Multimodal Multilingual Resources in the Subtitling Process

Stelios Piperidis[1,2], Iason Demiros[1,2], Prokopis Prokopidis[1,2], Peter Vanroose[3], Anja Hoethker[4], Walter Daelemans[4], Elsa Sklavounou[5], Manos Konstantinou[6], Yannis Karavidas[7]

[1]Institute for Language and Speech Processing, Artemidos 6 & Epidavrou, 151 25 Athens, Greece
[2]National Technical University of Athens
[3]Katholieke Universiteit Leuven, div. ESAT/PSI, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium
[4]CNTS Language Technology Group– Universiteit Antwerpen, Universiteitsplein 1, B-2610 Antwerpen Belgium
[5]Systran SA, 1 rue du Cimetiere-BP 7, 95230 Soisy Sous Montmorency, France
[6]Lumiere Cosmos Communications SA, Lazarou Sohou 5, 11525 Athens, Greece
[7]British Broadcasting Corporation (World Service), Bush House, Strand, London WC2 4PH
{spip, iason, prokopis}@ilsp.gr, Peter.Vanroose@esat.kuleuven.ac.be, { walter.daelemans, hoethker}@uia.ua.ac.be,
sklavounou@systran.fr, mkonstantinou@lumiere.gr, yannis.karavidas@bbc.co.uk

Abstract

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 project aims at the development of a system which combines speech recognition, advanced text analysis, and machine translation to help generate multilingual subtitles; a system that converts audio streams into text transcriptions, condenses the content to meet the spatio-temporal constraints of the subtitling process and produces draft translations in two language pairs. Three European languages are supported: English as source and target as far as subtitling 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 and "newsy" current affairs programmes. For each programme, the following data are captured: the actual video, its transcript or script, English, Greek and French subtitles, and topically relevant newspaper or web-sourced extracts.

1. Introduction

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 globalised media audience. In such a unified framework of mass communication, subtitling is playing a critical role. In many countries, subtitling is the most commonly used method for conveying the content of foreign language narrative or dialogue to the audience. However, subtitling is far from trivial and is deemed to be a very expensive and time-consuming task, 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 the various European countries. They cater for standardization along the following parameters: spatial parameters (layout), temporal parameters (duration), punctuation and letter case, and target text editing. These are the most language-technology-based and demanding requirements and include: single-line vs. two-line subtitles, subtitle segmentation at the highest linguistic nodes, subtitle segmentation and line length, spoken utterances and subtitled sentences, subtitles with more than one sentences, omission of linguistic items of the original (like padding expressions, tautological cumulative adjectives/adverbs, responsive expressions), retaining of linguistic items of the original, alterations of syntactic structures, etc. (Konstantinou, 2003)

3. Multilingual Subtitling System Architecture

The architecture of the multilingual subtitle production line includes the following functional blocks (Demiros *et al*, 2003):

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.

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 in principle 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. The subtitling subsystem comprises the constraint formulation and calculation module, the 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 or 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.

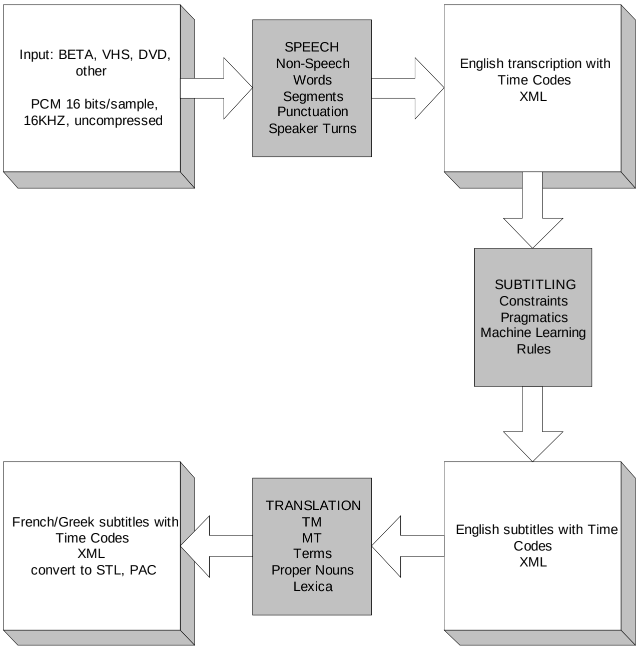
Input: BETA, VHS, DVD, other
PCM 16 bits/sample, 16KHZ, uncompressed
SPEECH
Non-Speech
Words
Segments
Punctuation
Speaker Turns
English transcription with Time Codes
XML
SUBTITLING
Constraints
Pragmatics
Machine Learning
Rules
English subtitles with Time Codes
XML
TRANSLATION
TM
MT
Terms
Proper Nouns
Lexica
French/Greek subtitles with Time Codes
XML
convert to STL, PAC

Figure 1: MUSA architecture

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. Available primary data have been filtered based on the existence of Greek subtitles. For each television programme, the following 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.

A total of ca 120 hours of running programmes have been sourced, associated with a total of 905.752 words in transcripts and/or scripts, 650.860 words in English subtitles, 552.575 words in Greek subtitles and 129.147 words in French subtitles (Karavidas, 2003).

