Proceedings of the Workshop on Iberian Cross-Language Natural Language Processings Tasks (ICL 2011)

held in conjunction with

27th Conference of the Spanish Society for Natural Language Processing

Editors

Paolo Rosso
Alberto Barrón-Cedeño
Marta Vila
Jorge Civera
Anabela Barreiro
Iñaki Alegria

Huelva, Spain, September 7th 2011

Preface

In the Iberian Peninsula, five official languages co-exist: Basque, Catalan, Galician, Portuguese and Spanish. Fostering multi-linguality and establishing strong links among the linguistic resources developed for each language of the region is essential. Additionally, a lack of published resources in some of these languages exists. Such lack propitiates a strong inter-relation between them and higher resourced languages, such as English and Spanish.

In order to favour the intra-relation among the peninsular languages as well as the inter-relation between them and foreign languages, different purpose multilingual NLP tools need to be developed. Interesting topics to be researched include, among others, analysis of parallel and comparable corpora, development of multilingual resources, and language analysis in bilingual environments and within dialectal variations.

With the aim of solving these tasks, statistical, linguistic and hybrid approaches are proposed. Therefore, the workshop addresses researchers from different fields of natural language processing/computational linguistics: text mining, machine learning, pattern recognition, information retrieval and machine translation.

The research in this proceedings includes work in all of the official languages of the Iberian Peninsula. Moreover, interactions with English are also included. Wikipedia has shown to be an interesting resource for different tasks and has been analysed or exploited in some contributions.

Most of the regions of the Peninsula are represented by the authors of the contributions. The distribution is as follows: Basque Country (2 authors), Catalonia (7 authors), Galicia (4 authors), Portugal (2 authors) and Valencia (5 authors). Interestingly, those regions where Spanish is the only official language are not represented. It is worth noting that authors working beyond the Peninsula have also contributed to this workshop, including: Argentina (3 authors), Finland (1 author), France (2 authors), Mexico (1 author), Singapore (1 author), and USA (6 authors).

The ICL workshop has been organised as one of the activities of the VLC/-CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems, EC WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework; the FPU Grant AP2008-02185 from the Spanish Ministry of Education; and MICINN Text-Enterprise 2.0 (TIN2009-13391-C04-03) and Text-Knowledge 2.0 (TIN2009-13391-C04-04) projects within the Plan I+D+i.

Organising Committee

Paolo Rosso	Universitat Politècnica de València
Alberto Barrón-Cedeño	Universitat Politècnica de València
Marta Vila	Universitat de Barcelona
Jorge Civera	Universitat Politècnica de València
Anabela Barreiro	INESC-ID Lisbon
Iñaki Alegria	Euskal Herriko Unibertsitatea

Program Committee

Eneko Agirre	University of the Basque Country
Amparo Alcina	Universitat Jaume I
Iñaki Alegria	Euskal Herriko Unibertsitatea
Jesús Andrés Ferrer	Universitat Politècnica de València
Alexandra Balahur	DLSI - University of Alicante
Anabela Barreiro	INESC-ID Lisbon
Alberto Barrón Cedeño	Universitat Politècnica de València
Yassine Benajiba	Philips Research North America, Briacliff Manor
Davide Buscaldi	Université d'Orléans
Paula Carvalho	University of Lisbon, Faculty of Sciences, LASIGE
Jorge Civera	Universitat Politècnica de València
Paul Clough	University of Sheffield
Iria Da Cunha	Institut Universitari de Lingüística Aplicada - UPF
Víctor Darriba	University of Vigo
Patrick Drouin	Université de Montréal
Antonio Ferrández	Universidad de Alicante
Mikel Forkada	DLSI - Universitat d'Alacant
Atsushi Fujita	Future University Hakodate
Miguel Angel García	University of Jaen
Veronique Hoste	University College Ghent - Ghent University
Zornitsa Kozareva	Information Sciences Institute
Sobha L.	AU-KBC Research Centre
Gorka Labaka	University of the Basque Country
François Laureau	Macquarie University
Codrina Lauth	Fraunhofer Inst. for Intelligent Analysis and Information Systems
Els Lefever	University College Ghent - Ghent University
Antonia Martí	Universitat de Barcelona
Fernando Martínez	Universidad de Jaen
Raquel Martínez	UNED
Mikhail Mikhailov	University of Tampere
Manuel Montes-Y-Gómez	INAOE
Lidia Moreno	Universitat Politècnica de València
Roberto Paredes	Universitat Politècnica de València
David Pinto	Benemérita Universidad Autónoma de Puebla
Horacio Rodriguez	Universitat Politecnica de Catalunya
Paolo Rosso	Universitat Politècnica de València
Horacio Saggion	Universitat Pompeu Fabra
Luís Sarmento	Universidade do Porto
Grigori Sidorov	CIC-IPN
Aberto Simões	Universidade do Minho
Thamar Solorio	University of Alabama at Birmingham
Mariona Taulé	Universitat de Barcelona
Dan Tufis	Research Inst. for Artificial Intelligence, Romanian Academy
Marta Vila	Universitat de Barcelona
Jesús Vilares	Universidade da Coruña

Luís Villaseñor INAOE
Michael Zock CNRS-LIF

Table of Contents

Measuring Comparability of Multilingual Corpora Extracted from Wikipedia *

Midiendo la comparabilidad de copus multilingües extraídos de la Wikipedia

Pablo Gamallo Otero
Centro de Investigación en Tecnoloxías da Información (CITIUS), Universidade de Santiago de Compostela
Galiza, Spain
pablo.gamallo@usc.es

Issac González López
Cilenis S.L.
Language Engineering Solutions
Santiago de Compostela
Galiza, Spain
isaacjgonzalez@cilenis.com

Resumen: Los corpus comparables son muy útiles en variadas tareas del procesamiento del lenguaje tales como la extracción de léxicos bilingües. Con la mejora de la calidad de los corpus comparables, podemos mejorar la calidad de la extracción. Este artículo describe algunas estrategias para construir corpus comparables a partir de la Wikipedia, y propone una medida de comparabilidad. Fueron realizados algunos experimentos utilizando la Wikipedia portuguesa, española e inglesa.
Palabras clave: Extracción de Información, Corpus Comparables, Léxicos Bilingües, Comparabilidad

Abstract: Comparable corpora can be used for many linguistic tasks such as bilingual lexicon extraction. By improving the quality of comparable corpora, we improve the quality of the extraction. This article describes some strategies to build comparable corpora from Wikipedia and proposes a measure of comparability. Experiments were performed on Portuguese, Spanish, and English Wikipedia.
Keywords: Information Extraction, Comparable Corpora, Bilingual Lexicons, Comparability

1. Introduction

Wikipedia is a free, multilingual, and collaborative encyclopedia containing entries (called "articles") for almost 300 languages (281 in July 2011). English is the more representative one with about 3 million articles. However, Wikipedia is not a parallel corpus as their articles are not translations from one language into another. Many works have been published in the last years focused on its use and exploitation for multilingual tasks in natural language processing: extraction of bilingual dictionaries (Yu y Tsujii, 2009; Tyers y Pieanaar, 2008), alignment and machine translation (Adafre y de Rijke, 2006; Tomás, Bataller, y Casacuberta, 2001), multilingual information retrieval (Pottast, Stein, y Anderka, 2008). There also exists theoretical work analysing symmetries and asymmetries among the different multilingual versions of an entry/article in Wikipedia (Filatova, 2009).

In addition, multilingual articles of Wikipedia have been used as a source to build comparable corpora (Gamallo y González, 2010). The EAGLES - Expert Advisory Group on Language Engineering Standards Guidelines (see `http://www.ilc.pi.cnr.it/EAGLES96/browse.html`) defines a "comparable corpus" as one which selects *similar* texts in more than one language or variety. One of the main advantages of comparable corpora is their versatility to be used in many linguistic tasks (Maia, 2003), like bilingual lexicon extraction (Gamallo y Pichel, 2008; Saralegui, Vicente, y Gurrutxaga, 2008), information retrieval, and knowledge engineering. Besides, they can also be used as training corpus to improve statistic machi-

* This work has been supported by Ministerio de Educación y Ciencia of Spain, within the project OntoPedia, ref: FF12010-14986.

ne learning systems, in particular when parallel corpora are scarce for a given pair of languages. Another advantage concerns their availability. In contrast with parallel corpora, which require (not always available) translated texts, comparable corpora are easily retrieved from the web. Among the different web sources of comparable corpora, Wikipedia is likely the largest repository of similar texts in many languages. We only require the appropriate computational tools to make them comparable.

By taking into account multilingual potentialities of Wikipedia, our main objective is to define a method to measure the similarity (or degree of comparability) of different comparable corpora built from Wikipedia. For this purpose, first we describe some strategies to extract monolingual corpora in Portuguese, Spanish, and English from Wikipedia, by making use of some categories ("Archaeology", "Biology", "Physics", etc.) to make them comparable according to a specific topic. These strategies were described in detail in (Gamallo y González, 2010). Then, we propose a measure of comparability to verify whether the corpora are lowly or highly comparable. For many extraction tasks, such as bilingual lexicon extraction, using highly comparable corpora often leads to better results. There are some works proposing comparability measures between monolingual corpora (Li y Gaussier, 2010; Saralegui y Alegria, 2007), based on the use of existing bilingual dictionaries. However, instead of exploiting dictionaries to compute the comparability degree, we take advantage of the translation equivalents inserted in Wikipedia by means of *interlanguage links*.

This paper is organized as follows. Section 2 introduces two strategies to build comparable corpora from Wikipedia. Next, in Section 3, we propose some comparability measures. Then, Section 4 describe some experiments performed in order to measure the comparability between different corpora built using the strategies defined in Sec. 2 . The last section discusses future tasks that will be implemented in order to extend and improve our tools.

2. Two strategies to Build Wikipedia-Based Comparable Corpora

The input of our strategies is CorpusPedia[1], a friendly and easy-to-use XML structure, generated from Wikipedia dump files. In CorpusPedia, all the internal links found in the text are put in a vocabulary list identified with the tag *links*. In the same way, all the categories (or topics) used to classify each article are inserted in the tag *category*. In addition, there is a tag called *translations* which codifies a list of interlanguage links (i.e., links to the same articles in other languages) found in each article. Categories and translations are very useful features to build comparable corpora. Given these features, we developed two strategies aimed to extract corpora with different degrees of comparability.

Not-Aligned Corpus This strategy extracts those articles in two languages having in common the same topic, where the topic is represented by a category and its translation (for instance, the English-Spanish pair "Archaeology-Arqueología"). It results in a not-aligned comparable corpus, consisting of texts in two languages. We called it "not-aligned" because the version of an article in one language may have not its corresponding version in the other language.

Aligned Corpus The goal is to extract pairs of bilingual articles related by interlanguage links if, at least, one of both contains a required category. It results in a comparable corpus that is aligned article by article.

In Section 4, we will measure the degree of comparability of corpora built by means of these two strategies. Before that, we will define how to measure comparability between Wikipedia-based corpora.

3. Comparability Measures

For a comparable corpus $\mathcal{C}$ of Wikipedia articles, constituted for instance by a Portuguese part $\mathcal{C}_p$ and a Spanish part $\mathcal{C}_s$, a comparability coefficient can be defined on the basis

[1]The software to build CorpusPedia, as well as CorpusPedia files for English, French, Spanish, Portuguese, and Galician, are freely available at `http://gramatica.usc.es/pln/`

of finding, for each Portuguese term t_p in the vocabulary $\mathcal{C}_p^v$ of $\mathcal{C}_p$, its interlanguage link (or translation) in the vocabulary $\mathcal{C}_s^v$ of $\mathcal{C}_s$. The vocabulary of a Wikipedia corpus is the set of "internal links" found in that corpus. So, the two corpus parts, $\mathcal{C}_p$ and $\mathcal{C}_s$, tend to have a high degree of comparability if we find many internal links in $\mathcal{C}_p^v$ that can be translated (by means of interlanguage links) into many internal links in $\mathcal{C}_s^v$. Let $Trans_{bin}(t_p, \mathcal{C}_s^v)$ be a binary function which returns 1 if the translation of the Portuguese term t_p is found in the Spanish vocabulary $\mathcal{C}_s^v$. The binary Dice coefficient, $Dice_{bin}$, between two parts of a comparable corpus $\mathcal{C}$ is then defined as:

$$Dice_{bin}(\mathcal{C}_p, \mathcal{C}_s) = \frac{2 \sum_{t_p \in C_p^v} Trans_{bin}(t_p, \mathcal{C}_s^v)}{|\mathcal{C}_p^v| + |\mathcal{C}_s^v|}$$

We consider that it is not necessary to define the counterpart of the translation function, since the number of ambiguous terms is very low in Wikipedia, and most cases of ambiguity are solved with the so-called "disambiguated pages".

To avoid a bias towards common internal links, that is, towards those links occurring in most articles, we define a specific version of tf_idf weight for each term. In particular, $tf\_idf(t_p)$ is the frequency of term t_p in the Portuguese part of the comparable corpus, multiplied by its inverse *article* frequency in the whole Portuguese Wikipedia. By taking into account the tf_idf of terms, we can define a weighted measure of comparability. Let $Trans_{tf\_idf}(t_p, \mathcal{C}_s^v)$ be a function which returns the smallest value (min) of two tf_idf scores, both $tf\_idf(t_p)$ and $tf\_idf(t_s)$, where t_s is the Spanish translation of t_p in the Spanish part $\mathcal{C}_s$. The weighted Dice coefficient, $Dice_{tf\_idf}$, between two parts of a comparable corpus $\mathcal{C}$ is then defined as follows:

$$Dice_{tf\_idf}(\mathcal{C}_p, \mathcal{C}_s) = \frac{2 \sum_{t_p \in C_p^v} Trans_{tf\_idf}(t_p, \mathcal{C}_s^v)}{\sum_{t_p \in C_p^v} tf\_idf(t_p) + \sum_{t_s \in C_s^v} tf\_idf(t_s)}$$

The experiments described in the next section will be performed with the two comparability measures defined here.

4. Experiments and Results

Taking CorpusPedia as input source, we performed several experiments to build different comparable corpora for three language pairs, namely Portuguese-Spanish, Portuguese-English, and Spanish-English. These corpora were built using the two strategies described in Section 2 and five domain specific seed terms (in the three languages) considered as representative of five domain topics: "Archaeology", "Linguistics", "Physics", "Biology", and "Sport".

Table 1 shows the (binary and tf_idf) Dice scores obtained from measuring the comparability degree of 30 different comparable corpora. For each corpus, the table also shows the size (in Mb) of its two parts. In particular, the first column introduces the two languages of the corpus (pt = Portuguese, sp = Spanish, en = English) and the type of strategy (aligned or not aligned) used to build it. In the second and third columns, we show the two Dice scores. The forth column shows the size of the two parts of the corpus, and the last column contains the two seed terms employed to generate the corpus. In Table 2, we show the Dice scores as well as the size of nine pairs of monolingual corpora randomly generated from Wikipedia.

