Scientific Report

First name / Family name	LUIS ALBERTO / BARRÓN CEDEÑO
Nationality	MEXICAN
Name of the *Host Organisation*	UPC + UPM
First Name / family name of the *Scientific Coordinator*	LLUÍS MÀRQUEZ (UPC) / MANUEL CARRO (UPM)
Period of the fellowship	18/07/2012 to 17/07/2013

This document reports on my activities as *ERCIM Alain Bensoussan* fellow (12 months). In the framework of such fellowship, I joined two institutions: *Universitat Politècnica de Catalunya (UPC)* and *Universidad Politécnica de Madrid (UPM)*. In the following sections I report on my activities and achievements in both institutions.

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

The main objective of my work during this fellowship was the development and application of models for multilingual natural language processing. In particular, we were interested in exploring the use of multilingual resources for the enhancement of machine translation, as well as the design of (cross-language) text similarity models. Note that although the activities appear for each institution, indeed they overlap significantly, and no clear border can be established.

I.1 Universitat Politècnica de Catalunya

During the first stage of the research our aim was determining the degree of similarity between two texts, at semantic level. In order to do so, we applied a manifold of features: from simple character-based measures up to semantic ones. The semantic

measures are particularly interesting. The first of them, probably the most promising, is based on a projection of the analysed texts to the space of Wikipedia [1]. The second one, indeed a set of features, is based on confidence estimation measures, borrowed from automatic machine translation [2]. This research work resulted in publication 1, in Section II.

We also focused on the problem of improving machine translation models. We did so by exploiting users' feedback to an online machine translation system. On the basis of a machine learning strategy we automatically filtered good instances of feedback and used them to enhance a state-of-the-art statistical machine translation system. This research produced publication 2 in Section II.

I.2 Universidad Politécnica de Madrid

Afterwards, we started working on the problem of exploiting Wikipedia as a source for human-made translations. Although plenty of parallel corpora are available for a bunch of well-resourced languages, we are focused on less-resourced ones; namely Basque and Catalan (and their links to both Spanish and English). We are designing a framework for the automatic extraction of translations from Wikipedia editions in these languages. This is an ongoing work.

Beside the discussed research, we further worked on my PhD topic: plagiarism detection. We particularly focused on two problems. Firstly, we analysed a set of plagiarism instances from a linguistic point of view: at the paraphrase level. Secondly, we developed a software framework for automatic cross-language plagiarism detection. This research produced publications 3 and 4 in Section II.

References

1. Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
2. Gesús Giménez and Lluís Màrquez. 2010b. Linguistic Measures for Automatic Machine Translation Evaluation. Machine Translation, 24(3–4):209–240

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

The following four papers are either published or accepted for publication.

1. Barrón-Cedeño, Màrquez, Fuentes, Rodríguez, Turmo. UPC-CORE: What Can Machine Translation Evaluation Metrics and Wikipedia Do for Estimating Semantic Textual Similarity? In: Proc. of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Atlanta, GA, 2013.

   Abstract: In this paper we discuss our participation to the 2013 Semeval Semantic Textual Similarity task. Our core features include (i) a set of metrics borrowed from automatic machine translation, originally intended to evaluate automatic against reference translations and (ii) an instance of explicit semantic analysis, built upon opening paragraphs of Wikipedia 2010 articles. Our similarity estimator relies on a support vector regressor with RBF kernel. Our best approach required 13 machine translation metrics + explicit semantic analysis and ranked 65 in the competition. Our post-competition analysis shows that the features have a good expression level, but overfitting and —mainly— normalization issues caused our correlation values to decrease.

2. Barrón-Cedeño, Màrquez, Henríquez, Formiga, Merino, May. Identifying Useful Human Correction Feedback from an On-line Machine Translation Service. In: Proc. of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), 2013

   Abstract. Post-editing feedback provided by users of on-line translation services offers an excellent opportunity for automatic improvement of statistical machine translation (SMT) systems. However, feedback provided by casual users is very noisy, and must be automatically filtered in order to identify the potentially useful cases. We present a study on automatic feedback filtering in a real weblog collected from Reverso.net. We extend and re-annotate a training corpus, define an extended set of simple features and approach the problem as a binary classification task, experimenting with linear and kernel-based classifiers and feature selection. Results on the feedback filtering task show a significant improvement over the majority class, but also a precision ceiling around 70-80%. This reflects the inherent difficulty of the problem and indicates that shallow features cannot fully capture the semantic nature of the problem. Despite the modest results on the filtering task, the classifiers are proven effective in an application-based evaluation. The incorporation of a filtered set of feedback instances selected from a larger corpus significantly improves the performance of a phrase-based SMT system, according to a set of standard evaluation metrics.

