JMLR: Workshop and Conference Proceedings 11 (2010) 103–110 Workshop on Applications of Pattern Analysis

Interactive Pattern Recognition and Human Language Technology for Digital Audiovisual Content Processing

Antonio Lagarda ALAGARDA@ITI.UPV.ES
Jorge Civera JCIVERA@DSIC.UPV.ES
Alfons Juan AJUAN@DSIC.UPV.ES
Francisco Casacuberta FCN@ITI.UPV.ES
DSIC/ITI, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain

Editors: Tom Diethe, Nello Cristianini, John Shawe-Taylor

Abstract

This paper describes ongoing research work by the Pattern Recognition and Human Language Technology (PRHLT) group (UPV PASCAL2 node) in two important technology transfer projects: `i3media` and `erudito.com`. On the one hand, `i3media` (2007-2010) is a 35 M€ "tractor" technology project within the Spanish *Programa CENIT-Ingenio 2010,* run through a consortium of 12 main enterprises of the media sector, which also involve 19 research groups, including PRHLT. `i3media` focuses on the creation and automated management of *intelligent audiovisual content,* so as to facilitate both, content personalisation and interaction with users (`i3media.barcelonamedia.org`). Our participation in `i3media` is centred on interactive machine translation, to transfer and adapt our experience on this technology to i3media-specific needs. On the other hand, `erudito.com` (2010-2012) is a 1.4 M€ experimental design project, supported by the Spanish Ministry of Industry, Tourism and Trade under the *Avanza I+D* program, aimed at developing a tool to encapsulate, distribute and intelligently use digital content such as that showed on thematic TV channels. In this project, PRHLT contributes to the development of interactive closed captioning (speech transcription) and machine translation tools.

Keywords: Interactive Machine Translation, Interactive Speech Transcription, Automatic Closed Captioning

1. Introduction

The Pattern Recognition and Human Language Technology (PRHLT) research group at the *Universitat Politècnica de València* (UPV PASCAL2 node) is involved in two important Spanish projects on the application of interactive PRHLT to digital audiovisual content: `i3media` and `erudito.com` .

`i3media` (2007-2010) is a 35 M€ "tractor" technology project within the Spanish *Programa CENIT-Ingenio 2010,* run through a consortium of 12 main enterprises of the media sector, which also involve 19 research groups, including PRHLT. It focuses on the creation and automated management of *intelligent audiovisual content,* so as to facilitate both, content personalisation and interaction with users [i3media]. The results are expected to have an important impact on all companies and areas of the media sector (including production,

post-production, storage, search, indexing and distribution) as well as on other fields (such as catering, leisure and automobiles).

`erudito.com` (2010-2012) is a 1.4 M€ experimental design project, supported by the Spanish Ministry of Industry, Tourism and Trade under the *Avanza I+D* program, aimed at developing a tool to encapsulate, distribute and intelligently use digital content such as that showed on thematic TV channels. In particular, `erudito.com` takes the challenge of working with children from 6 to 16 years old, and thus the level of user knowledge assumed is that of primary and secondary education. The project consortium includes companies from the media sector, research groups with experience in PRHLT, and final users (schools).

This paper describes the research goals, work done and future work of the PRHLT group in both projects. It must be noted that `i3media` is at its final year of execution, while `erudito.com` has just been started.

2. i3media

2.1. Natural language goals

An important activity in the project tackles natural language analysis and generation, both written and spoken, in several domains and languages. This activity develops procedures to automatically understand, create and modify linguistic aspects in audiovisual products, which are called in general written speechs, such as written screenplays, their spoken versions and some other linguistic elements.

This project applies state-of-the-art technologies for the automatic generation of the linguistic side of multimedia contents from a conceptual representation in specific genres:

- Automatic identification and annotation of relevant elements in written speechs.
- Summarisation of written speechs.
- Machine Translation (MT) applied to producing and personalising audiovisual content.
- Emotional inflection in synthetic voices.

The genres considered include, among others, news, sports, virtual commentators, dialogues in interactive games, forecast reports, economic reports, etc.

2.2. Machine translation goals

i3media works on the adaptation of a series of MT technologies to speech transcription texts. This task has a constructive approach, starting with clearly restricted domains and a given language pair in just one translation direction. The developed technologies will progressively expand to other more complex domains, language pairs, and directions. From a technological point of view, the main points considered are:

- Translation models from parallel corpora and dictionaries [Badia et al. (2005)].
- Integration of linguistic information [Hassan et al. (2006)].
- Application of interactive MT techniques [Barrachina et al. (2009)].
- Ordering of multiple translation alternatives depending on text and domain.

In the case of the PRHLT group, the main effort has focused first on the development of a statistical MT (SMT) prototype, and then, on the incorporation of interactive and predictive capabilities to this prototype.

