STELIOS PIPERIDIS[1,2]
IASON DEMIROS[1,2]
PROKOPIS PROKOPIDIS[1,2],
[1]**Institute for Language and Speech Processing,**
Artemidos 6 & Epidavrou, 151 25 Athens, Greece
[2]**National Technical University of Athens**
{spip, iason, prokopis}@ilsp.gr

Infrastructure for a multilingual subtitle generation system

Abstract

The expansion of digital television and the increasing demand to manipulate audiovisual content underlie the need for tools and systems that will automate the multilingual subtitle generation process. In this setting the MUSA project aims at providing a system which combines speech recognition, advanced text analysis, and machine translation to help generate multilingual subtitles. In its current version the system treats English as source and target language, as far as subtitle generation is concerned, French and Greek as subtitle translation target languages. In order to train and evaluate system components, an array of application specific resources is necessary. Primary audiovisual data consist in BBC TV documentaries. For each programme the set of multifaceted multimedia parallel data captured include: the actual video, its transcript or script, English, Greek and French subtitles, and topically relevant newspaper or web-sourced extracts.

1. Introduction

The changes to the audiovisual landscape are very rapid, linked simultaneously to the blossoming of mass communications to the evolution of technologies and the deregulation of the markets. New technological developments in mass media and communication, such as digital TV and DVD, are bound to overcome the limited physical borders of countries, leading to the creation of a world-wide media audience. In such a unified framework of mass communication, subtitling – as a means of overcoming linguistic barriers between the nations – is playing a critical role. In many countries, subtitling is the most commonly used method for conveying the content of foreign language dialogue to the audience; and a broadcaster's audience may now include several major linguistic groups (notably in the case of a satellite-broadcaster). Broadcasters, in order to meet the needs of the significant numbers of deaf and hard-of-hearing viewers, also provide subtitling increasingly. However, subtitling is far from trivial and is considered to be one of the most expensive and time-consuming tasks an interested company needs to perform, since experts mainly carry it out manually. Typically, a 1-hour programme needs around 7-15 hours of effort by humans.

In view of the expansion of digital television and the increasing demand to manipulate audiovisual content, tools producing subtitles in a multilingual setting become indispensable for the subtitling industry. Operating in this setting, the MUSA (Multilingual Subtitling of Multimedia Content) project (http://sifnos.ilsp.gr/musa/) aims at the development of a system that combines speech recognition, advanced text analysis, and machine translation to help generate multilingual subtitles. The system converts audio streams into text transcriptions, condenses and/or rephrases the content to meet the spatio-temporal constraints of the subtitling process, and produces draft translations in at least two language pairs. Three European languages are currently supported: English as source and target as far as subtitle generation is concerned, French and Greek as subtitle translation target languages.

2. Requirements and standards for subtitle production and visual presentation

Current practices and standards followed by the big media groups form the basis on which the subtitling component of the MUSA prototype converts transcripts to subtitles. They aim to provide a unifying formula based on the different subtitling conventions currently operating within various countries. They cater for standardization along the following parameters (Konstantinou, 2003):

spatial parameters (layout) including position on screen, number of lines, number of characters per line, typeface and distribution, font colour and background

temporal parameters (duration), maximum duration of a full two-line and full single-line subtitle, minimum duration of a single-word subtitle, leading-in time, lagging-out time, time between two subtitles, camera takes/cuts

punctuation and letter case, including sequence and linking dots (or ending/starting triple dots), use of ordinary punctuation marks, use of upper- and lower-case letters

target text editing including single-line vs two-line subtitle, segmentation at the highest-level linguistic nodes, segmentation and line length, relation between spoken utterances and subtitled sentences, subtitles with more than one sentence, omission or retainment of linguistic items of the original, simplification of syntactic structures, use of acronyms and other literals, use of dialects and taboo words, and use of culture-specific linguistic elements.

3. Multilingual Subtitling System Architecture

The architecture of the multilingual subtitle production line includes the following functional blocks (Piperidis *et al*, 2004, Demiros *et al*, 2004):

1. an English automatic speech recognition (ASR) subsystem for the transcription of audio streams into text, including separation of speech vs. non-speech, speaker identification and adaptation to speaker's style
2. a subtitling subsystem producing English subtitles from English audio transcriptions aiming to provide maximum comprehension while complying with spatio-temporal constraints and linguistic parameters
3. a multilingual translation subsystem integrating machine translation, translation memories and terminological banks, for English-Greek and English-French.