We can observe first that there are significant differences in terms of comparability between the Dice scores in Table 1 and those obtained from the randomly generated monolingual pairs in Table 2. It follows that corpora built by means of our strategies (not aligned and aligned) are actually *comparable*. Then, we should note that in the comparable corpora of Table 1, the Dice scores based on tf_idf are about 70 % higher than those based on the binary function. By contrast, in randomly generated corpora (Table 2), there are no significant differences between $Dice_{bin}$ and $Dice_{td\_idf}$. It means that our tf_idf makes the Dice similarity score higher if the two evaluated corpus parts are actually comparable.

As it was expected, not-aligned corpora tend to be larger than the aligned ones. However, if we just compare the smallest parts of each corpus, the differences are not very important: the smallest parts of not-aligned corpora are only 15 % larger than those of aligned corpora. This is in accordance with the fact that aligned corpora are more balanced in terms of size, since no part is much larger than the other one. As far the corpus size is concerned, let us note that, in average, English parts are clearly larger than the Spanish ones, which are slightly larger than the Portuguese ones. In general, English ar-

Corpora	Dice (bin)	Dice (tf-idf)	Size (in Mb)	Seed terms
pt-sp (not aligned)	.068	.086	0.6Mb/3.4Mb	Arqueologia, Arqueología
pt-en (not aligned)	.041	.067	0.6Mb/8.4Mb	Arqueologia, Archaeology
sp-en (not aligned)	.090	.140	0.4Mb/8.4Mb	Arqueología, Archaeology
pt-sp (aligned)	.179	.199	0.4Mb/0.2Mb	Arqueologia, Arqueología
pt-en (aligned)	.127	.140	0.4Mb/1.1Mb	Arqueologia, Archaeology
sp-en (aligned)	.181	.226	2.0Mb/2.9Mb	Arqueología, Archaeology
pt-sp (not aligned)	.078	.129	0.8Mb/1.7Mb	Linguística, Lingüística
pt-en (not aligned)	.054	.136	0.8Mb/5.1Mb	Linguística, Linguistics
sp-en (not aligned)	.074	.170	1.7Mb/5.1Mb	Lingüística, Linguistics
pt-sp (aligned)	.140	.214	0.6Mb/0.8Mb	Linguística, Lingüística
pt-en (aligned)	.128	.194	0.5Mb/1.2Mb	Linguística, Linguistics
sp-en (aligned)	.150	.257	0.9Mb/1.7Mb	Lingüística, Linguistics
pt-sp (not aligned)	.200	.374	4.4Mb/4.8Mb	Física, Física
pt-en (not aligned)	.123	.287	4.4Mb/12Mb	Física, Physics
sp-en (not aligned)	.270	.403	4.8Mb/12Mb	Física, Physics
pt-sp (aligned)	.237	.390	3.6Mb/4.7Mb	Física, Física
pt-en (aligned)	.178	.348	3.8Mb/11Mb	Física, Physics
sp-en (aligned)	.220	.387	3.4Mb/7.6Mb	Física, Physics
pt-sp (not aligned)	.130	.227	2.4Mb/1.5Mb	Biologia, Biología
pt-en (not aligned)	.102	.193	2.4Mb/9.4Mb	Biologia, Biology
sp-en (not aligned)	.068	.129	1.5Mb/9.4Mb	Biología, Biology
pt-sp (aligned)	.197	.328	1.6Mb/2.8Mb	Biologia, Biología
pt-en (aligned)	.186	.308	1.8Mb/4.5Mb	Biologia, Biology
sp-en (aligned)	.213	.294	0.9Mb/1.3Mb	Biología, Biology
pt-sp (not aligned)	.083	.148	11Mb/35Mb	Desporto, Deporte
pt-en (not aligned)	.026	.085	11Mb/333Mb	Desporto, Sport
sp-en (not aligned)	.047	.136	35Mb/333Mb	Deporte, Sport
pt-sp (aligned)	.175	.266	9.7Mb/15Mb	Desporto, Deporte
pt-en (aligned)	.189	.334	11Mb/20Mb	Desporto, Sport
sp-en (aligned)	.206	.290	20Mb/29Mb	Deporte, Sport
pt-sp (not aligned)	.111	.192	3.8Mb/9.3Mb	**Overall**
pt-en (not aligned)	.069	.153	3.8Mb/73Mb	**Overall**
sp-en (not aligned)	.109	.195	9.3Mb/73Mb	**Overall**
pt-sp (aligned)	.185	.279	3.2Mb/4.7Mb	**Overall**
pt-en (aligned)	.161	.264	3.5Mb/7.6Mb	**Overall**
sp-en (aligned)	.194	.290	6.2Mb/8.5Mb	**Overall**

Cuadro 1: Dice similarity between several comparable corpora in Portuguese, Spanish, and English.

Corpora	Dice (bin)	Dice (tf-idf)	Size (in Mb)
pt-sp1 (random)	.012	.012	2.2Mb/0.9Mb
pt-en1 (random)	.003	.003	2.2Mb/0.4Mb
sp-en1 (random)	.003	.003	0.9Mb/0.4Mb
pt-sp2 (random)	.016	.014	1.5Mb/3.0Mb
pt-en2 (random)	.017	.014	1.5Mb/42Mb
sp-en2 (random)	.017	.015	3.0Mb/42Mb
pt-sp3 (random)	.008	.006	0.2Mb/0.5Mb
pt-en3 (random)	.001	.001	0.2Mb/1.4Mb
sp-en3 (random)	.005	.005	0.5Mb/1.4Mb

Cuadro 2: Dice similarity between randomly generated pairs of monolingual corpora.

ticles tend to have more words than Spanish and Portuguese articles. As it was suggested by one of the reviewers of the article, one of the reasons for the difference in size in the case of aligned corpora is that Spanish and Portuguese entries seem to be summaries of the English ones. So, to increase comparability between an aligned pair of articles, the longer article could be shortened by removing those parts which are not present in the other language, obtaining, this way, a more comparable pair of articles.

Finally, as it was expected, aligned corpora are significantly more comparable (i.e., higher Dice coefficient) than not-aligned corpora. In average, $Dice_{td\_idf}$ increases 80 % the comparability of aligned-corpora with regard to not-aligned ones. So, considering that aligned corpora only decreases 15 % in size in relation to not-aligned corpora, we can conclude that the aligned strategy seems to be more appropriate to build comparable corpora from Wikipedia.

5. Conclusions and Future Work

The emergence of multilingual resources, such a Wikipedia, makes it possible to design new methods and strategies to compile corpus from the web, methods that are more efficient and powerful than the traditional ones. In particular, the semi-structured information underlying Wikipedia turns out to be very useful to build comparable corpora. In this article, we proposed two strategies to build comparable corpora from Wikipedia and a way to measure their degree of comparability. The experiments led us to conclude that corpora aligned article by article are more comparable than not aligned corpora. Besides, they consist of two balanced corpus parts in terms of size. Finally, they are not much smaller than not aligned corpora.

In future work, we will be focused on how to improve the strategies to build comparable corpora by extending coverage (more articles) without losing comparability. For this purpose, we will test and evaluate techniques to expand categories using a list of similar terms identified as hyponyms or co-hyponyms of the source category. In order to find hyponyms and co-hyponyms of a term, it will be required to build an ontology of categories using the semi-structured information of Wikipedia (Chernov et al., 2006; Ponzetto y Navigli, 2009; de Melo y Weikum, 2010). On the other hand, we will evaluate comparability in an indirect way. In particular, we will use the generated corpora on tasks requiring comparable corpora as input (e.g., bilingual lexicon extraction). The better the extracted lexicon, the more comparable the input corpus should be. Finally, we believe that our method for aligning pairs of articles could be useful for related tasks, such as Wikipedia infoboxes alignment in different languagues (Adar, Skinner, y Weld, 2009).

Bibliografía

Adafre, S.F. y M. de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. En *11th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 62–69.

Adar, Eytan, Michael Skinner, y Daniel S. Weld. 2009. Information arbitrage across multi-lingual wikipedia. En *Second ACM International Conference on Web Search and Data Mining , WSDM*.

Chernov, Sergey, Tereza Iofciu, Wolfgang Nejdl, y Xuan Zhou. 2006. Extracting semantic relationships between wikipedia categories. En *SemWiki2006 - From Wiki to Semantics*, Budva, Montenegro.

de Melo, Gerard y Gerhard Weikum. 2010. Menta: inducing multilingual taxonomies from wikipedia. En *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, páginas 1099–1108.

Filatova, Elena. 2009. Directions for Exploiting Asymmetries in Multilingual Wikipedia. En *CLEAWS3*, páginas 30–37, Colorado.

Gamallo, Pablo y Isaac González. 2010. Wikipedia as a multilingual source of comparable corpora. En *LREC 2010 Workshop on Building and Using Comparable Corpora*, páginas 19–26, Valeta, Malta.

Gamallo, Pablo y José Ramom Pichel. 2008. Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary. *LNCS*, 4919:413–423.

Li, Bo y Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. En

20th International Conference on Computational Linguistics (COLING 2010, páginas 644–652.

Maia, Belinda. 2003. What Are Comparable Corpora. En *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, páginas 27–34, Lancaster, UK.

Ponzetto, Simone Paolo y Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. En *Proceedings of the 21st international jont conference on Artifical intelligence*, páginas 2083–2088.

Pottast, M., B. Stein, y M. Anderka. 2008. A wikipedia-based multilingual retrieval model. En *Advances in Information Retrieval*, páginas 522–530.

Saralegui, X. y I. Alegria. 2007. Similitud entre documentos multilíngües de carácter científico-técnico en un entorno Web. En *Procesamiento del Lenguaje Natural*, página 39.

Saralegui, X., I. San Vicente, y A. Gurrutxaga. 2008. Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. En *LREC 2008 Workshop on Building and Using Comparable Corpora*.

Tomás, J., J. Bataller, y F. Casacuberta. 2001. Mining Wikipedia as a Parallel and Comparable Corpus. En *Language Forum*, volumen 1, página 34.

Tyers, M.F. y J.A. Pieanaar. 2008. Extracting Bilingual Word Pairs from Wikipedia. En *LREC 2008, SALTMIL Workshop*, Marrakesh, Marocco.

Yu, Kun y Junichi Tsujii. 2009. Bilingual dictionary extraction from wikipedia. En *Machine Translation Summit XII*, Ottawa, Canada.

Extracción de corpus paralelos de la Wikipedia basada en la obtención de alineamientos bilingües a nivel de frase*

Extracting Parallel Corpora from Wikipedia on the basis of Phrase Level Bilingual Alignment

Joan Albert Silvestre-Cerdà, Mercedes García-Martínez, Alberto Barrón-Cedeño, Jorge Civera y Paolo Rosso
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
mgarcia@iti.upv.es,{jsilvestre,lbarron,jcivera,prosso}@dsic.upv.es

Resumen: Este artículo presenta una nueva técnica de extracción de corpus paralelos de la Wikipedia mediante la aplicación de técnicas de traducción automática estadística. En concreto, se han utilizado los modelos de alineamiento basados en palabras de IBM para obtener alineamientos bilingües a nivel de frase entre pares de documentos. Para su evaluación se ha generado manualmente un conjunto de test formado por pares de documentos inglés-español, obteniéndose resultados prometedores.
Palabras clave: corpus comparables, extracción de oraciones paralelas, traducción automática estadística

Abstract: This paper presents a proposal for extracting parallel corpora from Wikipedia on the basis of statistical machine translation techniques. We have used word-level alignment models from IBM in order to obtain phrase-level bilingual alignments between documents pairs. We have manually annotated a set of test English-Spanish comparable documents in order to evaluate the model. The obtained results are encouraging.
Keywords: comparable corpora, parallel sentences extraction, statistical machine translation

1. Introducción

La extracción automática de corpus paralelos a partir de recursos textuales multilingües es, hoy por hoy, una tarea de especial interés debido al creciente auge de la traducción automática estadística. La web es una fuente inmensa de documentos en múltiples lenguas que tiene muchas posibilidades de explotación. No obstante, encontrar frases paralelas a nivel global en la web es una tarea muy dispersa y extremadamente difícil, aunque no imposible (Uszkoreit et al., 2010).

La Wikipedia es uno de los pocos recursos web que nos provee de forma explícita gran cantidad de textos multilingües comparables, pues sus contenidos se presentan como artículos en múltiples idiomas que describen un mismo concepto. El objetivo es, pues, explotar los contenidos comparables de dichos documentos con la finalidad de extraer frases paralelas que puedan ser utilizadas en el entrenamiento de sistemas de traducción automática.

En este trabajo se propone una aproximación heurística a la extracción de corpus paralelos de la Wikipedia basada en técnicas de Traducción Automática Estadística (TAE).

* Este trabajo se ha llevado a cabo en el marco del VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems, financiado parcialmente por parte de la EC (FEDER/FSE; WIQEI IRSES no. 269180 / FP 7 Marie Curie People), por el MICINN como parte del proyecto Text-Enterprise 2.0 (TIN2009-13391-C04-03) en el Plan I+D+i, y por la beca 192021 del CONACyT. También ha recibido apoyo por parte del EC (FEDER/FSE) y del MEC/MICINN bajo el programa MIPRCV "Consolider Ingenio 2010" (CSD2007-00018) y el proyecto iTrans2 (TIN2009-14511), por el MITyC en el marco del proyecto erudito.com (TSI-020110-2009-439), por la Generalitat Valenciana con las ayudas Prometeo/2009/014 y GV/2010/067, y por el "Vicerrectorado de Investigación de la UPV" con la ayuda 20091027.

En la siguiente sección analizaremos los trabajos previos que han servido de inspiración a este trabajo. Posteriormente, en la Sección 3 se describe ampliamente el sistema propuesto. La Sección 4 muestra los resultados experimentales y finalmente, una serie de conclusiones son expuestas en la Sección 5.

2. Trabajos relacionados

Debido a su creciente necesidad e importancia, la extracción automática de corpus paralelos es una tarea bastante explorada en la actualidad, aunque los primeros trabajos se realizaron hace ya más de dos décadas (Brown, Lai, y Mercer, 1991; Gale y Church, 1991), si bien éstos se ceñían a encontrar alineamientos entre frases en textos paralelos. Estos trabajos proponen métodos de alineamiento muy rápidos pero poco precisos, pues para detectar relaciones entre frases utilizaban únicamente la información de longitud de las oraciones. Posteriormente, Chen propuso utilizar información léxica mediante un sencillo modelo de traducción estadístico basado en palabras, demostrando una mejora significativa de la calidad de los alineamientos extraídos (Chen, 1993), y unos años más tarde, Moore combinó ambas aproximaciones (Moore, 2002). Más recientemente, González propuso un modelo de alineamiento entre frases y palabras inspirado en el modelo 1 de IBM (González-Rubio et al., 2008).