2.2.1. Statistical machine translation prototype

PRHLT has developed a SMT prototype based on Moses [Koehn et al. (2007)]. Moses is a LGPL toolkit that implements state-of-the-art SMT techniques. Given a corpus, Moses trains and combines statistical phrase-based models to provide a translation from a sentence.

Corpus. SMT systems learn a model of how to translate between languages taking examples from parallel corpora with sentences in a source language and their translations into a target language. This model is employed to translate new sentences between those languages. One of the main advantages of these systems is their ability to work with any language pair and domain, with the only requirement of providing a corpus with examples of the desired languages and domains. Thus, one of our first steps in `i3media` was the collection of an appropriate corpus to learn a model.

PRHLT collected a parallel corpus in the main domain and languages of the project: journalistic texts in Catalan and Spanish. These texts have been downloaded from *El Periódico de Catalunya* website (*www.elperiodico.com*), a newspaper that publishes articles in both languages. Table 1 shows a set of basic statistics for the partition of this corpus in training and test. These figures reflect its large size with 2 millions sentence pairs and more than 30 millions running words.

Table 1: Raw corpus statistics (K stands for thousand, and M for million)

		Spanish	Catalan
Training	Sentence pairs	2M	
	Running words	33M	33M
	Vocabulary size	752K	780K
	Average sentence length	17.5	17.3
Test	Sentence pairs	10K	
	Running words	175K	174K
	Vocabulary size	36K	36K
	Perplexity (Trigram)	192	195

Statistical models. Due to the size of the corpus, the statistical models learnt are too large. As a result, translation search time and memory requirements are not suitable to be used in a real-time prototype. In order to reduce the size of the model and searching time, pruning techniques were applied [Sanchis-Trilles and Casacuberta (2008)]. These techniques reduce model size in more than 95%, improving search speed without significantly affecting the translation quality. Table 2 compares the model size, search speed, and translation quality in terms of BLEU [Papineni et al. (2002)] between the original and reduced model.

The reduced model achieves important space savings and faster search with little translation quality degradation.

Table 2: Comparison of translation results between original and reduced model in terms of BLEU scores, translated words per second and model sizes.

	BLEU	words/second	model size
Original model	87.7	0.01152	2 GBytes
Reduced model	82.8	0.00245	82 MBytes

Graphical interface. The SMT prototype has a graphical interface in GTK+, which provides the user a SMT and post-editing environment. The user chooses a text file in the source language, whose sentences are translated by the Moses toolkit. Then, the proposed translations can be post-edited and stored. As shown in Figure 1, the interface is organised in three areas In the centre, a blue text box shows the working space with the current sentence and its translation being corrected by the user. All the sentences that appear before the current sentence in the source text file are shown in the upper part of the window, along with their respective translations if they have been previously translated. The sentences displayed after the current sentence are shown on the bottom of the window.

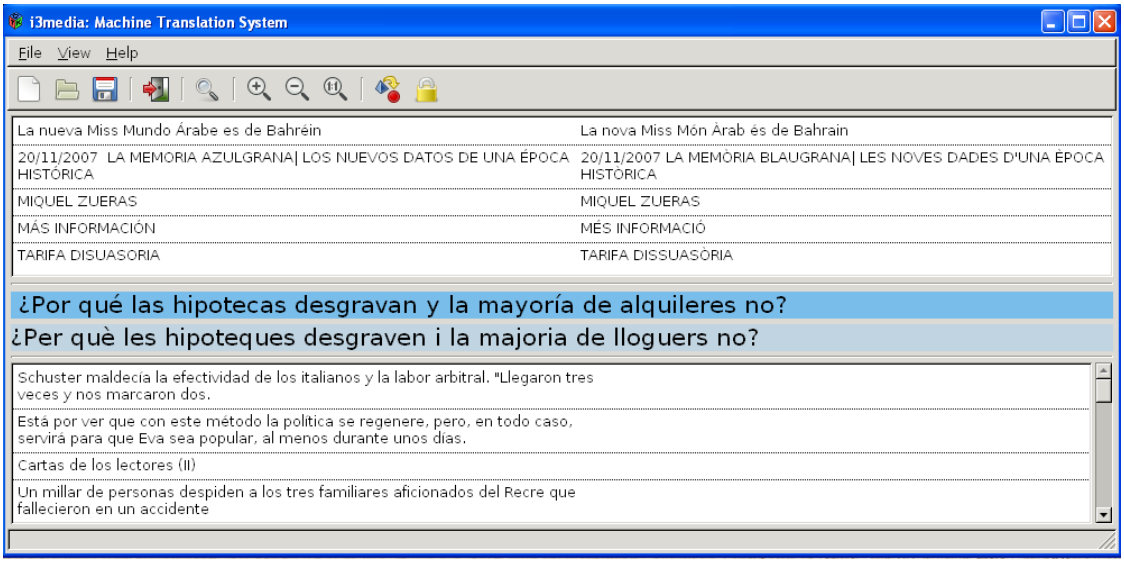

Figure 1: Graphical interface for the SMT prototype in `i3media` project.