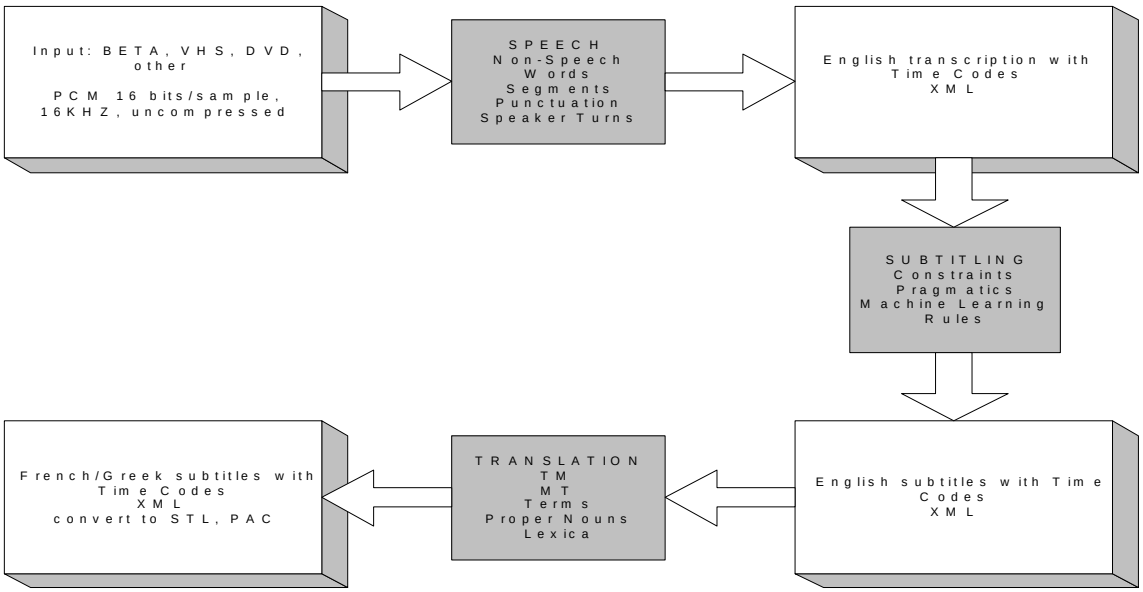

Figure 1: MUSA architecture

The component modules of the automatic speech recognizer, developed by K.U.Leuven/ESAT, include a pre-processing stage, the acoustic model (AM), the language model (LM), the lexicon and the search engine. The input to the speech recogniser is an audio file (PCM, big-endian) of 16-bit samples at 16 kHz, and the output a time-tagged text that is the word-by-word transcript of the input audio, with segments of transcript corresponding to sentences. In case the programme's transcript is available, then the ASR module aligns the audio with the transcript and provides timecodes (Figure 2). The subtitling subsystem comprises the constraint formulation and calculation module, the CNTS text condensation module and the subtitle editing module. The input to the subtitling subsystem is English transcript with time codes, words, segments, internal punctuation and speaker turns, and the output is English subtitles. The translation subsystem comprises the TrAID translation memory module (Piperidis *et al*, 1999) and the Systran machine translation engine (Systran White papers, 2003). The input to the translation subsystem is English subtitles and the output French/Greek subtitles with time codes. All data exchange between the system components is performed via XML files obeying predefined DTDs. Greek/French subtitles are linguistically processed and converted into the STL format. Formatted subtitles are then viewed and edited in a subtitle editor.