Con el problema de alinear frases en textos paralelos bien estudiado, y ante la creciente demanda de corpus paralelos para TAE, los principales esfuerzos se centraron en la extracción de corpus paralelos (Eisele y Xu, 2010; Uszkoreit et al., 2010; Varga et al., 2005), en incluso monolingües (Barzilay y Elhadad, 2003; Quirk, Brockett, y Dolan, 2004), a partir de la web. En éste ámbito, la Wikipedia ha sido un recurso bastante explotado, presentándose una gran variedad de aproximaciones, desde métodos heurísticos (Adafre y de Rijke, 2006; Mohammadi y GhasemAghaee, 2010) hasta aproximaciones basadas en clasificación estadística utilizando combinaciones lineales de características (Smith, Quirk, y Toutanova, 2010; Tomás et al., 2008). También se han llevado a cabo algunos trabajos en la vertiente monolingüe (Yasuda y Sumita, 2008). Ahora bien, ninguno de los trabajos previos ha explorado la utilización de modelos de traducción estadísticos como sistemas de evaluación de alineamientos en recursos comparables como la Wikipedia, y es precisamente este vacío experimental el que se pretende cubrir en este trabajo.

3. Descripción del sistema

Para la tarea de extracción de corpus paralelos de la Wikipedia consideraremos pares de documentos de Wikipedia $X = (x_1, \ldots, x_j, \ldots, x_{|X|}) \in \mathcal{X}^*$ e $Y = (y_1, \ldots, y_i, \ldots, y_{|Y|}) \in \mathcal{Y}^*$ que representen un mismo concepto, siendo x_j la j-ésima frase del documento X, y_i la i-ésima frase del documento Y, y $\mathcal{X}$ e $\mathcal{Y}$ los vocabularios de los lenguajes en los que se encuentran los respectivos documentos. Definimos (x_j, y_i) como un alineamiento entre la j-ésima frase del documento X y la i-ésima frase del documento Y, y A un conjunto finito de alineamientos.

Inicialmente asumiremos que $A = (X \times Y)$, es decir, el conjunto A contiene todo alineamiento posible entre las frases de X y de Y. La probabilidad de cada alineamiento $(x_j, y_i) \in A$ se calcula de acuerdo con el modelo 4 de IBM (Brown y others, 1993), que es un modelo de alineamiento a nivel de palabra ampliamente utilizado en Traducción Automática Estadística. Un alineamiento recibirá una probabilidad alta si el grado de coocurrencia de las palabras que componen las frases es alto, pero por contra recibirá una probabilidad baja si las palabras involucradas tienen poca o ninguna correlación. Cabe decir que las puntuaciones otorgadas por los modelos de IBM provienen de una serie de productos de probabilidades, tantos como el número de palabras que conforman la frase de destino y_i, por lo que dicha puntuación debe ser normalizada convenientemente para que no sea dependiente de la longitud. De no ser así, los alineamientos con frases destino y_i de menor número de palabras tenderían a ser más probables, pudiendo darse casos de alineamientos (x_j, y_i) con altos valores de probabilidad con $|x_j| = 8$ e $|y_i| = 1$, por ejemplo.

Una vez se han evaluado todos los alineamientos del conjunto A, se obtiene el conjunto de alineamientos más probables $B \subseteq A$ mediante la siguiente maximización:

$$(x_j, y_i) \in B \,/\, p_{IBM}(x_j \mid y_i) > p_{IBM}(x_j \mid y_{i'}) \quad (1)$$
$$\forall i' = 1 \ldots |Y| \quad \forall j = 1 \ldots |X|$$

Es decir, para cada frase x_j del documento X, conservaremos el alineamiento (x_j, y_i)

que maximice la probabilidad del modelo 4 de IBM para toda posible frase y_i. Esto implica añadir una restricción importante en el proceso de alineamiento, pero que no obstante nos permite definir un sistema base o inicial que tenemos previsto mejorar en el futuro mediante el cálculo y la posterior combinación de los alineamientos en ambas direcciones.

Por último, se genera el conjunto final de alineamientos filtrados $C \subseteq B$, formado por aquellos alineamientos cuya puntuación supere un cierto umbral α, es decir:

$$(x_j, y_i) \in C \,/\, p_{IBM}(x_j \mid y_i) > \alpha \qquad (2)$$

El umbral α puede interpretarse como un parámetro que afecta a la calidad de los alineamientos extraídos, ya que cuanto mayor es el umbral, mayor es nuestra exigencia sobre el sistema, extrayéndose en consecuencia un menor número de alineamientos. En la Sección 4 estudiaremos la influencia de este parámetro en las prestaciones de nuestro sistema.

4. Experimentación

Con el objetivo de evaluar las prestaciones que ofrece nuestro método de extracción de corpus paralelos de la Wikipedia, hemos realizado un estudio experimental en el que se evalúa la calidad de los pares de frases extraídos automáticamente por nuestro sistema a partir de un conjunto de prueba que tuvimos que generar de forma manual, debido a la inexistencia de corpus adecuadamente etiquetados para esta tarea. La generación de dicho conjunto, formado por pares de documentos de la Wikipedia en inglés y español, es detallada en las Secciones 4.1 y 4.2.

El modelo 4 de IBM fue entrenado con MGIZA, un software basado en el popular GIZA++ que nos ofrece la posibilidad de evaluar un conjunto de prueba con los modelos ya entrenados, además de que permite realizar un entrenamiento paralelo de los mismos. Con el fin de minimizar los problemas relacionados con las palabras fuera de vocabulario y generalizar el dominio del sistema, los modelos de IBM se entrenaron con un subconjunto de pares de frases, definido en (Sanchis-Trilles et al., 2010), de tres corpus de referencia en el área de la Traducción Automática Estadística: Europarl-v5 (Koehn, 2005), News-Commentary y United Nations (Rafalovitch y Dale, 2009). Las estadísticas de este subconjunto pueden ser consultadas en la Tabla 1. Cabe destacar la gran cantidad de pares de frases empleados para el entrenamiento de los modelos, así como el considerable tamaño de los vocabularios de cada una de las lenguas.

Tabla 1: Estadísticas básicas del corpus empleado para el entrenamiento de los modelos IBM.

	Entrenamiento	
Idioma	En	Es
Número de frases	2.8M	
Tamaño Vocabulario	118K	164K
Número Total Palabras	54M	58M

El resto de esta sección se estructura como sigue: la Sección 4.1 muestra el procedimiento de extracción de documentos y su preproceso. Posteriormente, las Secciones 4.2 y 4.3 presentan la metodología de etiquetado y las métricas de evaluación empleadas, respectivamente. Finalmente, la Sección 4.4 expone los resultados obtenidos al evaluar el conjunto de entrenamiento generado manualmente.

4.1. Selección de documentos y preproceso

La Wikipedia alberga miles de artículos disponibles en inglés y español, y abarcan un dominio extremadamente amplio. Por ese motivo, y con el objetivo de realizar una prueba optimista con el sistema, se realizó una selección de pares de documentos cuyos dominios se asemejaran al dominio del corpus empleado en el entrenamiento del modelo de alineamiento. En concreto, se seleccionaron un total de 15 pares de documentos inglés-español relacionados con la economía y procesos administrativos de la Unión Europea. De dichos documentos se extrajo el texto plano, que posteriormente fue sometido a un preproceso consistente en la separación de frases en líneas (sentence-splitting), aislamiento de palabras y signos de puntuación (tokenizing) y conversión a minúsculas (lowercasing). Las estadísticas de dicho corpus después de ser sometido a este preproceso se muestran en la Tabla 2.

4.2. Metodología de etiquetado

A continuación se describe la metodología seguida para generar el conjunto de evalua-

Tabla 2: Estadísticas básicas del conjunto de evaluación construido de forma manual.

	Evaluación	
Idioma	En	Es
Número de documentos	15	
Número de frases	661	341
Alineamientos posibles	22680	
Tamaño Vocabulario	3,4K	2,8K
Número Total Palabras	24,5K	16,2K

ción etiquetado, partiendo de un conjunto de pares de documentos previamente preprocesados. Esta metodología está inspirada en (Och y Ney, 2003), pero tomando alineamientos entre frases en lugar de alineamientos entre palabras.

Dos personas se encargaron de etiquetar manualmente y independientemente todo el conjunto de pares de documentos. Se les pidió que anotaran aquellos alineamientos, de entre todos los posibles para cada par de documentos, que guardaran una relación de paralelismo.

Adicionalmente, los etiquetadores fueron instruidos para que asignaran cada uno de los alineamientos a uno de los siguientes dos conjuntos:

- P: Conjunto de alineamientos probables. Definen alineamientos entre frases que conforman traducciones similares, aunque no exactas, en las que se expresa la misma idea semántica, o bien para indicar que un determinado alineamiento forma parte de una relación 1-a-muchos o muchos-a-1.
- S: Conjunto de alineamientos seguros, siendo $S \subseteq P$. Define alineamientos entre frases que son traducciones exactas o casi exactas (paralelas).

En este contexto, el etiquetador 1 genera los conjuntos S_1 y P_1, mientras que el etiquetador 2 genera S_2 y P_2. Entonces, los conjuntos S_1, P_1, e S_2, P_2 se combinan en S y P de la siguiente forma:

$$S = S_1 \cap S_2$$
$$P = P_1 \cup P_2$$

El conjunto P (que incluye S) representa los pares de frases que deberían ser extraídos por el sistema, y por tanto son tomados como referencia para la tarea. Para el caso concreto de este corpus, el conjunto S está formado por 10 alineamientos, mientras que el conjunto P engloba un total de 115 alineamientos.

4.3. Medidas de Evaluación

La evaluación de la calidad del conjunto filtrado de alineamientos C obtenido de forma automática mediante nuestro sistema se ha realizado mediante la métrica Sentence Alignment Error Rate, claramente inspirada en la presentada en (Och y Ney, 2003).

Dado un par de documentos X e Y, los conjuntos de alineamientos entre ambos documentos S y P etiquetados manualmente, y el conjunto filtrado de alineamientos C, se define la métrica Sentence Alignment Error Rate (SAER) como sigue:

$$SAER(S, P, C) = 1 - \frac{|C \cap S| + |C \cap P|}{|C| + |S|} \quad (3)$$

Al igual que (Och y Ney, 2003), también hemos empleado las medidas de cobertura y precisión para obtener más información acerca de las prestaciones del sistema:

$$\text{Cobertura} = \frac{|C \cap S|}{|S|}, \ \text{Precisión} = \frac{|C \cap P|}{|C|} \quad (4)$$

4.4. Resultados

En la presente sección se presentan los resultados de las pruebas experimentales llevadas a cabo con nuestro sistema, utilizando el conjunto de evaluación generado de forma manual. En la Sección 3 hemos resaltado la necesidad de estudiar la influencia del parámetro α, puesto que radica directamente en la calidad de la frases extraídas. Un valor alto para dicho umbral puede conllevar a que el sistema no sea capaz de extraer ningún alineamiento. Por contra, un valor pequeño de α se traduciría en la extracción de un gran número de pares de frases, e idealmente en un aumento del número de alineamientos correctos (Verdaderos Positivos, VP), aunque hay que tener en cuenta que el número de casos de Falsos Positivos (FP), es decir, alineamientos que no existen en la referencia, aumenta generalmente en mayor proporción que los VPs. La clave está pues en encontrar un valor de α que garantice la obtención de la mayor proporción posible de Verdaderos Positivos (VPR) y que minimice el ratio de Falsos Positivos (FPR). Ambas proporciones se calculan de la siguiente forma:

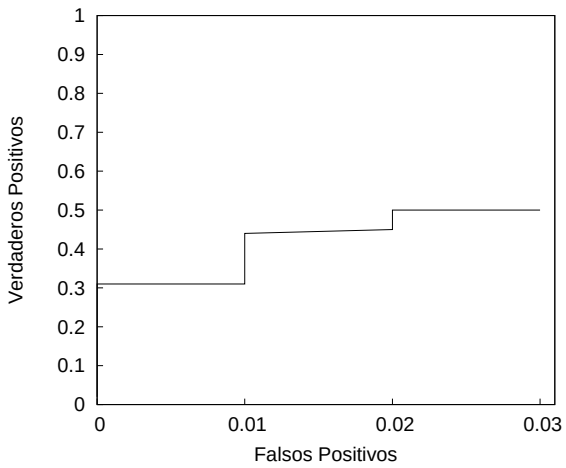


Figura 1: Curva ROC para constatar la relación entre Verdaderos Positivos y Falsos positivos en función del parámetro α.

$$VPR = \frac{VP}{P} = \frac{VP}{VP + FN} \quad (5)$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + VN} \quad (6)$$

donde P representa el número de muestras positivas, que es igual al número de casos de Verdaderos Positivos (VP) más el número de casos de Falsos Negativos (FN), mientras que N representa el número de muestras negativas, que es igual al número de casos de Falsos Positivos (FP) más el número de casos de Verdaderos Negativos (VN).

Con esta finalidad, hemos realizado una exploración exhaustiva del parámetro α, y posteriormente hemos dibujado una curva ROC, mostrada en la Figura 1, en la que se observa la relación entre los Verdaderos Positivos (VPR, eje vertical) y los Falsos Positivos (FPR, eje horizontal) en función del umbral α, cuyo valor es inversamente proporcional al desplazamiento de ambos ejes. Cabe decir que dicha exploración debería de haberse llevado a cabo mediante un conjunto de desarrollo, pero debido a la ausencia del mismo tuvimos que emplear el conjunto de evaluación. En el futuro planeamos ampliar dicho corpus para poder generar un conjunto de desarrollo.

De la Figura 1 cabe destacar varias cosas. En primer lugar, la gráfica tiene un aspecto degenerado debido a que la proporción relativa de Falsos Positivos nunca podrá llegar a valer 1, puesto que está acotada superiormente por $FP/(FP + VN)$ teniendo en cuenta que $FP \leq |X|$ (como máximo se darán lugar tantos FPs como número de frases del documento de entrada) y que $VN \leq |X \times Y|$ (el sistema puede llegar a descartar el conjunto de todos los posibles alineamientos), por lo que el valor del cociente será muy pequeño. En segundo lugar, podemos observar que para valores más altos del umbral α la relación de Falsos Positivos llega a ser casi cero para un ratio del 0.3 de Verdaderos Positivos, mientras que para valores de α más pequeños podemos llegar a conseguir un 0.5 de VPR con un ratio del 0.02 de FPR. En términos relativos, este segundo punto parece ser el óptimo, pero si tomamos en cuenta los valores absolutos, nos encontramos con diferencias del orden de centenares de FPs. Es por este motivo por el cual nos decantaremos por el primer de ellos, con $\alpha = 1{,}1 \cdot 10^{-3}$.

En la Tabla 3 se muestran los valores de las métricas, presentadas en la Sección 4.3, tras la evaluación del conjunto de prueba, además de otras estadísticas de interés, para el valor del umbral que hemos considerado como óptimo ($\alpha = 1{,}1 \cdot 10^{-3}$) y para dos casos extremos, con el objetivo de apreciar más notoriamente la influencia de dicho parámetro en las prestaciones del sistema. La primera fila muestra el tamaño del conjunto de alineamientos filtrados C, mientras que las cuatro filas siguientes muestran el número de muestras clasificadas como Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN). Por último, se muestran los valores de las tres métricas empleadas para evaluar las prestaciones del sistema: cobertura, precisión y SAER.