2.2.2. Future work: Interactive machine translation prototype

MT is far from perfect, especially in a general domain task as the i3media project. As a result, the user usually needs to correct (post-edit) the translations provided by the SMT system. Interactive MT takes advantage of these corrections to provide improved

translations in an interactive fashion. A collaborative interaction is established between the user and the system, which incrementally builds the correct translation reducing the user effort. One of the objectives in `i3media` is to extend the current SMT prototype to build an interactive MT prototype using the Thot toolkit [Ortiz et al. (2005)].

3. erudito.com

3.1. Scientific and Technical Goals

The erudito.com considers different scientific and technical goals related to the application described in Section 1, however we will focus on the following two goals:

- To develop a closed captioning (CC) system.
- To develop a machine translation (MT) system for closed captions.

Figure 2 contextualises the CC and MT prototype with the other two main components of the erudito.com system. These components are the content and semantic manager shown on the left and right side, respectively. The CC and MT prototype, shown in the centre, provides closed captions integrated and synchronised with audiovisual content provided as input, allowing to switch these captions to other languages.

3.1.1. CC system

This system considers two scenarios. The first one assumes that the captions are provided, so the system should be able to automatically segment and synchronise them. In the second scenario captions are not available, then they will be generated with the help of an interactive computer-assisted CC system under the supervision of a human expert.

Regarding the first scenario, automatic processing methods for audiovisual content will be developed to provide a timed synchronisation of closed captions using speech recognition technology. The result will be a timed caption file defining the time interval in which each word is uttered. The CC system will be designed bearing in mind the peculiarities of the audiovisual content to process. More precisely, it is necessary to develop the necessary technology to face the problems dealing with audiovisual content of long duration.

Silence or soundtrack periods can be employed as segmentation tags to simplify signal processing and improve the results. Furthermore, speech recognition technology must be adapted to take into account the specific context (acoustic and language register) in audiovisual content. The fact that professional broadcasters are employed can significantly improve the quality of the captions automatically generated. On the other hand, it is necessary to filter non speech sounds such as noise, music, applause, etc., that could interfere in the synchronisation process.

In the second scenario, a human expert will supervise the captioning, segmentation and synchronisation process in an interactive manner with the help of a computer-assisted CC system. This system will be able to learn from the user corrections and validations in order to refine and adapt the subsequent captions provided. The idea behind this behaviour is to perform a dynamic context adaptation of the initial statistical models employed by the speech recognition system to improve its performance. At the same time, human experts

