy to parse and check.

A DTD for bilingual aligned corpora was defined and successfully applied. It can be easily extended to deal with more information.

The spontaneous nature of speech does not allow for the definition of golden translation rules that can be applied in any circumstance, and would be equally applied by everyone. To minimize the impact of this factor, translation project managers should be involved from the early stages of the development of translation resources. They would help to define the translation and validation criteria, provide useful input on how the corpus should be obtained, and introduce elements to make the task of translation easier.

7. References


Table 3: Example of a dialog.

<table>
<thead>
<tr>
<th>5. Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>This section presents the validation process of the aligned bilingual corpora described so far. All validation checks were carried out by the validation centre following the specifications described in [5]. The validation process was divided into two parts:</td>
</tr>
<tr>
<td>- Checks considered as language-independent: completeness of the documentation and files, and compliance of the corpus to the generic rules for XML files and for the ones specified in the DTD.</td>
</tr>
<tr>
<td>- Checks considered as language-dependent: validation of the translation. This is presented next.</td>
</tr>
</tbody>
</table>

5.1. Validation of the translation

First, a subset of the corpora was selected for validation. This subset consisted of 3000 word-pairs chosen at random but with the following constraints: 500 words (50 words from short dialogues, 150 words from medium size dialogues and 300 words from long dialogues) were chosen from each of the three source languages (English, Spanish and Catalan). In randomly chosen sample dialogues, this represents well over 5% of each dialogue that will be checked. In total, the number of word-pairs to check was: 500 words per language x 3 languages x 2 translations = 3000 entries.

For each source language, three dialogues (short, medium and long) were randomly selected, and for each dialogue a starting turn was also randomly selected. From this starting turn, a fragment was chosen to accomplish the minimum number of words to be validated (50, 150 and 300).

The imposed validation criterion [1] was that misalignment errors should be less than 5% and that the addition of naturalness, syntactic and semantic errors should also be less than 5%. Errors were counted on sentence level.

5.2. Results of the validation process

The selected fragments of the corpus were validated by a trilingual (English, Catalan and Spanish) expert.

The total number of validated sentences was 370, and the following errors were found:

- Misalignment errors: 0 (0%)
- Naturalness errors: 2 (0.54%)
- Syntactic errors: 1 (0.27%)
- Semantic errors: 13 (3.51%)

Therefore, the validated fragments fulfill the rule that the alignment error must be below 5% (0% was achieved), and the naturalness, syntactic and semantic errors must be below 5% (4.32% was achieved). Thus, the corpus was considered valid.

Some additional remarks can be made about the translations and the validation process in order to illustrate that naturalness errors are difficult to define:

- In the process of validation, sometimes a choice has to be made between naturalness and closeness to the structure of the source language. For example, to translate the Catalan “quina sort!” into English, “a real stroke of luck!” seemed to be more natural here, although “how lucky!” is still correct, and closer to the source structure.

- In some cases, lack of naturalness was perceived in the context of the dialogue. Without the context (i.e. taken in isolation) the utterances cease to be unnatural. To take an example, “let’s start” is a perfectly natural phrase in English and an accurate translation of the Spanish “comencemos”. However, “here we go” is preferred on the grounds of the context: the speaker is about to dictate a credit card number, and in this context “let’s start” seems unnaturally brisk.