En ella se puede ver como, a pesar de la simplicidad de nuestro planteamiento, se obtienen unos resultados bastante aceptables para el valor óptimo de α, con una tasa del 0.36 de error de alineamiento, un 0.59 de grado de precisión, y sobretodo un 0.9 de cobertura, aunque cabe decir que esta última no es una medida fiable dado que en el corpus sólo existen 10 alineamientos etiquetados como seguros. A continuación se muestran algunos ejemplos de los pares de frases extraídos por nuestro sistema:

En: On 20 april 2005, the European Commission adopted the communication on Kosovo to the council “a european futu-

Tabla 3: Resultados del sistema para el conjunto de test generado manualmente, con $\alpha = \{1 \cdot 10^{-4}, 1{,}1 \cdot 10^{-3}, 5 \cdot 10^{-2}\}$.

	$\alpha = 1 \cdot 10^{-4}$	$\alpha = 1{,}1 \cdot 10^{-3}$	$\alpha = 5 \cdot 10^{-2}$
$\|C\|$	656	59	4
VP	58	35	2
VN	21967	22541	22563
FP	598	24	2
FN	57	80	113
Cobertura	1,00	0,90	0,1
Precisión	0,09	0,59	0,50
SAER	0,90	0,36	0,79

re for Kosovo" which reinforces the commission's commitment to Kosovo.

Es: El 20 de abril de 2005, la Comisión Europea adoptó la comunicación sobre kosovo en el consejo "un futuro europeo para Kosovo" que refuerza el compromiso de la comisión con Kosovo.

En: He added that the decisive factor would be the future and the size of the eurozone, especially whether Denmark, Sweden and the UK would have adopted the euro or not.

Es: Añadió que el factor decisivo será el futuro y el tamaño de la zona del euro, especialmente si Dinamarca, Suecia y el Reino Unido se unen al euro o no.

En: Montenegro officially applied to join the EU on 15 december 2008.

Es: Oficialmente, Montenegro pidió el acceso a la UE el 15 de diciembre de 2008.

Si observamos nuevamente la Tabla 3 y nos fijamos en las diferencias existentes entre el caso óptimo y los casos extremos, se pueden extraer algunas conclusiones interesantes. Para $\alpha = 1 \cdot 10^{-4}$ no se filtra ningún alineamiento, esto es, $C = B$, y por tanto nos damos cuenta que nuestro sistema nunca será capaz de encontrar 57 alineamientos que sí están en la referencia. Para evitar esta severa limitación tenemos previsto obtener los alineamientos entre frases en ambos sentidos (X a Y, e Y a X), y posteriormente aplicar un algoritmo heurístico inspirado en (Och y Ney, 2003) que los combine, partiendo de la intersección entre ambos alineamientos y añadiendo alineamientos adicionales. Esto nos llevará, en primer lugar, a obtener alineamientos más robustos, y en segundo lugar, a capturar relaciones entre frases de muchas-a-1, 1-a-muchas, e incluso muchas-a-muchas.

5. Conclusiones y Trabajo Futuro

En este trabajo hemos presentado una aproximación heurística alternativa a las ya existentes para la extracción automática de corpus paralelos a partir de los contenidos multilingües comparables que ofrece la Wikipedia. La evaluación experimental ha mostrado unos resultados francamente prometedores para nuestro sistema inicial. Como extensión de este trabajo planeamos obtener de forma heurística los alineamientos entre frases en ambas direcciones con el objetivo de mejorar la calidad del sistema, una mejora que creemos que será sustancial. Otra alternativa de cara al futuro sería emplear la variante del modelo 1 de IBM presentada en (González-Rubio et al., 2008) en esta tarea, ya que nos permitiría obtener los alineamientos bidireccionales de forma no heurística mediante un entrenamiento Expectation-Maximization (Dempster, Laird, y Rubin, 1977). Con la implementación de estas mejoras, realizaremos un estudio comparativo de nuestro sistema con otros sistemas del estado del arte.

Cabe destacar, además, que en este trabajo hemos adaptado una metodología existente para la evaluación de alineamientos a nivel de frase. Para ello, hemos definido una metodología de etiquetado adecuada para generar un conjunto de evaluación, así como una serie de métricas para cuantificar las prestaciones del sistema. Como trabajo futuro pre-

tendemos aumentar el tamaño del corpus y el número de anotadores, con el fin de hacer más robusto el proceso de etiquetado manual de los alineamientos.

Bibliografía

Adafre, S. F. y M. de Rijke. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 62–69.

Barzilay, Regina y Noemie Elhadad. 2003. Sentence Alignment for Monolingual Comparable Corpora. En *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, páginas 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brown, P. F. y others. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Brown, Peter F., Jennifer C. Lai, y Robert L. Mercer. 1991. Aligning Sentences in Parallel Corpora. En *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, páginas 169–176, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chen, Stanley F. 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. En *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, páginas 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dempster, A. P., N. M. Laird, y D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Statistical Society. Series B*, 39(1):1–38.

Eisele, Andreas y Jia Xu. 2010. Improving Machine Translation Performance using Comparable Corpora. En *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora LREC 2010*, páginas 35–41. ELRA.

Gale, William A. y Kenneth W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. En *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, páginas 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

González-Rubio, Jesús, Germán Sanchis-Trilles, Alfons Juan, y Francisco Casacuberta. 2008. A Novel Alignment Model Inspired on IBM Model 1. En *Proceedings of the 12th conference of the European Association for Machine Translation*, páginas 47–56.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *Proc. of the MT Summit X*, páginas 79–86, September.

Mohammadi, M. y N. GhasemAghaee. 2010. Building Bilingual Parallel Corpora Based on Wikipedia. En *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, volumen 2, páginas 264 –268, march.

Moore, Robert C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. En *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, páginas 135–144, London, UK, UK. Springer-Verlag.

Och, Franz Josef y Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51, March.

Quirk, Chris, Chris Brockett, y William Dolan. 2004. Monolingual Machine Translation for Paraphrase Generation. En *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, páginas 142–149.

Rafalovitch, Alexandre y Robert Dale. 2009. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus.

Sanchis-Trilles, Germán, Jesús Andrés-Ferrer, Guillem Gascó, Jesús González-Rubio, Pascual Martínez-Gómez, Martha-Alicia Rocha, Joan-Andreu Sánchez, y Francisco Casacuberta. 2010. UPV-PRHLT English–Spanish System for WMT10. En *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, páginas

172–176, Uppsala, Sweden, July. Association for Computational Linguistics.

Smith, Jason R., Chris Quirk, y Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. En *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, páginas 403–411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomás, Jesús, Jordi Bataller, Francisco Casacuberta, y Jaime Lloret. 2008. Mining Wikipedia as a Parallel and Comparable Corpus. *LANGUAGE FORUM*, 34(1). Article presented at CICLing-2008, 9th International Conference on Intelligent Text Processing and Computational Linguistics, February 17 to 23, 2008, Haifa, Israel.

Uszkoreit, Jakob, Jay M. Ponte, Ashok C. Popat, y Moshe Dubiner. 2010. Large Scale Parallel Document Mining for Machine Translation. En *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, páginas 1101–1109, Stroudsburg, PA, USA. Association for Computational Linguistics.

Varga, Dániel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, y Viktor Nagy. 2005. Parallel Corpora for Medium Density Languages. En *Proceedings of the RANLP 2005*, páginas 590–596.

Yasuda, Keiji y Eiichiro Sumita. 2008. Method for Building Sentence-Aligned Corpus from Wikipedia. En *Proceedings of the 33th AAAI workshop on Artificial Intelligence (AAAI-08)*.

Pivot strategies as an alternative for statistical machine translation tasks involving iberian languages*

Estrategias pivote como alternativa a las tareas de traducción automática estadística entre idiomas ibéricos

Carlos Henríquez†, Marta R. Costa-jussà⋆, Rafael E. Banchs‡, Lluis Formiga† and José B. Mariño†
† Universitat Politècnica de Catalunya-TALP
C/Jordi Girona, 08034, Barcelona
{carlos.henriquez,lluis.formiga,jose.marino}@upc.edu
⋆Barcelona Media Innovation Center
Av Diagonal, 177, 9th floor, 08018 Barcelona, Spain
marta.ruiz@barcelonamedia.org
‡ Institute for Infocomm Research
1 Fusionopolis Way 21-01, Singapore 138632
rembanchs@i2r.a-star.edu.sg

Resumen: Este artículo describe diferentes aproximaciones para construir sistemas de traducción automática estadísticas (SMT por sus siglas en inglés) entre idiomas de escasos recursos paralelos. La estrategia es especialmente interesante para España, un país con tres idiomas oficiales (catalán, vasco y gallego) aparte del castellano, en donde es difícil conseguir corpus paralelo entre cualquiera de los tres primeros pero es comparativamente fácil hacerlo entre castellano y cualquiera de ellos. Tal particularidad nos permite aprovechar el castellano como puente o pivote para construir sistemas que traduzcan entre catalán e inglés, por ejemplo. Estos sistemas son de gran utilidad para los idiomas minoritarios pues ayudan a darles una presencia global y a promover su uso. Como caso de uso, se describe un sistema catalán-inglés siguiendo la estrategia pivote de corpus sintético, la comparamos con una aproximación de cascada y comentamos sobre mejoras adicionales que pudieran implementarse para este par de idiomas en particular.
Palabras clave: idioma pivote, traducción automática estadística, corpus paralelo escaso, cascada, pseudo-corpus, modelos de traducción, frases, n-gramas

Abstract: This paper describes different pivot approaches to built SMT systems for language pairs with scarce parallel resources. The strategy is particularly interesting for Spain, a country with three official languages (Catalan, Basque, and Galician) besides Spanish, where it is difficult to find parallel corpora between two of the first three mentioned languages but it is relatively easy to collect it between Spanish and any of them. This characteristic, however, allow us to develop machine translation systems from major languages like English, to Catalan for instance, using Spanish as pivot. Such systems help these minority languages giving them global presence and promoting their use in content collaboration. We describe a English-Catalan baseline system built following the synthetic approach, we compare it with the transfer approach and comment about future enhancement that could be implemented for this language pair.
Keywords: pivot language, statistical machine translation, scarse parallel corpora, cascade, pseudo-corpus, phrase-based, ngram-based, translation models

1. Motivation

Spain is a multilingual country with four official languages: Catalan, Euskera, Galician and Spanish. Catalan is spoken by 11.5 million people, Euskera by 1.2 million people, Galician by 3.2 million people and Spanish by 400 million people. Given the high number of Spanish speakers compared to the other languages, Spanish has much more linguistic and data resources.

The quantity of resources is relevant in statistical machine translation. The more parallel text we have, the better the translation quality. In order to face the lack of resources in translation, there are many research works on pivot approaches which consist on using a pivot language to perform a source to target translation (Bertoldi et al., 2008a) (Costa-jussà, Henríquez, y Banchs, 2011). For example, in order to translate from Galician to Catalan, we could use Spanish as pivot language. There are much more resources in Galician-Spanish and Spanish-Catalan than between Galician and Catalan directly. The same could happen when interested in translating Catalan, Euskera or Galician into English. In this work, we introduce a state-of-the-art English-Catalan translation system recently built for the free web translator N-II[1].

The main differences with the Catalan-English SMT system presented in (de Gispert y Mariño, 2006) are that in this paper we use an extended corpus and we propose to build a hybrid system which uses an Ngram-based system for Catalan-Spanish and a phrase-based system for Spanish-English. The Ngram-based system outperforms the phrase-based system in Catalan-Spanish (Farrús et al., 2009) while the opposite occurs for the case of Spanish-English (Costa-Jussà y Fonollosa, 2009). Additionally, for the Catalan-Spanish system we are using a further competitive system using rules and statistical features (Farrús et al., 2011).

The remainder of this paper is organized as follows. Section 2 reports a brief description of the phrase-based and Ngram-based translation approaches. Section 3 presents the pivot approaches used in this paper. Section 4 describes the English-Catalan SMT system. Section 5 compares the pivot strategies in terms of translation quality and Section 6 presents the most relevant conclusions.

2. Statistical Machine Translation approaches

As mentioned in the previous section, we are working with two SMT systems: the phrase-based (Koehn, Och, y Marcu, 2003) and Ngram-based systems (Mariño et al., 2006; Casacuberta y Vidal, 2004), which are briefly described as follows.

2.1. Phrase-based

This approach to SMT performs the translation splitting the source sentence in segments and assigning to each segment a bilingual phrase from a phrase-table. Bilingual phrases are translation units that contain source words and target words, e.g. $<$ *unidad de traducción* | *translation unit* $>$, and have different scores associated to them. These bilingual phrases are then selected to maximize a linear combination of feature functions. Such strategy is known as the log-linear model (Och y Ney, 2002) and it is formally defined as:

$$\hat{e} = \arg\max_{e} \left[\sum_{m=1}^{M} \lambda_m h_m (e, f) \right] \qquad (1)$$

where h_m are different feature functions with weights λ_m. The two main feature functions are the translation model (TM) and the target language model (LM). Additional models include POS target language models, lexical weights, word penalty and reordering models among others.

Moses (Koehn et al., 2007) was used to build the phrase-based system.

2.2. Ngram-based

The base of the Ngram approach is the concept of tuple. Tuples are bilingual units with consecutive words both on the source and target side that are consistent with the word alignment. They must provide a unique monotonic segmentation of the sentence pair and they cannot be inside another tuple

* The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 247762 (FAUST) and from the Spanish Ministry of Science and Innovation through the Juan de la Cierva research program and the Buceador project (TEC2009-14094-C04-01).

[1] available at http://www.n-ii.org

in the same sentence. This unique segmentation allows us to see the translation model as a language model, where the language is composed of tuples instead of words. That way, the context used in the translation model is bilingual and implicitly works as a language model with bilingual context as well. In fact, while a language model is required in phrase-based and hierarchical phrase-based systems, in Ngram-based systems it is considered just an additional feature.

This alternative approach to a translation model defines the probability as:

$$P(f,e) = \prod_{n=1}^{N} P\left((f,e)_n \,|\, (f,e)_{n-1}, \ldots, (f,e)_1\right) \quad (2)$$

where $(f,e)_n$ is the n-th tuple of hypothesis e for the source sentence f.

As additional features, we used a Part-Of-Speech (POS) language model for the target side and a target word bonus model.

We used the open source decoder MARIE (Crego, de Gispert, y Mariño, 2005) to build the Ngram-based system.

3. Pivot Approaches

The best approaches to build a SMT system through a pivot language are: the cascade system, also known as the transfer approach and the pseudo-corpus or synthetic approach. Other pivot approaches do not outperform these two (Wu y Wang, 2007) (Cohn y Lapata, 2007). The cascade and the pseudo-corpus approaches have been evaluated and compared in works such as (de Gispert y Mariño, 2006; Bertoldi et al., 2008a; Bertoldi et al., 2008b). Consistently, both works have shown that the pseudo-corpus approach is the best performing strategy.