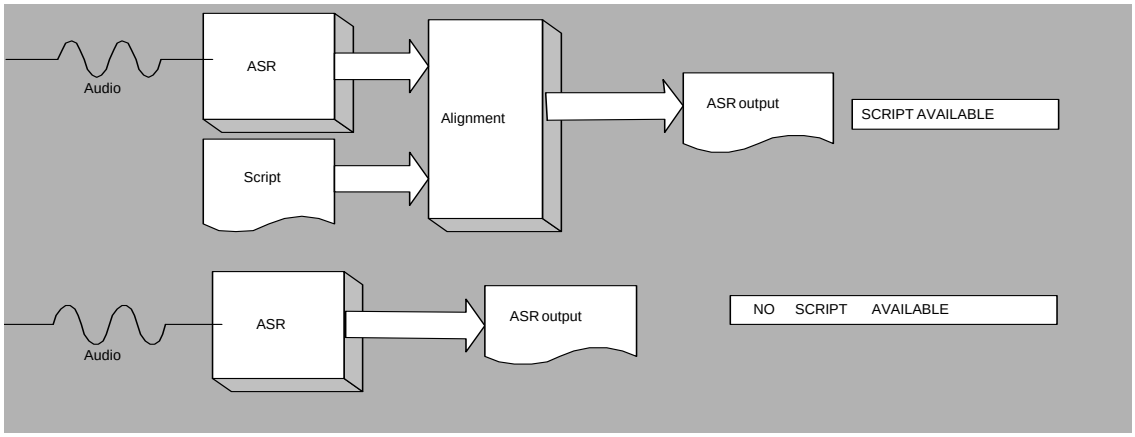

Figure 2 : Modes of ASR operation

4. Resources for multilingual subtitling

In order to train and evaluate system components, a complex array of application specific resources is necessary. The primary audiovisual data used in MUSA consist in BBC TV programmes of the type of documentaries and "newsy" current affairs programmes. For each television programme, the following multifaceted parallel data have been captured: a) the actual video of the programme, b) its transcript or script, c) English, Greek and French subtitles, d) topically relevant newspaper and web-sourced extracts. Table 1 presents the current size of the data collection.

Programmes (video-audio)	Scripts + Trasncripts	English subtitles	Greek subtitles	French subtitles
120 hours	905.752 words	650.860 words	552.575 words	129.147 words[1]

Table 1 : Size of MUSA multifaceted parallel data

5. Speech processing for subtitling purposes

Two types of data have been useful for automatic speech recognition: audio and text. Portions of the data (31 documentaries, 37-59 minutes each, total of 23h40m) have been used to create a new audio corpus since this data is more similar to what the speech recogniser will have to operate on, as compared to the WSJ corpus originally used by the speech recognition system deployed. For the same set of data, accurate transcripts were also captured and used to align the audio at phoneme level. A second type of data collected within MUSA for improving the performance of the speech recogniser is text data, necessary to build new language models but also necessary for making speech audio useful for acoustic model training. Finally, newspaper texts covering the documentaries at hand have been made available and have proved useful for tuning the language model with important keywords (like proper names) in a given documentary.

With the audio data described, context-independent acoustic models were built (Vanroose and Wambacq, 2004). To this end, the raw data were processed into a structured data corpus, involving accurate phoneme alignment of the audio with the transcripts. A new language model was built using BBC transcripts. One of the important design choices has been to include punctuation (full-stops, question marks and commas) as entries in the language model, and to make a distinction between capitalised and non-capitalised words (like e.g. "turkey" vs. "Turkey"). In addition to this generic language model, an add-on "sub-model" to this LM to cope with numbers, ordinals, and words which are not in the lexicon (mainly proper names) was developed.

Problems encountered during the speech recognition process mainly include noise conditions of the data, and the problem of non-native speakers. Most of the speech contained background music, which is not the kind of audio data on which the speech recogniser will be operating in a studio setup. In a "real-life" application, use of the unmixed audio from the speakers can be envisaged, to alleviate the noise problem. For training purposes, however, work proceeded with only those audio fragments where music or background voices were absent or very low level. Similarly, for non-native speakers, filtering out non-British speech from the data was the adopted solution.

In case the programme's transcript is available, then the ASR module aligns the audio with the transcript and provides timecodes. The presence of such transcript leads to a significant reduction of the ASR error rate to almost zero. However, if there exist large omissions in the audio, for instance when a truncated version of the original targets foreign markets, as is the case in many TV productions, then the discrepancies between the transcript and the audio could cause the synchronization to be lost.

6. Subtitling

The task of automatic subtitling presupposes the condensation of sentences, or segments, to a shorter length, in number of words and number of characters, as a function of the available space on the screen and the pace of the transcript being subtitled. The MUSA subtitling component comprises a) the constraint formulation and calculation module, b) the text condensation - segment compression - module and c) the subtitle editing module.

Available space on screen and pace of transcript are translated into a set of external constraints by the Constraint Formulation and Calculation module. Constraints are passed on to the text condensation module. Given for each segment (output from the speech recogniser) an XML file containing the segment and constraints at the word and character level, the text condensation module generates a subtitle conforming to these constraints as much as possible, and provides this output in an XML file for further processing. The compressed and linguistically processed segments are passed on to the subtitle editing module that decides where to split subtitles, if more than one subtitles have to be produced corresponding to a single segment, or subtitle lines, if the compressed segment cannot fit in a single-line subtitle.