3. Barrón-Cedeño, Vila, Martí, and Rosso. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics 39(4) (to appear; accepted November 2012)

Abstract. Although paraphrasing is the linguistic mechanism underlying many plagiarism cases, little attention has been paid to its analysis in the framework of automatic plagiarism detection. Therefore, state-of-the-art plagiarism detectors find it difficult to detect cases of paraphrase plagiarism. In this article, we analyze the relationship between paraphrasing and plagiarism, paying special attention to which paraphrase phenomena underlie acts of plagiarism and which of them are detected by plagiarism detection systems. With this aim in mind, we created the P4P corpus, a new resource that uses a paraphrase typology to annotate a subset of the PAN-PC-10 corpus for automatic plagiarism detection. The results of the Second International Competition on Plagiarism Detection were analyzed in the light of this annotation. The presented experiments show that (i) more complex paraphrase phenomena and a high density of paraphrase mechanisms make plagiarism detection more difficult, (ii) lexical substitutions are the paraphrase mechanisms used the most when plagiarizing, and (iii) paraphrase mechanisms tend to shorten the plagiarized text. For the first time, the paraphrase mechanisms behind plagiarism have been analyzed, providing critical insights for the improvement of automatic plagiarism detection systems.

4. Barrón-Cedeño, Gupta, and Rosso. **Methods for cross-language plagiarism detection**. Knowledge-Based Systems (to appear; available online 3 July 2013)

Abstract. Three reasons make plagiarism across languages to be on the rise: (i) speakers of under-resourced languages often consult documentation in a foreign language, (ii) people immersed in a foreign country can still consult material written in their native language, and (iii) people are often interested in writing in a language different to their native one. Most efforts for automatically detecting cross-language plagiarism depend on a preliminary translation, which is not always available.

In this paper we propose a freely available architecture for plagiarism detection across languages covering the entire process: heuristic retrieval, detailed analysis, and post-processing. On top of this architecture we explore the suitability of three cross-language similarity estimation models: Cross-Language Alignment-based Similarity Analysis (CL-ASA), Cross-Language Character n-Grams (CL-CNG), and Translation plus Monolingual Analysis (T+MA); three inherently different models in nature and required resources.

The three models are tested extensively under the same conditions on the different plagiarism detection sub-tasks—something never done before. The experiments show that T+MA produces the best results, closely followed by CL-ASA. Still CL-ASA obtains higher values of precision, an important factor in plagiarism detection when lesser user intervention is desired.

Potential publications regarding the last work at UPM are still under preparation.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

1. *Primer taller latinoamericano de tratamiento automático del lenguaje* (First Latinamerican Workshop on Natural Language Processing). July 3-5, 2013. Puebla, Mexico
2. 23rd International Joint Conference on Artificial Intelligence. August 3-9, 2013. Beijing, China.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

IV.1 INRIA Sophia Antipolis

Country:	France
Project:	Wimmics
Coordinator:	Dr. Fabien L. Gandon
Dates:	April 25th to May 3rd, 2013

In this institution I mainly interacted with Drs. Elena Cabrio and Fabien Gandon. We discussed potential collaboration on similarity metrics and question answering issues. Moreover, I offered a talk in the framework of the Wimmics seminar: "Uncovering Good Feedback Instances from an On-line Machine Translation System."

IV.2 NTNU

Country:	Norway
Project:	Department of Computer and Information Science
Coordinator:	Dr. Kjetil Norvag
Dates:	May 24th to June 1st, 2013

In this institution I mainly interacted with Drs. Dirk Ahlers and Kjetil Norvag. Their research on information retrieval was indeed very interesting. We identified a good potential collaboration project: automatic cross-language knowledge acceleration. It seems like we can consolidate this collaboration and come out with some interesting work. Moreover, I offered the talk "Detection of (Cross-Language) Text Re-Use and Plagiarism."