3.1. Cascade or transfer method

This approach considers the language pairs source-pivot and pivot-target independently. It consists in training and tuning two different SMT systems and combine them in a two-step process: first, we translate a source sentence using the source-pivot system; then, we use the resulting sentence as input for the pivot-target translation. A common variation for this strategy presented in (Khalilov et al., 2008) considers a n-best output instead of the single-best during the first translation and then produce a m-best translation in the last step. At the end, mn-best hypotheses are produced, which are reranked by using Minimum Bayes Risk (MBR) (Kumar y Byrne, 2004), allowing the introduction of additional features such as new language models.

3.2. Pseudo-corpus or synthetic approach

Instead of considering the two language pairs independently, this approach produces a single source-target SMT system. Assuming we have a source-pivot and a pivot-target parallel corpus, we build and tuned a pivot-target SMT system and we use it to translate the pivot part from the source-pivot corpus. This results in a source-target synthetic corpus (hence the name) which is finally used to build the source-target SMT system. For the tuning process, we could also use a synthetic development corpus but an actual source-target corpus is prefered, if possible. A simple variation for this approach is to build a pivot-source SMT system in order to translate the pivot part of the pivot-target corpus, and use the resulting source-target synthetic corpus to build the final system.

4. Building an English-Catalan SMT using Spanish as pivot

We present an English-Catalan SMT baseline system, using Spanish as the pivot language. In this case, the parallel corpus available for the Catalan-Spanish language pair was provided by the bilingual newspaper "El Periódico"[2] and the English-Spanish corresponds to the train corpora provided during the 2010 WMT's translation task[3], i.e. Europarl and News Commentary. We followed the synthetic approach described before to build the final system. Therefore, the Spanish part from the WMT Corpus was translated into Catalan and a English-Catalan phrase-based SMT system was built using the resulting synthetic corpus. Table 1 shows a summary of the statistics of both corpora. We also used the Catalan-Spanish baseline together with the Spanish-English baseline system presented in the 2010's WMT (Henríquez Q. et al., 2010) to build the other direction and compare the different approaches in it.

[2] http://www.elperiodico.es

[3] http://www.statmt.org/wmt10/translation-task.html

Corpora	Catalan	Spanish
Training sents.	4,6M	4,6M
Running words	96,94M	96,86M
Vocabulary	1,28M	1,23M
Development sents.	1966	1966
Running words	46765	44667
Vocabulary	9132	9426
Corpora	**Spanish**	**English**
Training sents.	1,18M	1,18M
Running words	26,45M	25,29M
Vocabulary	118073	89248
Development sents.	1729	1729
Running words	37092	34774
Vocabulary	7025	6199
Test sents.	2525	2525
Running words	69565	65595
Vocabulary	10539	8907

Cuadro 1: Catalan-Spanish and Spanish-English corpora (M stands for Millions)

4.1. Spanish-Catalan baseline system

As mentioned before, the Spanish-Catalan SMT system (named N-II) is based on the corpus provided by the bilingual newspaper "El Periódico". It is a Ngram-based SMT system that includes several improvements specific to the language pair: a homonym disambiguation for the Catalan verb 'soler' and Catalan possessives, special consideration for pronominal clitics, upper-case words and the Catalan apostrophe, gender concordance, numbers and time categorization and text processing for common mistakes found when writing in Catalan. The full description can be found in (Farrús et al., 2011).

4.2. English-Catalan system description

Once obtained the Catalan translation from the Spanish section of the WMT corpus, a phrase-based SMT system was built using Moses as the decoder. Apart from the baseline pipeline, the system also includes a POS target language model computed with TnT (Brants, 2000), numbers and time categorization similar to N-II and the parallel corpus was aligned considering the Catalan lemmas computed with Freeling (Padró et al., 2010) and the English stems of words obtained with Snowball[4].

[4]http://snowball.tartarus.org

Pivot approach	Direction	BLEU
Cascade	cat-eng	21,63
Cascade	eng-cat	24,29
Pseudo-corpus	cat-eng	23,19
Pseudo-corpus	eng-cat	26,97

Cuadro 2: English-Catalan results

5. Results

Table 2 shows the BLEU score of the cascade and pseudo-corpus approaches in both directions. The test set was the one provided as internal test set during the WMT translation task. It is also important to mention that the score was computed using one reference.

The final quality of the Catalan-English system is determined by the quality of the Spanish-English corpus, whose baseline has a BLEU around 24 (Henríquez Q. et al., 2010). The Catalan-Spanish baseline has a BLEU around 80 (Farrús et al., 2009). Also there is a negative effect given the difference in domain between the Catalan-Spanish corpus (a regional newspaper) and Spanish-English corpus (Europarl).

Using paired bootstrap resampling (Koehn, 2004), we can see that for these systems, the Pseudo-corpus approach is better than Cascade with 95 % statistical significance.

6. Conclusions and further work

We have presented an English-Catalan SMT system built using Spanish as pivot language, given the scarce resources for English-Catalan.

Similarly to previous research work, we have seen here that, in the particular translation task under consideration, the pseudo-corpus approach constitutes the best strategy for pivot translation. Although the cascade approach clearly performs worse than the pseudo-corpus approach, it could be also beneficial to consider a system combination between these two strategies to further boost the quality of the translations.

Further work should focus on building Spanish-pivot systems between all the official languages and English, as well as among them. The similarities between the languages (except Basque) and the availability of parallel corpora between Spanish and the others encourage the approach.

Bibliografía

Bertoldi, N., R. Cattoni, M. Federico, y M. Barbaiani. 2008a. FBK @ IWSLT-2008. En *Proc. of the International Workshop on Spoken Language Translation*, páginas 34–38, Hawaii, USA.

Bertoldi, Nicola, Madalina Barbaiani, Marcello Federico, y Roldano Cattoni. 2008b. Phrase-Based Statistical Machine Translation with Pivot Languages. En *Proceedings of IWSLT*.

Brants, T. 2000. TnT – a statistical part-of-speech tagger. En *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Casacuberta, F. y E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225.

Cohn, T. y M. Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. En *Proc. of the ACL*.

Costa-Jussà, M. R. y J. A. R. Fonollosa. 2009. Phrase and ngram-based statistical machine translation system combination. *Applied Artificial Intelligence: An International Journal*, 23(7):694–711, August.

Costa-jussà, M.R., C. Henríquez, y R. Banchs. 2011. Evaluación de estrategias para la traducción automática estadística de chino a castellano con el inglés como lengua pivote. En *Proc. of the SEPLN*, Huelva.

Crego, J.M., A. de Gispert, y J.B. Mariño. 2005. An Ngram-based Statistical Machine Translation Decoder. En *Proceedings of 9th European Conference on Speech Communication and Technology (Interspeech)*.

de Gispert, A. y J.B. Mariño. 2006. Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish. En *Proc. of LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages (SALTMIL'06)*, páginas 65–68, Genova.

Farrús, M., M. R. Costa-jussà, J. B. Mariño, M. Poch, A. Hernández, C. Henríquez, y J. A. R. Fonollosa. 2011. Overcoming statistical machine translation limitations: error analysis and proposed solutions for the catalan-spanish language pair. *Language Resoures and Evaluation*, 45(2):181–208.

Farrús, M., M. R. Costa-jussà, M. Poch, A. Hernández, y J. B. Mariño. 2009. Improving a catalan-spanish statistical translation system using morphosyntactic knowledge. En *Proceedings of European Association for Machine Translation 2009*.

Henríquez Q., C. A., M.R. Costa-jussà, V. Daudaravicius, R. E. Banchs, y J. B. Mariño. 2010. Using collocation segmentation to augment the phrase table. En *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, páginas 104–108, Uppsala, Sweden, July.

Khalilov, M., M. R. Costa-Jussà, C. A. Henríquez, J. A. R. Fonollosa, A. Hernández, J. B. Mariño, R. E. Banchs, B. Chen, M. Zhang, A. Aw, y H. Li. 2008. The TALP & I2R SMT Systems for IWSLT 2008. En *Proc. of the International Workshop on Spoken Language Translation*, páginas 116–123, Hawaii, USA.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. En *Proceedings of EMNLP*, volumen 4, páginas 388–395.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, y E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. En *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, páginas 177–180, Morristown, NJ, USA.

Koehn, P., F.J. Och, y D. Marcu. 2003. Statistical phrase-based translation. En *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*.

Kumar, S. y W. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. En *Proceedings of the Human Language Technology and North American*

Association for Computational Linguistics Conference (HLT/NAACL'04), páginas 169–176, Boston, USA, May.

Mariño, José B., Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, y Marta R. Costa-jussà. 2006. Ngram-based Machine Translation. *Computational Linguistics*, 32(4):527–549.

Och, F. J. y H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. En *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Padró, Ll., M. Collado, S. Reese, M. Lloberes, y I. Castellón. 2010. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. En *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, La Valleta, Malta, May.

Wu, H. y H. Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. En *Proc. of the ACL*, páginas 856–863, Prague.

A Bilingual Summary Corpus for Information Extraction and other Natural Language Processing Applications*

Un corpus bilingüe para la extracción de información y otras tareas de procesamiento de lenguaje natural.

Horacio Saggion and Sandra Szasz
Universitat Pompeu Fabra
Departament de Tecnologies de la Informació i les Comunicacions
Grupo TALN
C/Tanger 122 - Barcelona - 08018
Spain
horacio.saggion@upf.edu, sandra.szasz@upf.edu

Resumen: Presentamos un corpus bilingüe comparable en español e inglés de pares de resúmenes de tres tipos de eventos: accidentes aéreos, accidentes ferroviarios y terremotos. Cada resumen es un texto que describe de manera sucinta un evento particular. El corpus fue anotado manualmente con información semántica sobre cada evento y resulta apropiado para la experimentación en extracción de información monolingüe así como también cros-lingue.
Palabras clave: Extracción de informaciones, corpus bilingüe, resúmenes

Abstract: Cross-lingual information extraction, the task of extracting information from multiple-multilingual sources, can benefit from the availability of a corpus of equivalent documents in various languages. We present a dataset of pairs of summaries in Spanish and English in various application domains and demonstrate its use in information extraction experiments. The dataset has been manually annotated with semantic information.
Keywords: Cross-lingual information extraction, biligual corpus, summaries

1 Introduction

Cross-lingual information extraction, the task of extracting information from multiple-multilingual sources, is a problem which has received considerably less attention than extraction from mono-lingual sources. In this paper, we are concerned with the creation of a dataset for the development and evaluation of *cross-lingual information extraction* systems. Our corpus is a set of pairs of summaries in Spanish and English in various domains. An example of the dataset is shown below:

> *17 julio 2006 Isla de Java: un maremoto de magnitud 7,7 Richter de magnitud provoca un 'tsunami' que causó la muerte de 596 personas.*

> *On 17 July at 03:19:25 p.m. local time an earthquake measuring 7.7 on the Richter scale struck offshore immediately south of West Java at a depth of 10 km. The areas affected by the earthquake and resultant tsunami included the districts of Taskimalaya, Ciamis, Sukabumi and Garut in West Java province, Cilacap, Kebumen and Banyumas in Central Java and the Gunung Kidul and Bantul districts in the province of Yogyakarta. No. Deaths 500.*

These elements in the dataset are non-translated equivalent summaries which have been found on the Web. They report on the same event, in this case an earthquake, but because they are not translations of one another, they contain different information, for example the Spanish summary reports 596 people dead while the English summary

* We are grateful to Programa Ramón y Cajal from Ministerio de Ciencia e Innovación, Spain.

reports 500 people dead. The English summary is more verbose and contains information about the time of the event and various locations affected by the tremor thus being the two elements complementary. The dataset can be used for training information extraction systems, studying template-to-text bilingual generation, and automatic knowledge modelling.

This paper gives an overview of the dataset and initial experiments showing its potential application. The rest of this paper is structured as follows: Section 2 we explain related work and then, in Section 3 we describe the data set created. After that, in Section 4 we illustrate how we have used the corpus and in Section 5 we present our conclusions.

2 Related Work

There are various multilingual datasets in the machine translation field such as the Europarl Multilingual Corpus (Koehn, 2005) or the United Nations Parallel Corpus (Eisele y Chen, 2010). Related to the work presented here are those datasets prepared for text summarization or information extraction research. Among them we have identified the SummBank corpus (Saggion et al., 2002) created for the study of multi-lingual summarization in Chinese and English. The documents in this corpus are translations of one another and contain announcements of a local administration. The corpus has been used in text summarization and information retrieval experiments (Radev et al., 2003). Because of the content and annotation provided with the dataset, this corpus is probably less suitable for information extraction. The CAST corpus (Orăsan, Mitkov, y Hasler, 2003) contains newswire texts and popular science articles in English where annotations are added to indicate: (i) essential sentences, (ii) unessential fragments in sentences, and (iii) links between sentences when one sentence is needed to understand another. Because of the particular annotation schema used, the corpus has potential applications for sentence compression. The SumTime-Meteo Corpus (Reiter y Sripada, 2002) provides weather summaries in English from numerical data and is potentially useful in data to text generation applications and information extraction. The Ziff-Davis corpus contains technical documents in English and their human created summaries and has been used in text summarization experiments (Knight y Marcu, 2000). The dataset of the Message Understanding Conferences (ARPA, 1993) is probably the best known set for the development of information extraction systems.

3 Data Set Creation and Annotation

The dataset under development is a comparable corpus of Spanish and English summaries for four different domains: aviation accidents, rail accidents, earthquakes, and terrorist acts; this later subset is still under development. Further domains will be incorporated in the future for researchers interested in evaluating the robustness and adaptation capabilities of different natural language processing techniques. In order to collect the summaries, a keyword search strategy was used to search for documents on the Internet using Google Search. Keywords per domain were defined and used to select a set of Web pages in Spanish, for example the keywords "lista de terremotos" could be used to search for documents in the earthquake domain. The pages returned by the search engine were examined to verify if they actually contained an event summary and in that case a document was created for the summary (it is not unusual to find multiple summaries in a single Web page). The documents were given names indicating the type of event and the date of the event/incident. A set of around 50 summaries per domain in Spanish were collected in this manner. After this, for each event summary originally in Spanish the Internet was searched for an equivalent English summary (not a translation) using keywords in English, this time manually derived from the Spanish summary. For example if an earthquake event mentioned a particular date and intensity, then those elements were used as keywords. Following this procedure we found equivalent English summaries for most of the Spanish ones.

For each domain (event or incident) a set of semantic components (i.e., slots) were identified based on intuition and on the actual data observed in a set of summaries for the domain. The slots/components making

Information	# Spa	# Eng
City	23	16
Country	47	31
DateOfEarthquake	53	36
Depth	1	4
Duration	1	3
Epicentre	7	7
Fatalities	50	35
Homeless	7	11
Injured	9	11
Magnitude	47	32
OtherPlacesAffected	27	29
Province	10	9
Region	25	25
Survivors	1	2
TimeOfEarthquake	4	21
TotalVictims	2	0

Table 3: Number of Semantic Concepts in Spanish and English Earthquake's Summaries

up the templates which model the domain are shown in Table 1.