[1] French subtitle files were not readily available and had to be produced anew. Their number corresponds to 25% of the total number of programmes.

Constraints take into account available space (layout) and time (duration), and are expressed in terms of word rate, leading-in time, brain delay, delay between subtitles, characters and words in full two-line and single-line subtitles. In the current version of the system the following constraints have been implemented :

Constraint	**Range**
Word Rate	2.5 – 3 words/sec
Leading-in Time	0.25 sec
Delay between subtitles	0.20 sec
Characters in full 2-line subtitle	70-75 chars
Words in full 2-line subtitle	14 words
Words in full 1-line subtitle	7 words

Combining time information provided by the speech recognizer, a set of two constraints for each segment has been designed: the *number of words* that have to be removed, and the *number of characters* that have to be removed.

The input to the Subtitling Component is the output of the Speech Recognizer, in XML format. The transcript contains words and pauses, as well as time information such as the start and the duration of each element. In order to create valid subtitles, the stream of transcribed words is segmented to semantically meaningful units that roughly correspond to sentences, although the proper notion of sentence is not applicable to spoken data as it is to written data. The development of the subtitling engine required perfect transcript segmentation into chunks that would feed the engine. Text condensation techniques have been applied to gold (error-free) transcript enriched with punctuation that was the result of the alignment of the speech recognizer output to the corresponding script.

The CNTS text condensation module implements a hybrid approach that combines a pragmatic and linguistic approach (Daelemans, Hoethker & Tjong Kim Sang, 2004). The pragmatic approach consists in looking up for paraphrases in a paraphrase table, extracted semi-automatically from available transcripts and their hand-made subtitles. In order to extract paraphrases, the transcripts of 87 documentaries (~480K words) were automatically aligned at sentence level with their subtitles in the corresponding subtitle files. Alignment was performed using an algorithm developed in the framework of the Atranos (http://atranos.esat.kuleuven.ac.be/) project (Tjong Kim Sang, 2003). For each aligned transcript-subtitle pair, word alignment was performed by linking identical words in the two alignment parts. If the word sequences between two such anchors in the transcript and the subtitle were different, the pair was added to the list of paraphrase candidates. Paraphrase candidates were manually checked, resulting in 1500 entries for the current paraphrase table. Similar approaches are presented in Barzilay and McKeown, 2001) and (Lin and Pantel, 2001). In cases where constraints are weak, paraphrasing seems to be a reliable way to achieve text condensation. If paraphrasing alone cannot meet the computed constraints, the linguistic approach is invoked. The linguistic approach consists in hand-crafted deletion rules which use a shallow-parse of the segments and surprise values for each word, computed on the basis of a large text corpus. Even if, compression may have been achieved by invoking just the paraphrase module, segments are shallow parsed, with MBSP (Daelemans et al., 1999; Buchholz et al, 1999), so that the subsequent subtitle editing module is provided with the necessary information as to where subtitles should be split, both at inter- and intra-subtitle level.

7. Translation data

The translation subsystem of MUSA integrates the TrAID translation memory component with the SYSTRAN machine translation engine. To populate the translation memory module databases, English-Greek subtitle files have been aligned using the TrAID aligner tool and loaded in the translation memory database.

To improve the performance of the machine translation module, the module had to be endowed with customised lexical resources. Customisation of the lexical resources of the translation subsystem consisted in a) customization by selection of the system dictionaries appropriate to the content, and b) customization by creation of external-to-the system dictionaries. The entries of the translation dictionaries in question are: a) not found words (NFW), i.e. missing words from the system's lexical resources, b) do not translate entries (DNT), i.e. proper nouns and frozen sequences that must not be translated, and c) terminological dictionaries. The first two were extracted from the BBC corpus and enriched by the feedback of the automatic speech recognition component, especially regarding proper names. These entries have been coded with elementary grammatical information and taken the form of a textual bilingual dictionary ready for compilation. In the compiled version lexical entries are enriched with more morphosyntactic and semantic properties rendering their integration into the translation output accurate.

Terminological dictionaries were obtained as a result of exploitation of the aligned bilingual subtitle files. These were processed using the TrAID bilingual term extraction tools and resulted in the extraction of e.g. 5.174 English-Greek lexical equivalences. The MUSA parallel aligned corpus consisted of 120 subtitle files. On one side of the corpus, e.g. Greek, a term extractor was applied producing a list of candidate terms. This list was subsequently fed to the TrAID bilingual concordancing tool (Antonopoulos *et al*, 2003) extracting all English translation equivalents. At the end, all automatically produced results were hand validated. These terms were used to update the terminological and lexical resources (including entries that were not found in dictionaries) that the translation system utilizes in order to advance its translation accuracy.