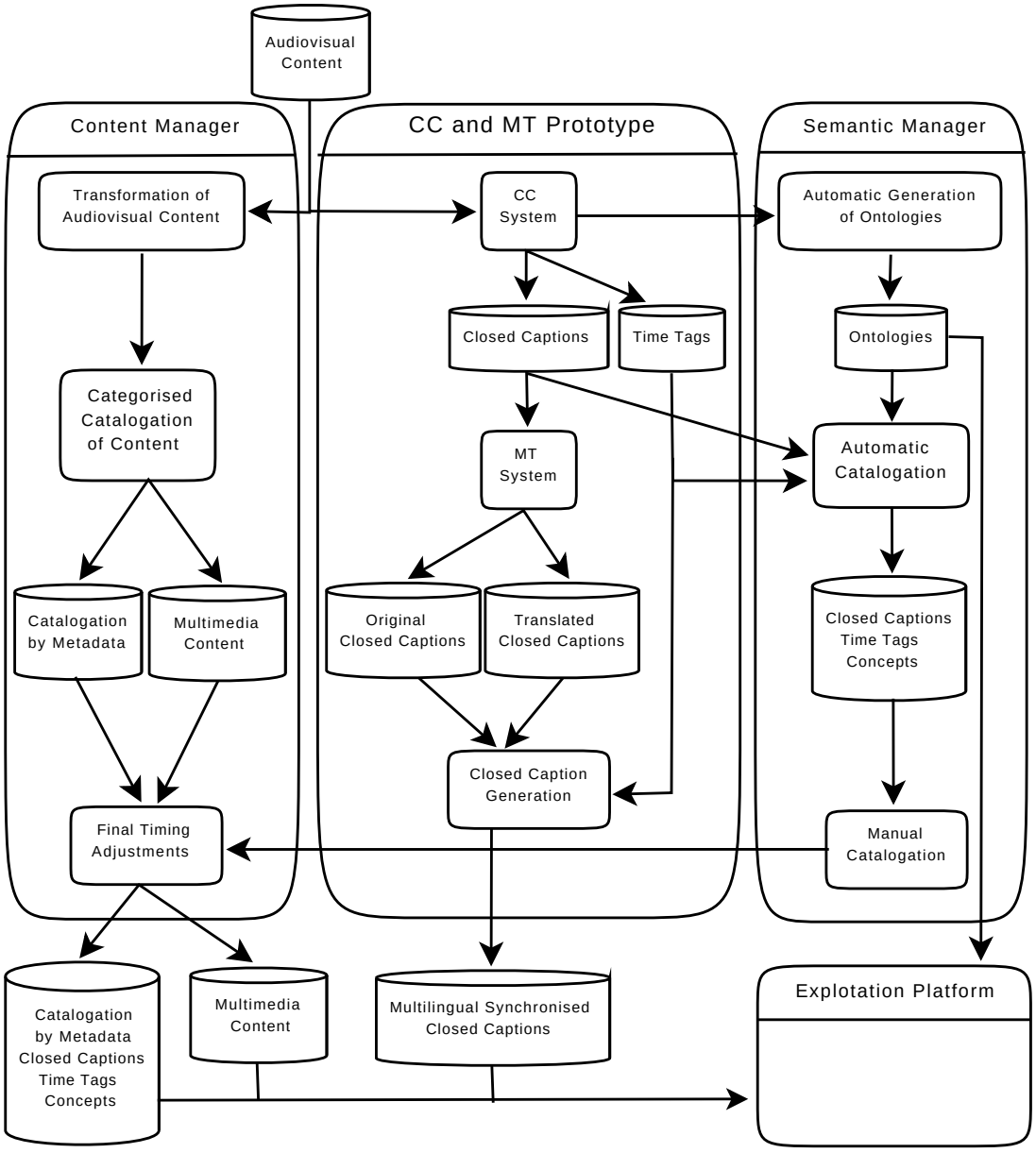

Figure 2: The system architecture for the `erudito.com` project encloses three main components: the content manager, the CC and MT prototype, and the semantic manager whose workflows are shown from left to right. The CC and MT prototype receives as input audiovisual content, and provides as output closed captions and time tags for the semantic manager, and multilingual synchronised closed captions to the exploitation platform.

can take advantage of this adaptive and interactive computer-assisted CC system to increase his/her productivity. In this sense, confidence measures are also interesting in this framework since they can guide the expert in the supervision process highlighting those captions more likely to be incorrect.

Once captions are made available, the user can search for audiovisual content based on keywords. This search should be able to be performed at the level of short time segments (a few seconds or minutes), so that the user just watches what is interested in. Nevertheless, timed captions are not only used for searching and selection purposes, but also they can be used to automatically generate captions for deaf people. In this case, it is necessary to solve an additional problem regarding syntactic independence and adequate duration.

3.1.2. MT system for closed captions

Another goal of this project is to develop an integrated audiovisual management system including the option to automatically translate captions. Initially, an MT system to translate from Spanish into Catalan will be developed, although the translation to other language is possible. Besides, multimodality will be explored as the translation from more than one language (English and Spanish) into other language (Catalan) aiming at improving the translation quality. The aim is to adapt the existing MT technology to the CC context improving the translation quality offered by commercial systems. To this purpose, we will take into account the peculiarities of the language used in audiovisual content to carry out a theme adaptation of the initial MT models. Ideally, this adaptation requires a large parallel corpus, in our case a Spanish-Catalan corpus, related to the theme and language register. However, given the limited availability of parallel texts for a specific context, monolingual techniques for adaptation will be explored. The incorporation of linguistic information and preprocessing rules are also considered in the development of this MT system.

Acknowledgments

Work supported by the EC (FEDER, FSE) and the Spanish Government (MICINN, MITyC, "Plan E", under grants MIPRCV "Consolider Ingenio 2010", iTrans2 TIN2009-14511, Mediapro-i3media CENIT-2007-1012 and erudito.com TSI-020110-2009-439).

References

T. Badia et al. An n-gram approach to exploiting a monolingual corpus for Machine Translation. In *Proc. of the MT Summit X*, pages 1–7, 2005.

S. Barrachina et al. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28, 2009.

H. Hassan, M. Hearne, A. Way, and K. Sima'an. Syntactic Phrase-Based Statistical Machine Translation. In *Proc. of the WSLT*, pages 238–241, 2006.

i3media. `i3media.barcelonamedia.org`.

P. Koehn et al. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL*, pages 177–180, 2007.

D. Ortiz, I. Garcia-Varea, and F. Casacuberta. Thot: a Toolkit To Train Phrase-based Statistical Translation Models. In *Proc. the MT Summit*, pages 141–148. 2005.

K. Papineni, S. Roukos, T. Ward, and W.-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, 2002.

G. Sanchis-Trilles and F. Casacuberta. Increasing Translation Speed in Phrase-Based Models via Suboptimal Segmentation. In *Proc. of PRIS*, pages 135–143, 2008.