Corpus examples (pairs of summaries in the two languages) for the three domains are shown in Table 2. In order to manually annotate the summaries with semantic information, we have used the GATE annotation framework (Maynard et al., 2002). To facilitate the annotation process an annotation schema was used so that in the GATE Graphical User Interface the target text span to be annotated can be selected, and annotated with one valid category from the annotation schema. The summaries are annotated by one person, however a second person checks the annotations for any inconsistency. Note that because we are dealing with short texts, the annotation process is easier than that of annotating a full event report.

The number of event components found in the set of summaries is reported in Tables 3, 4 and 5.

4 Uses of the Corpus

We have started using the corpus in monolingual as well as in cross-lingual information extraction. Information extraction is the mapping of natural language texts (e.g. news articles, web pages, e-mails) into predefined structured representations or templates (Grishman, 1997) such as those we defined in Table 1. Various techniques have been used

Information	# Spa	# Eng
Airline	26	31
Cause	16	13
DateOfAccident	30	29
Destination	8	7
FlightNumber	26	31
NumberOfVictims	21	23
Origin	11	5
Passenger	5	9
Place	24	28
Survivors	5	10
Tripulation	8	6
TypeOfAccident	28	29
TypeOfAircraft	18	32
Year	31	31

Table 4: Number of Semantic Concepts in Spanish and English Aviation Accident's Summaries

Information	# Spa	# Eng
Cause	18	23
DateOfAccident	43	36
Destination	8	12
NumberOfVictims	43	37
Origin	9	13
Place	45	40
Survivors	25	20
TypeOfAccident	41	36
TypeOfTrain	30	33

Table 5: Number of Semantic Concepts in Spanish and English Train Accident's Summaries

in the development of information extraction systems including rule-based approaches relying on robust partial syntactic analysis (Appelt et al., 1993), Hidden Markov Models (Leek, 1997; Freigtag y McCallum, 1999), and a combination of supervised machine learning (Ciravegna, 2001) and weakly supervised machine learning (Yangarber, 2003; Riloff, 1996). In recent years there has been an increasing interest in the application of information extraction for the "Semantic Web" using ontologies as knowledge representation formalisms (Maynard et al., 2007; Saggion et al., 2007) as well as on multilingual and cross-lingual information extraction (Poibeau y Saggion, 2007; Poibeau, Saggion, y Yangarber, 2008). It has been shown that extraction from multiple

Incident	Semantic Schema
Aviation Accident	Airline; Cause; DateOfAccident; Destination; FlightNumber; Origin; Passenger; Place; Survivors; Tripulation; TypeOfAccident; TypeOfAircraft; Victims; Year
Railway Accident	Cause; DateOfAccident; Destination; Origin; Passenger; Survivors; TrainLine; Tripulation; TypeOfAccident; TypeOfTrain; Victims; Year
Earthquake	City; Country; DateOfEarthquake; Depth; Epicentre; Fatalities; Homeless; Injured; Magnitude; OtherPlacesAffected; Province; Region; Survivors; TimeOfEarthquake; TotalVictims

Table 1: Conceptual Information in Summaries

Aviation Accident
2009 30 de junio: el vuelo 626 de Yemenia chocó en cercanías a Comoras, en el Océano Indico.
2009 June 30 Yemenia Flight 626, an Airbus A310-300 flying from Sana'a, Yemen to Moroni, Comoros, crashes into the Indian Ocean with 153 people aboard; one 12-year-old is found clinging to the wreckage.
Railway Accident
12 enero 1997 8 muertos y 25 heridos en el descarrilamiento del tren rápido Milán-Roma en las proximidades de Piacenza (Italia).
January 12, 1997 A Pendolino train derails just before a train station at Piacenza, Italy, killing 8 people and injuring 29 others.
Earthquake
27 mayo 2006 Isla de Java (Indonesia): un terremoto de magnitud 6,2 Richter causa al menos 6.234 muertos, 20.000 heridos y 340.000 desplazados.
May 27, 2006 A powerful earthquake struck Indonesia's central province of Java early Saturday morning at 0554 Hrs local time (26 May 2254 Hrs GMT), flattening buildings and killing over 4900 people.

Table 2: Sample of the Parallel Corpus

multilingual sources can lead to improved semantic indexing (Saggion et al., 2003) when compared to monolingual or single source extraction. It has also been shown that cross-lingual extraction (Hakkani-Tür, Ji, y Grishman, 2007) can be used as a filtering step to improve retrieval in a target language.

4.1 Experiments

Our cross-lingual information extraction experiments involve the use of a system trained in a source language to extract information from translations from another language. However, to test how useful the dataset is, we have started with monolingual experiments per domain and language (e.g., six systems in total). The systems are a pipeline of text processing tools followed by a process of token classification based on Support Vector Machines (Li et al., 2002). The machine learning component was adjusted through testing and evaluation cycles. The text analysis components are as follows:

- For English: we used default processors from the GATE system: tokenizer, parts-of-speech tagger, rule-based morphological analysis, dictionary lookup, and named entity recognition and classification;

Event	Prec	Rec	F
Train Accident Spanish	0.49	0.41	0.44
Train Accident English	0.76	0.56	0.64
Aviation Accident Spanish	0.64	0.47	0.53
Aviation Accident English	0.68	0.62	0.65
Earthquake Spanish	0.62	0.48	0.54
Earthquake English	0.49	0.36	0.41

Table 6: Overall Extraction Performance in Spanish and English

- For Spanish: we used the TreeTagger software (Schmid, 1995) and our own trainable named entity recognizer.

Basic linguistic features were used to train Spanish and English extraction systems. Both the Spanish and English systems use for each token to be classified a context window of five positions containing the following token features: orthography (e.g., word capitalization), word root, parts-of-speech, named entity type, and dictionary (gazetteer lookup) information.

Because each dataset is relatively small, we have performed 10-fold cross-validation experiments reporting here aggregated precision, recall, and f-score figures. Table 6 presents the results. The English extraction system performs better than the Spanish system in the train and aviation accident domains, while the Spanish system performs better than the English one in the earthquake domain. This could be due to the fewer human annotations in the English earthquakes compared to the Spanish counterpart. It is worth noting that the English summaries are more verbose in this domain making extraction more difficult. Although the obtained results are modest, they have to be assessed taken into account the limited syntactic and semantic information available from the text processors. In order to test how the systems cope with noisy data we have translated the Spanish summaries into English and the English summaries into Spanish using Google Translator and have applied the information extraction systems to each translation. In these experiments, for each translation T in a domain D, the extraction system is trained with all documents except the document which is equivalent to T and the resulting system is applied to summary T. Evaluation metrics are also computed and aggregated over all documents. In these experiments we have obtained in most domains and languages f-scores over 0.60 which although not directly comparable with the mono-lingual results are certainly encouraging, full details on these experiments can be found in (Saggion y Szasz, 2011).

5 Conclusions

In this paper we have presented an overview of a dataset with potential interest for cross-lingual natural language processing applications. To the best of our knowledge this is one of the few datasets in this field for the pair Spanish/English. We have shown information extraction and cross-lingual extraction as potential applications of the dataset. Our current work involves the expansion of the dataset to cover additional domains such as terrorism and sports. In future work we will address automatic domain modelling from summaries and information extraction induction. We also plan to use the cross-lingual extraction results to improve mono-lingual mono-document extraction.

References

Advanced Research Projects Agency. 1993. *Proceedings of the Fifth Message Understanding Conference (MUC-5)*. Morgan Kaufmann, California.

Appelt, D.E., J.R. Hobbs, J. Bear, D. Israel, M. Kameyama, y M. Tyson. 1993. Description of the JV-FASTUS system as used for MUC-5. En *Proceedings of the Fourth Message Understanding Conference MUC-5*, páginas 221–235. Morgan Kaufmann, California.

Ciravegna, F. 2001. Adaptive information extraction from text by rule induction and generalisation. En *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*.

Eisele, Andreas y Yu Chen. 2010. MultiUN: A Multilingual Corpus from United

Nation Documents. En Nicoletta Calzolari (Conference Chair) Khalid Choukri Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Mike Rosner, y Daniel Tapias, editores, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Freigtag, D. y A. K. McCallum. 1999. Information Extraction with HMMs and Shrinkage. En *Proceesings of Workshop on Machine Learnig for Information Extraction*, páginas 31–36.

Grishman, R. 1997. Information extraction: Techniques and challenges. En Maria Teresa Pazienza, editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School (SCIE-97)*, volumen 1299 de *Lecture Notes in Computer Science*, páginas 10–27, Frascati, Italy, Jul. Springer Verlag.

Hakkani-Tür, D., Heng Ji, y R. Grishman. 2007. Using Information Extraction to Improve Cross-lingual Document Retrieval. En *Proceedings of the 1st Intl. Workshop on Multi-source Multi-lingual Information Extraction and Summarization Workshop*.

Knight, K. y M. Marcu. 2000. Statistics-based summarization - step one: Sentence compression. En *AAAI/IAAI*, páginas 703–710, Austin, Texas.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *Conference Proceedings: the tenth Machine Translation Summit*, páginas 79–86, Phuket, Thailand. AAMT, AAMT.

Leek, T.R. 1997. Information Extraction Using Hidden markov Models. Informe técnico, University of California, San Diego, USA.

Li, Y., H. Zaragoza, R. Herbrich, J. Shawe-Taylor, y J. Kandola. 2002. The Perceptron Algorithm with Uneven Margins. En *Proceedings of the 9th International Conference on Machine Learning (ICML-2002)*, páginas 379–386.

Maynard, D., H. Saggion, M. Yankova, K. Bontcheva, y W. Peters. 2007. Natural Language Technology for Information Integration in Business Intelligence. En W. Abramowicz, editor, *10th International Conference on Business Information Systems*, Poland, 25-27 April.

Maynard, D., V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, y Y. Wilks. 2002. Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 8(2/3):257–274.

Orăsan, C., R. Mitkov, y L. Hasler. 2003. CAST: a Computer-Aided Summarisation Tool. En *Proceedings of EACL2003*, páginas 135 – 138, Budapest, Hungary, April.

Poibeau, T. y H. Saggion, editores. 2007. *1st International Workhop on Multi-Source, Multi-Lingual Information Extraction and Summarization*. RANLP, September.

Poibeau, T., H. Saggion, y R. Yangarber, editores. 2008. *2nd International Workhop on Multi-Source, Multi-Lingual Information Extraction and Summarization*. COLING, September.

Radev, Dragomir Radev, Wai Lam, Arda C Elebi, Simone Teufel, John Blitzer, Danyu Liu, Horacio Saggion, Hong Qi, Elliott Drabek, y Johns Hopkins U. 2003. Evaluation challenges in large-scale document summarization. En *In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, páginas 375–382.

Reiter, E. y S. Sripada. 2002. Squibs and discussions: human variation and lexical choice. *Computational Linguistics*.

Riloff, E. 1996. Automatically generating extraction patterns from untagged text. *Proceedings of the Thirteenth Annual Conference on Artificial Intelligence*, páginas 1044–1049.

Saggion, H., H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, y Y. Wilks. 2003. Multimedia Indexing through Multisource and Multilingual Information Extraction; the MUMIS project. *Data and Knowledge Engineering*, 48:247–264.

Saggion, H., A. Funk, D. Maynard, y K. Bontcheva. 2007. Ontology-based information extraction for business applications. En *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*, Busan, Korea, November.

Saggion, H., D. Radev, S. Teufel, L. Wai, y S. Strassel. 2002. Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment. En *3rd International Conference on Language Resources and Evaluation (LREC 2002)*, páginas 747–754, Las Palmas, Gran Canaria, Spain.

Saggion, H. y S. Szasz. 2011. Multi-domain cross-lingual information extraction from clean and noisy texts. En *Proceedings of the Brazilian Symposium on Information and Human Language Technology*, Cuiabá, Brazil, 24-26 October. SBC.

Schmid, H. 1995. Improvements in part-of-speech tagging with an application to german. En *In Proceedings of the ACL SIGDAT-Workshop*, páginas 47–50.

Yangarber, R. 2003. Counter-Training in Discovery of Semantic Patterns. En *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*.

Extracción automática de léxico bilingüe: experimentos en español y catalán*

Automatic Bilingual Lexicon Extraction: Experiments in Spanish and Catalan

Raphaël Rubino **Iria da Cunha** **Georges Linarès**

Laboratoire Informatique d'Avignon
339, chemin des Meinajaries
84911 Avignon Cedex 9, Francia
raphael.rubino@univ-avignon.fr
georges.linares@univ-avignon.fr

Institut Universitari de
Lingüística Aplicada
Roc Boronat 138
08018 Barcelona, España
iria.dacunha@upf.edu

Resumen: En este artículo presentamos un sistema de extracción automática de léxico bilingüe catalán-español. Evitamos el empleo de corpus paralelos y usamos la información ofrecida por la Wikipedia como un corpus comparable entre el español y el catalán. Empleamos la similitud contextual para traducir unidades léxicas que no pueden traducirse por la distancia de edición. Los resultados obtenidos son positivos y confirman que este método podría aplicarse a las lenguas ibéricas.
Palabras clave: extracción automática, léxico bilingüe, traducción automática, español, catalán

Abstract: In this paper, we propose an automatic bilingual lexicon extraction system for Catalan and Spanish languages. Parallel corpora are not employed and Wikipedia is used as Catalan-Spanish comparable corpora. A contextual similarity approach is used to translate lexical units that are not translated by an edition distance. The obtained results are positive and confirm that this method could be applied to Iberian languages.
Keywords: Automatic Extraction, Bilingual Lexicon, Machine Translation, Spanish, Catalan

1. Introduction

En la Península Ibérica coexisten cinco lenguas oficiales: español, catalán, gallego, euskera y portugués. Para establecer vínculos entre estas lenguas y favorecer el multilingüismo, es necesario desarrollar recursos para todas ellas. Además, es indispensable crear recursos que permitan relacionarlas. Actualmente, hay una carencia de recursos de Procesamiento del Lenguaje Natural (NLP) para algunas de ellas, especialmente el gallego, el catalán y el euskera. Uno de los recursos necesarios para interrelacionar estas lenguas y diseñar herramientas de PLN (como sistemas de traducción automática) son los léxicos multilingües. Sin embargo, su desarrollo y actualización es costoso y lento, ya que normalmente supone la intervención humana.