Translated subtitles are linguistically annotated following the principles outlined above. The ILSP linguistic processing tools (Papageorgiou *et al,* 2000; Boutsis *et al*, 2000) are used for annotating Greek translated subtitles, while for French the Systran tools have been tried.

The resources infrastructure cycle completes with bibliographic data for each broadcast that, as well as annotations from all software components, are stored in XML documents. The final output is converted into files that obey the European Broadcast Union subtitle file specifications (European Broadcasting Union, 1991).

8. Conclusion

An infrastructure for multilingual subtitle generation has been presented. The infrastructure consists of a set of multifaceted multimedia parallel data, in the sense that the same content is conveyed in different media (video, audio, text), while for component development purposes different facets of the parallel corpus are deployed. Parallel data facets include : audio-script/transcript, transcript-english subtitles, English subtitles-Greek subtitles and English subtitles-French subtitles. It will be part of our future work to exploit the parallelness between video segments-audio segments-transcripts, which has not as yet been tackled.

Acknowledgement

This work has been supported by the IST-2001-38299 MUSA project.

References

Antonopoulos, V., Malavazos, C., Triantafyllou, I., Piperidis, S., (2003) Enhancing Translation Systems with Bilingual Concordancing Functionalities, Workshop on Balkan Language Resources and Tools, Greece,http://iit.demokritos.gr/skel/bci03_workshop/pages/programme.html

Barzilay, R., McKeown, K. (2001), Extracting Paraphrases from a Parallel Corpus, in Proceedings of ACL/EACL

Boutsis, S., P. Prokopidis, V. Giouli & Piperidis, S. (2000) A Robust Parser for Unrestricted Greek Text. Proceedings of the 2nd LREC Conference, (pp. 467-473), Athens, Greece.

Buchholz, S., Veenstra, J., Daelemans, W. (1999) Cascaded Grammatical Relation Assignment, In: Proceedings of EMNLP/VLC-99, University of Maryland, USA, June 21-22

Daelemans, W., Buchholz, S., Veenstra, J. (1999) Memory-Based Shallow Parsing, in Proceedings of CoNLL-99, Bergen, Norway, June 12

Demiros, I., Vanroose, P., Daelemans, W., Sklavounou, E., Prokopidis, P., Piperidis, S., (2004) MUSA project, D3 Descriptions of the software component modules and technical specifications for their integration

European Broadcasting Union (1991). Specification of the EBU Subtitling data exchange format, TECH. 3264-E.

Daelemans, W., Hoethker, A., Tjong Kim Sang, E. (2004) Automatic Sentence Simplification for Subtitling in Dutch and English, in the Proceedings of the 4th International Language Resources and Evaluation Conference (LREC 2004), Lisbon.

Konstantinou, M., (2003) MUSA project, D2 User requirements for content production & visual presentation

Lin, D., Pantel, P. (2001). DIRT - Discovery of Inference Rules from Text, in: Knowledge Discovery and Data Mining

Papageorgiou, H., P. Prokopidis, V. Giouli and S. Piperidis, S., (2000). A Unified POS Tagging Architecture and its Application to Greek. Proceedings of the 2nd LREC Conference (pp 1455-1462), Athens, Greece.

Piperidis, S., Malavazos, C., Triantafyllou, Y., (1999). A Multi-level Framework for Memory-Based Translation Aid Tools, Aslib, Translating and the Computer 21, London

Piperidis, S., Demiros, I., Prokopidis, P., Vanroose, P., Hoethker, A., Daelemans, W., Sklavounou, E., Konstantinou, M., Karavidas, Y. (2004) Multimodal Multilingual Resources in the Subtitling Process, Proceedings of Fourth International Conference on Language Resources and Evaluation-LREC2004, 26-28 May 2004, Lisbon, Portugal

Systran White papers (2003) http://www.systransoft.com/Technology/WhitePapers.html

Vanroose, P. Wambacq, P. (2004) MUSA project, D4.1 Continuous speech recognition module with large vocabulary, running in real time with efficient acoustic models and with a good language model for general use, operating in English

Tjong Kim Sang, E. F. (2003). Alignment of Transcribed Text with Subtitles - ATraNoS WP4-01