4.1 Speech processing

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. 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 to be 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. Since for building a language model a much larger amount of textual data is required, transcripts of all BBC documentaries have been sourced. 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.

4.1.1 Creation of a speech data corpus

With the audio data described, first of all context-independent acoustic models have been built. To this end, the raw data has been processed into a structured data corpus: this involved accurate phoneme alignment of the audio with the transcripts, which therefore first needed to be tokenised. Automatic alignment software was available and has been used, while tokenisation scripts were developed from scratch since they had to be specific to the format of the BBC transcripts, and to reflect the design choices made for the lexicon, viz. the distinction of capitalised and non-capitalised words, and the explicit representation of pauses such as commas and full stops. Both of these are unusual in the classical framework of speech recognition but are of crucial importance to a subtitling framework.

4.1.2 Construction of language models

The most important design choice 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"). First, a new trigram language model was built from the WSJ data, now containing punctuation and capitalisation (which was not the case with the existing WSJ models). Next, a language model was built using BBC transcripts. This was necessary since the WSJ models are based on American English newspaper texts, which have too different idiomatics to be acceptable for transcribing BBC programmes. On the other hand, the WSJ corpus is much larger than the newly created BBC corpus, hence its "coverage" of general language phenomena is better. Therefore, we tried to combine the advantages of the two. In addition to this generic language model, we have also developed an add-on "sub-model" to this LM to cope with numbers, ordinals, and words which are not in the lexicon (mainly proper names).

4.1.3 Alignment of audio with transcripts

A context independent acoustic model essentially consists of a set of probability distributions in 39-dimensional feature space, one PDF per phoneme. To build an acoustic model from an audio corpus, it is thus necessary and sufficient to know for each frame of the audio data to which phoneme it belongs.

To this end, first the audio must be aligned with an accurate transcript, giving a word-level alignment that must then be further refined to a phoneme-level alignment. In practice, only phoneme-level alignment is done, since the transcript is first converted to a phoneme string, or actually a phoneme graph, as the lexicon may specify more than one phoneme transcription for a word. One may look at this alignment as a linear mapping problem, where the only uncertainty is the duration of each of the phonemes in the phoneme string.

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 contains 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 proceeds with only those audio fragments where music or background voices are absent or very low level. Similarly, for non-native speakers, filtering out non-British speech from the data was the adopted solution.

4.2 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.

4.2.1 Constraint formulation and calculation

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 Speech Recognition) 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.

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. 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.

4.2.2 Text condensation

The text condensation module implements a hybrid approach that combines the following modules (Hothker, *et al*, 2003).

1. Lookup from a table of paraphrases, extracted semi-automatically from available transcripts and their hand-made subtitles. (Pragmatic Approach).
2. Hand-crafted deletion rules (Linguistic Approach). These rules use as information: a shallow-parse of the segments and surprise values for each word, computed on the basis of a large text corpus.

First the pragmatic module is invoked and if it cannot achieve the required condensation, the linguistic approach module is activated. Given that sometimes the condensation constraints are fairly weak, paraphrases provide a reliable and accurate way of achieving sentence compression. Examples include: [Within the next few years -> Soon, During the years when -> While, Whether or not -> If].

The paraphrases used by the pragmatic module were derived from the actual BBC data. For this purpose the transcripts of 87 documentaries (~480K words) were automatically aligned on sentence level with their subtitles in the corresponding subtitle files. This was done using an algorithm developed in the framework of the Atranos project (Tjong Kim Sang, 2003). The alignment algorithm estimates the probability that sentences belong together based on the words that they contain. The system benefits from making several passes over the data. To minimise the number of misalignments, we checked alignments in which the ratio of transcript words with no counterpart in the subtitling sentence exceeded a threshold.

For every transcript-subtitle pair, a 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 to decide which of them were suitable for our purposes, resulting in 1500 entries for the current paraphrase table. See Barzilay *et al*. (2001) and Lin and Pantel, (2001) for similar approaches.

Paraphrases are applied in no particular order, since they do not change the content of a sentence. The lookup process is repeated until either compression rates are satisfied or no more paraphrases are found. After that the segment is shallow-parsed with MBSP (Daelemans et al., 1999; Buchholz et al, 1999). This is necessary even if no compression is required, since linguistic information is passed to the subtitling editing module to aid the decision where to split a subtitle. If more compression is needed, the rule-based module is called. This module uses the linguistic annotation to mark parts of the sentences that can be deleted without affecting the syntactical correctness of the segment. Examples for such deletions are adverbs, prepositional phrases etc. The different suggestions are rated in order to first delete the less informative sections and delete more important parts only if necessary. To estimate the importance of word sequences we used the BNC. The surprise value for a word is determined by the log-likelihood of its unigram frequency in this corpus. To compute the informativeness of a sequence of words we calculate the average surprise value of all its content words. We keep deleting subsentences until either compression rates are satisfied or no suggestions are left.

4.3 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

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 in section 4.2.2. 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).

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., (2003) 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.

Hothker, A., Daelemans, W., Demiros, I., Piperidis, S., (2003) MUSA project, D5.1 Initial prototype of the Subtitling core engine

Karavidas, Y. (2003) MUSA Project D2a Data Collection

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

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

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