El diseño de herramientas automáticas que ayuden en la construcción de léxicos bilingües (o multilingües) supone un reto en el ámbito del PLN. Existen trabajos que tratan este tema empleando diferentes estrategias. La mayor parte utilizan corpus paralelos (Brown et al., 1990; Wu y Xia, 1994; Koehn, 2005). No obstante, la creación de este tipo de corpus es costosa, lo cual encarece la investigación y no permite trabajar sobre todas las combinaciones de lenguas. Otra línea de investigación se basa en la utilización de un recurso más accesible, los corpus bilingües comparables, es decir, conjuntos de textos no paralelos con temáticas comunes pero escritos en cada lengua de manera independiente. Diversos autores han estudiado la

* Esta investigación ha sido parcialmente financiada por la Agence Nationale de la Recherche (ANR, Francia), proyecto AVISON (ANR-007-014); y los proyectos RICOTERM (FFI2010-21365-C03-01) y APLE (FFI2009-12188-C05-01) en España.

posibilidad de extraer unidades léxicas a partir de estos corpus, basándose en la hipótesis de que una unidad léxica y su traducción comparten similitudes en cuanto a su contexto (Fung, 1995; Rapp, 1995). Además de corpus comparables, esta aproximación emplea un léxico bilingüe preliminar de las lenguas analizadas.

La mayoría de las investigaciones sobre este tema se han realizado para relacionar el inglés con otras lenguas. Para las lenguas ibéricas, encontramos algunos trabajos, que utilizan principalmente métodos basados en corpus paralelos: para inglés-gallego (Guinovart y Fontenla, 2004), para portugués, español e inglés (Caseli y Nunes, 2007), y para inglés-gallego e inglés-portugués (Guinovart y Simoes, 2009).

Como se afirma en (Gamallo Otero y Pichel Campos, 2007), "desgraciadamente, no hay todavía una gran cantidad de texto paralelo, especialmente en lo que se refiere a lenguas minorizadas". Por esto, trabajar con lenguas como el gallego, catalán o euskera se hace más complicado. En (Gamallo Otero y Pichel Campos, 2007) se propone un método basado en corpus comparables de la Web, usando la idea de la similitud contextual. Lo aplican al español y el gallego, y, aunque sus resultados no superan los obtenidos usando corpus paralelos, son elevados. Esto refuerza la idea de que la gran cantidad de datos incluidos en la Web es una fuente de información importante y explotable para la construcción automática de léxicos bilingües. En esta línea, en (Gamallo y González, 2010) se propone un método automático para construir corpus comparables empleando la Wikipedia. En (Tomás et al., 2008) se construye un corpus que incluye dos tipos de artículos de la Wikipedia (paralelos y comparables) en español y catalán. En (Vivaldi y Rodríguez, 2010) se presenta un método de extracción de terminología bilingüe que emplea las categorías y estructura de la Wikipedia. La extracción de frases paralelas de la Wikipedia es también una tarea interesante que ha sido explorada por (Smith, Quirk, y Toutanova, 2010), por ejemplo, realizando diferentes experimentos a partir de la estructura de la Wikipedia.

El objetivo de este trabajo es desarrollar un sistema de extracción automática de léxico bilingüe para las lenguas de la Península Ibérica. Concretamente, trabajamos el par de lenguas español-catalán. Para ello, evitamos el empleo de corpus paralelos y aplicamos la idea de la similitud contextual entre una unidad léxica y su traducción (Fung, 1995; Rapp, 1995), empleando textos de la Wikipedia como corpus comparable. La metodología descrita en este trabajo está basada en el empleo de recursos y heurísticas existentes, pero aplicadas concretamente a la extracción de léxico bilingüe en estas dos lenguas.

2. Metodología

La metodología de nuestro trabajo incluye dos fases principales: Preprocesamiento y creación de recursos léxicos (FASE 0) y Aplicación del algoritmo (FASE 1).

2.1. FASE 0: Preprocesamiento y creación de recursos léxicos

Ya que nuestro trabajo se basa en un corpus comparable y un léxico bilingüe, en esta fase se construyen estos recursos. Concretamente, necesitamos dos léxicos bilingües: I) un léxico con candidatos a la traducción (con sus correspondientes traducciones) y II) un léxico "pivote" utilizado como elemento de relación entre las dos lenguas.

2.1.1. Preprocesamiento del corpus comparable

El preprocesamiento del corpus comparable incluye:

- Descarga de un fichero con todos los artículos de la Wikipedia (Wikipedia Dump) en las dos lenguas de trabajo (español y catalán).
- Eliminación de "páginas redirigidas" en Wikipedia, es decir, artículos que tienen un título pero no contienen texto en su interior. Por ejemplo, en la Wikipedia en español, la unidad "Proyección Azimutal" está vacía y redirigida a "Proyección azimutal" (simplemente cambia una "a" en mayúscula o minúscula); el año "4450" está redirigido al artículo sobre el "V milenio", etc.
- Eliminación de las stopwords en las dos lenguas. La lista de stopwords en catalán se ha obtenido del área de Ingeniería Lingüística del Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF)[1]. La lista

[1] http://latel.upf.edu/morgana/altres/pub/ca_stop.htm

de stopwords en español se ha obtenido del Laboratoire Informatique d'Avignon (LIA-UAPV)[2].

- Formateo de este fichero en Trec-text[3]. En el siguiente ejemplo se muestra un ejemplo de este tipo de formato, en donde la etiqueta <DOCNO> indica el número de documento, <TITLE> el título y <TEXT> el contenido:

```
<DOC>
 <DOCNO> 22 </DOCNO>
 <TITLE> Astronomía galáctica </TITLE>
 <TEXT>
   se denomina 'astronomía galáctica' a
   la investigación astronómica de nuestra
   galaxia, la vía láctea [...] seguros
   posee un agujero negro, etc.
 </TEXT>
</DOC>
```

- Indexación de los artículos con Lemur Indexation Toolkit[4]. Usamos esta herramienta para facilitar el cálculo de coocurrencias entre la unidad léxica que se quiere traducir y su contexto (es decir, las palabras del léxico II).

Actualmente, la Wikipedia en español contiene 761.727 artículos y en catalán 341.142. Depués de este preprocesamiento, nuestro corpus incluye 701.423 artículos en español y 296.465 en catalán. Esta reducción se debe a la eliminación de artículos redirigidos. No se realizó una selección temática de los artículos incluidos en el corpus, sino que se emplearon todos los temas de la Wikipedia. Tampoco se usó la estructura de la Wikipedia.

2.1.2. Recopilación del léxico I

En esta fase, creamos nuestro propio léxico bilingüe, que contiene los candidatos a la traducción en la lengua de partida (catalán), acompañados de su traducción en la lengua de llegada (español). Construimos estos recursos dada la carencia de léxicos bilingües extensos y actualizados gratuitos disponibles para el par de lenguas empleadas. Así, nuestro léxico podrá contener neologismos de reciente creación (como, por ejemplo, "mileurista")[5]. Esta fase incluye dos subfases:

1. Extracción de relaciones de correspondencia entre los títulos de los artículos de la Wikipedia en español y catalán, para obtener una lista preliminar de léxico bilingüe. Las relaciones entre los artículos en estas dos lenguas se establecen mediante enlaces interlengua (en el menú "En otros idiomas" de la Wikipedia en español). Establecemos las correspondencias en los dos sentidos (español-catalán y catalán-español) porque, en ocasiones, la estructura de la Wikipedia no correlaciona de la misma forma las entradas en los dos sentidos. Por ejemplo, en la Wikipedia en catalán encontramos la entrada "Prestige", que está correlacionada en la Wikipedia en español con "Desastre del Prestige". Sin embargo, la Wikipedia en español también ofrece la entrada "Prestige" (que se refiere al mismo petrolero), que solo muestra su correspondencia al inglés y al ruso, pero no al catalán. Vemos así que la estructura de la Wikipedia en español es más compleja que la de otras lenguas con menos entradas.
2. Filtrado de la lista preliminar de los dos léxicos bilingües mediante la eliminación automática de:

   - Pares de unidades léxicas que no mantienen la misma correlación en la estructura de la Wikipedia en los dos sentidos.
   - Pares de unidades léxicas que coinciden en las dos lenguas. Este criterio se aplica por dos motivos. Primero, porque consideramos que no es interesante evaluar los pares de unidades que son idénticas. Segundo, porque una gran cantidad de las unidades de este léxico bilingüe extraído de la Wikipedia serán entidades nombradas iguales en ambas lenguas, como por ejemplo "Harry Potter".
   - Pares de elementos numéricos, ya que no nos interesa traducir cifras, años, fechas, etc., aunque somos conscientes de que estás entidades podrían servir para poder paralelizar de forma eficiente frases en corpus comparables.
   - Pares de elementos en que solo uno tiene un signo de puntuación: generalmente

[2] http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/torres/logiciels/fonctionnels_esp.txt

[3] http://trec.nist.gov

[4] http://www.lemurproject.org

[5] Para más información sobre neología véase (Cabré y Estopà, 2009)

indican un error en la traducción (excepto el punto de la geminada del catalán). - Pares de elementos que pueden traducirse por la distancia de edición (Levenshtein, 1966). Por ejemplo, las siguientes unidades léxicas del catalán (a la izquierda) fueron traducidas correctamente al español por la distancia de edición (a la derecha), ya que las similitudes ortográficas son evidentes:

CATALÁN	ESPAÑOL
palau de westminster	*palacio de westminster*
lateralitat	*lateralidad*
fagocitosi	*fagocitosis*
província de bilecik	*provincia de bilecik*

En cambio, las siguientes unidades del catalán no se tradujeron adecuadamente:

CATALÁN	ESPAÑOL
surquillo	*bordillo*
floquet neu	*alquino*
tupaia	*tucana*
eratostenià	*río eno*

Comenzamos con un léxico de 140.137 unidades. Después del filtrado, antes de aplicar la distancia de edición, obtenemos 57.859 unidades y, después de la distancia de edición, 8.045 unidades, con las que trabajamos finalmente. Este léxico final contiene las unidades léxicas más difíciles de traducir, porque no pueden ser traducidas por una distancia de edición tradicional. Por este motivo, consideramos que la traducción automática de estas 8.045 unidades es el principal reto. Partimos de la idea de que el léxico bilingüe creado en esta fase es correcto. Sin embargo, no hemos realizado una revisión manual, dada su gran extensión. Esta revisión sería óptima para eliminar errores, pero intentamos evitar al máximo la intervención humana.

2.1.3. Recopilación del léxico II

Como ya hemos comentado, este léxico "pivote" se utiliza como elemento de relación entre las dos lenguas del trabajo. Por este motivo, este léxico debe ser correcto necesariamente, ya que gracias a él se realizan las corrrespondencias entre lenguas. Por esto, hemos decidido utilizar un léxico bilingüe español-catalán existente en la colección AULEX[6], que contiene vocabularios breves en línea de lenguas con recursos limitados, dirigida por Manuel Rodríguez Villegas, especialista compilador de diccionarios en línea.

2.2. FASE 1: Aplicación del algoritmo

El proceso de identificación de traducciones puede ser visto como un alineamiento palabra por palabra. Esta tarea se aborda normalmente mediante algoritmos basados en corpus paralelos, como el modelo IBM (Brown et al., 1993; González-Rubio et al., 2008). Sin embargo, como nosotros basamos nuestro proceso de extracción en corpus comparables (no paralelos), necesitamos otro método. Esta es la razón por la que nos centramos en la información contextual de la palabra que se quiere traducir y candidatos a traducciones. Nuestra aproximación se basa en las palabras adyacentes, asumiendo que podemos traducir parte del contexto del vocabulario. De hecho, como no se pueden traducir todas las unidades léxicas existentes alrededor de los candidatos en la lengua fuente y la lengua de llegada, necesitamos capturar la información más importante en las coocurrencias detectadas. Usamos medidas de normalización para resaltar las particularidades de las coocurrencias entre una palabra (léxico I) y las palabras del léxico "pivote" (léxico II).

En resumen, el método para identificar traducciones basado en la información contextual incluye cuatro pasos:

- cálculo de las coocurrencias entre una palabra (léxico I) y las palabras del léxico "pivote" (léxico II),
- normalización de las coocurrencias con una medida de asociación,
- construcción de un vector de contexto,
- comparación de los vectores de la lengua de partida y la lengua de llegada con una medida de similitud.

La Figura 1 resume el proceso general de extracción de traducción que presentamos en este trabajo.

El primer paso está basado en la premisa de que una palabra y su traducción comparten similitudes contextuales en corpus comparables. Las palabras del léxico "pivote"

[6] http://aulex.org/aulex.php

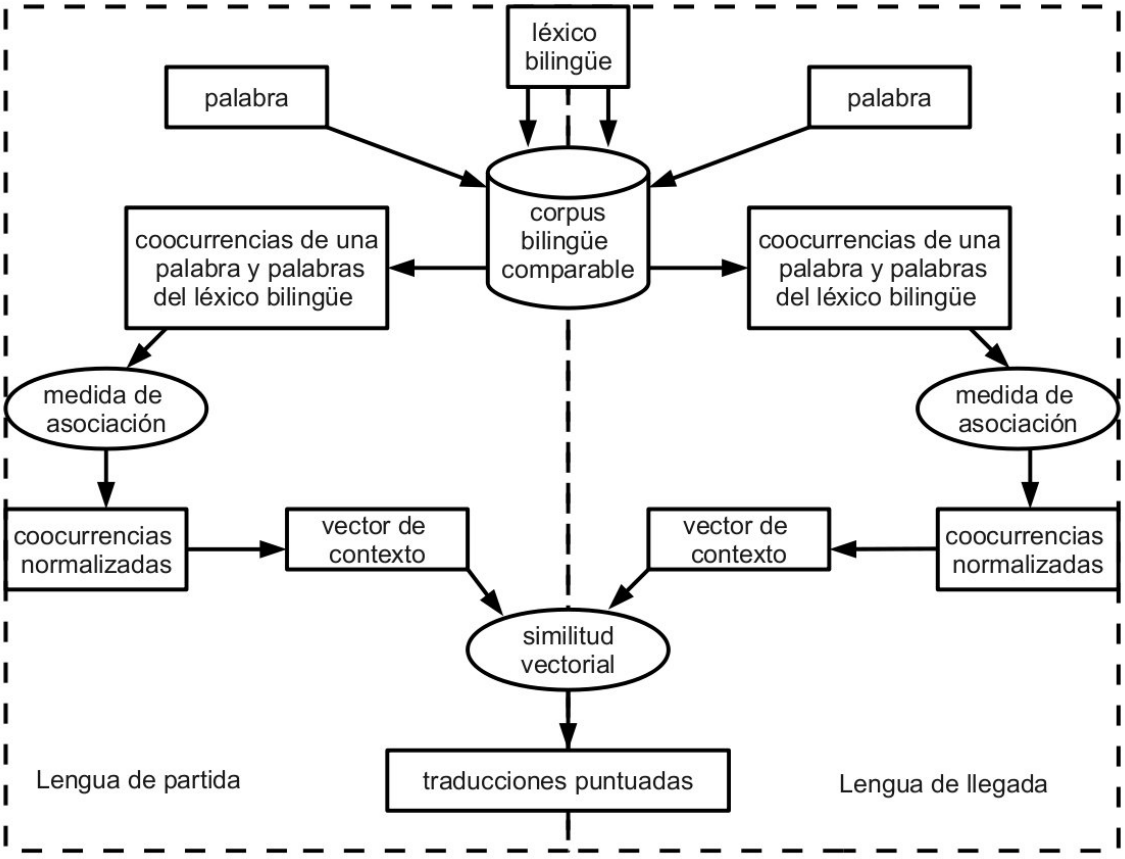


Figura 1: Esquema general del proceso de extracción de traducciones.

(léxico II) son los elementos de relación en ambas lenguas para modelizar el espacio contextual de donde vamos a extraer las traducciones. Las coocurrencias entre una palabra (léxico I) y las palabras del léxico "pivote" (léxico II) se contabilizan dentro de una ventana deslizante de un tamaño fijo (de 10 a 30 palabras en cada ejemplo) o dinámico (oraciones, párrafos, etc.).

El segundo paso ha sido ya ampliamente estudiado en la literatura. Se han probado diversas medidas de asociación, basadas en tablas de contingencia 2*2 como la mostrada en el Cuadro 1, y se observa que las más efectivas son información mutua (Church y Hanks, 1990), *log-likelihood* (Dunning, 1993) y *odds-ratio* (Evert, 2004). En la Sección 3 presentamos los resultados obtenidos con las medidas de información mutua y *odds-ratio*, cuyas fórmulas ofrecemos en la Ecuación 1 y 2, respectivamente.

$$mi(w,s) = log\frac{a}{(a+b)(a+c)} \qquad (1)$$

$$odds(w,s) = log\frac{(a+\frac{1}{2})(d+\frac{1}{2})}{(b+\frac{1}{2})(c+\frac{1}{2})} \qquad (2)$$

	s	$\overline{s}$
w	$a = occ(\ w\ ,\ s\ )$	$b = occ(\ w\ ,\ \overline{s}\ )$
$\overline{w}$	$c = occ(\ \overline{w}\ ,\ s\ )$	$d = occ(\ \overline{w}\ ,\ \overline{s}\ )$

Cuadro 1: Tabla de contingencias entre dos palabras

El Cuadro 1 contiene las coocurrencias comunes en una ventana de una palabra del léxico I (reflejada como w) y las palabras del léxico "pivote" o II (reflejadas como s), pero también los casos en los que w aparece sin s, s aparece sin w, y finalmente en los que no aparecen juntas. Este paso de normalización es particularmente útil para tratar diferencias entre lenguas en corpus comparables. Por ejemplo, el corpus extraído de la Wikipedia

	Documentos	Unidades léxicas
Candidatos	-	300
Léxico "pivote"	-	1.944
Wikipedia CA	296.465	1.461.325
Wikipedia ES	701.423	3.931.243

Cuadro 2: Recursos empleados para los experimentos

en español contiene una mayor cantidad de unidades léxicas, por eso el número de ocurrencias de palabras es mayor que el número de ocurrencias de su traducción en una lengua con menos recursos (como el catalán).

El tercer paso se refiere básicamente a la modelización de una palabra (léxico I) en un espacio contextual. Para cada palabra (léxico I) en la lengua de partida y de llegada, el contexto se modeliza como un vector de contexto. Cada componente de este vector contiene un cálculo de coocurrencias normalizado. Los componentes tienen que ser fijos porque queremos que las dimensiones sean comparables entre los vectores de la lengua de partida y de llegada.

El cuarto paso se basa en medidas de vectores de similitud para comparar los vectores de contexto en la lengua de partida y de llegada. El objetivo es detectar similitudes entre las asociaciones contextuales de las palabras. Los vectores más similares son traducciones posibles. Estas medidas son otro parámetro bien estudiado en la literatura, y las más populares son el coseno, la distancia euclidiana y la métrica *City Block* (Morin et al., 2007). La fórmula de la distancia del coseno entre los vectores de la lengua de partida y de llegada, con la medida de asociación *odds-ratio*, se detalla en la Ecuación 3 (donde V es un vector, s es la lengua de partida, t es la lengua de llegada, y n es una unidad del léxico "pivote").

$$cosine_{V_s}^{V_t} = \frac{\sum_n odds_n^s odds_n^t}{\sqrt{(\sum_n odds_n^s)^2}\sqrt{(\sum_n odds_n^t)^2}} \quad (3)$$

3. Experimentos y resultados

Para evaluar nuestro método, hemos empleado los recursos incluidos en el Cuadro 2. Hemos extraído aleatoriamente 300 candidatos a traducir del léxico I.

Hemos realizado diversos experimentos empleando las medidas de asociación y las medidas de similitud vectorial, presentadas en 2.2. Observamos que los mejores resultados se obtienen con la utilización de la medida de asociación *odds-ratio* y la similitud de *cosenos*. Los resultados se presentan en el Cuadro 3 (P = Precisión, C = Cobertura, F = F-measure). Consideramos que es interesante presentar también los resultados obtenidos con las otras medidas de asociación, como las coocurrencias y la información mutua.

A continuación mostramos algunos ejemplos de traducciones correctas:

CATALÁN	ESPAÑOL
formatge blau	*queso azul*
floridura	*moho*
momificació	*embalsamamiento*
senglar calidó	*jabalí calidón*
vaga	*huelga*

Y también ejemplos de traducciones incorrectas:

CATALÁN	ESPAÑOL
creu nòrdica	*idioma islandés*
castellà mèxic	*alfabetización*
bombeta elèctrica	*cuenco*
astúries	*labor*
bitxo	*salsa pescado*

Los resultados obtenidos muestran la eficacia en cuanto a la precisión en el rango 1 de la medida *odds ratio* combinada con la similitud de *cosenos*. El aumento de la cobertura según el número de candidatos tenidos en cuenta (un rango entre 5 y 10) implica un descenso significativo de la precisión. El cálculo de la precisión tiene en cuenta el número de unidades léxicas de la lengua de llegada consideradas como una buena traducción. Para el rango 10, por ejemplo, una sola traducción es válida según la referencia (léxico I), pero el sistema ofrece 10. En este rango, la información mutua y *odds ratio* son equivalentes en cuanto a precisión y cobertura.

Estos resultados son difícilmente comprables con los de otros trabajos. Sin embargo, observamos que, para el dominio periodístico, los experimentos de (Rapp, 1999) muestran una precisión del rango 1 del 72 % sobre 100 candidatos evaluados. El autor utiliza un corpus en alemán que contiene 135 millones de palabras y un corpus en inglés que incluye 163 millones. Además, el léxico "pivote" que emplea en sus experimentos contiene 16.380 entradas, es decir, que es muy superior al léxico "pivote" que nosotros empleamos en este tra-

	TOP 1			TOP 5			TOP 10		
	P	*C*	*F*	*P*	*C*	*F*	*P*	*C*	*F*
Coocurrencias	45,00	45,00	45,00	15,33	76,67	25,56	8,17	81,67	14,85
Información mutua	57,67	57,67	57,67	16,60	83,00	27,67	9,07	90,67	16,48
Odds ratio	58,00	58,00	**58,00**	16,47	82,33	27,44	9,07	**90,67**	16,48

Cuadro 3: Resultados obtenidos a tres rangos (mejores 1, 5 y 10 traducciones) por similitud de cosenos entre los vectores de contexto

bajo. De hecho, creemos que la precisión del rango 1 del 58 %, que hemos obtenido, podría mejorarse con un léxico con un mayor número de entradas. Este aspecto está relacionado con la cantidad de recursos disponibles para el catalán, menos dotado que otras lenguas. La evaluación de los candidatos ubicados en el primer rango es el modo más apropiado de observar si el léxico bilingüe extraído podría ser incluido en un sistema de traducción automática. Sin embargo, es necesario mejorar la precisión de los resultados con el objetivo de aportar recursos robustos.

En nuestro trabajo no abordamos la construcción de modelos estadísticos de traducción, sino que nos centramos en la tarea de la extracción de léxico bilingüe. Sin embargo, existen diversos trabajos que se están realizando actualmente por otros autores en relación con el entrenamiento de sistemas de traducción automática con datos no paralelos, obteniendo resultados prometedores (Ravi y Knight, 2011).

4. Conclusiones y trabajo futuro

En este trabajo presentamos un sistema de extracción automática de léxico bilingüe, que aplicamos a un par de lenguas de la Península Ibérica: español-catalán. Para los experimentos no empleamos corpus paralelos, sino corpus comparables usando como recurso la información ofrecida por la Wikipedia, aplicando la idea de las similitudes contextuales entre una unidad léxica y su traducción. Los resultados obtenidos son positivos, dado que se logró traducir correctamente más de la mitad de los candidatos. Además, consideramos que la precisión del rango 1 podrá mejorarse mediante un léxico "pivote" que incluya más unidades léxicas, lo cual planeamos hacer como trabajo futuro.

Creemos que este trabajo es relevante, dado que proponemos un sistema que casi no requiere esfuerzo humano, es rápido y, sobre todo, permite la actualización constante del léxico bilingüe, ya que la Wikipedia se amplía cada día con nuevas entradas. Tomando la Wikipedia como un corpus abierto y en constante evolución, podremos emplear este método para aumentar el léxico de cualquier lengua de la Península Ibérica de una manera dinámica y, así, favorecer el multilingüismo, las relaciones entre lenguas y el desarrollo de herramientas de PLN, como los sistemas de traducción automática. La principal ventaja de la metodología empleada en este trabajo es que es independiente de lengua. Para emplearla en diferentes lenguas solo se necesita un corpus comparable y un léxico "pivote" entre las dos lenguas que se quieren tratar.

Como trabajo futuro, nos gustaría aplicar el sistema sobre otros pares de lenguas. Especialmente, estamos interesados en el español-euskera, dada la gran diferencia ortográfica entre las unidades léxicas de estas dos lenguas. Además, nos gustaría incorporar nuestro sistema de extracción a un sistema de traducción automática, para:

1. realizar una evaluación extrínseca de nuestro sistema,
2. aumentar la cobertura de vocabulario de un traductor automático.

Bibliografía

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, y P.S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Brown, P.F., S.D. Pietra, V.J.D. Pietra, y R.L. Mercer. 1993. The Mathematic of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Cabré, M.T. y R. Estopà. 2009. *Les paraules noves criteris per detectar i mesurar els neologismes*. Eumo editorial.

Caseli, HM y MGV Nunes. 2007. Automatic Induction of Bilingual Lexicons for Machi-

ne Translation. *International Journal of Tranlation*, 19:29–43.

Church, K.W. y P. Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Evert, S. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. tesis, Universität Stuttgart. 353 páginas.

Fung, P. 1995. Compiling Bilingual Lexicon Entries from a Non-parallel English-Chinese Corpus. En *Workshop on Very Large Corpora*, páginas 173–183.

Gamallo, P. y I. González. 2010. Wikipedia as a Multilingual Source of Comparable Corpora. En *LREC Workshop on Building and Using Comparable Corpora*, páginas 19–26.

Gamallo Otero, P. y J.R. Pichel Campos. 2007. Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego. *Lenguaje Natural*, páginas 241–248.

González-Rubio, J., G. Sanchis-Trilles, A. Juan, y F. Casacuberta. 2008. A Novel Alignment Model Inspired on IBM Model 1. En *EAMT*, páginas 47–56.

Guinovart, X.G. y E.S. Fontenla. 2004. Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos. *Lenguaje Natural*, 33:133–140.

Guinovart, X.G. y A. Simoes. 2009. Parallel Corpus-Based Bilingual Terminology Extraction. En *International Conference on Terminology and Artificial Intelligence*.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *MT Summit X*, páginas 79–86.

Levenshtein, V.I. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. En *Soviet Physics Doklady*, páginas 707–710.

Morin, E., B. Daille, K. Takeuchi, y K. Kageura. 2007. Bilingual Terminology Mining-Using Brain, not Brawn Comparable Corpora. En *ACL*, páginas 664–671.

Rapp, R. 1995. Identifying Word Translations in Non-parallel Texts. En *ACL*, páginas 320–322.

Rapp, R. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. En *ACL*, páginas 519–526.

Ravi, S. y K. Knight. 2011. Deciphering Foreign Language. En *ACL*, páginas 12–21.

Smith, J.R., C. Quirk, y K. Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. En *NAACL/HLT*, páginas 403–411.

Tomás, J., J. Bataller, F. Casacuberta, y J. Lloret. 2008. Mining wikipedia as a parallel and comparable corpus. En *Language Forum*.

Vivaldi, J. y H. Rodríguez. 2010. Finding domain terms using wikipedia. En *LREC*, páginas 386–393.

Wu, D. y X. Xia. 1994. Learning an English-Chinese lexicon from a Parallel Corpus. En *AMTA*, páginas 206–213.

A Particle Swarm Optimizer to Cluster Parallel Spanish-English Short-text Corpora*

Un Optimizador basado en Cúmulo de Partículas para el Agrupamiento de Textos Cortos de Colecciones Paralelas en Español-Inglés

Diego Ingaramo, Marcelo Errecalde, Leticia Cagnina
LIDIC Research Group
Universidad Nacional de San Luis
Ej. de los Andes 950
5700 San Luis, Argentina.
{daingara,merreca,lcagnina}@unsl.edu.ar

Paolo Rosso
Natural Language Engineering Lab.
ELiRF, DSIC
Universidad Politécnica de Valencia
Camino de Vera s/n
46022 Valencia, España.
prosso@dsic.upv.es

Resumen: El agrupamiento de textos cortos es actualmente un área importante de investigación debido a su aplicabilidad en la recuperación de información desde la web, generación automática de resúmenes y minería de texto. Estos textos con frecuencia se encuentran disponibles en diferentes lenguajes y en colecciones paralelas multilingüe. Algunos trabajos previos han demostrado la efectividad de un algoritmo optimizador basado en Cúmulo de Partículas, llamado CLUDIPSO, para el agrupamiento de colecciones monolingües de documentos muy cortos. En todos los casos considerados, CLUDIPSO superó la prestación de diferentes algoritmos representativos del estado del arte en el área. Este artículo presenta un estudio preliminar mostrando la prestación de CLUDIPSO en colecciones paralelas en Español-Inglés. La idea es analizar cómo la información bilingüe puede ser incorporada al algoritmo CLUDIPSO y en qué medida esta información puede mejorar los resultados del agrupamiento. Con el objetivo de adaptar CLUDIPSO al ambiente bilingüe, se proponen y evalúan algunas alternativas. Los resultados fueron comparados considerando CLUDIPSO en ambos ambientes, bilingüe y monolingüe. El trabajo experimental muestra que la información bilingüe permite obtener resultados comparables con aquellos obtenidos con colecciones monolingües. Se requiere más trabajo de forma tal de hacer un uso efectivo de esta clase de información.
Palabras clave: Agrupamiento de Textos Cortos, Colecciones Paralelas en Español-Inglés, Optimizador basado en Cúmulo de Partículas.

Abstract: Short-texts clustering is currently an important research area because of its applicability to web information retrieval, text summarization and text mining. These texts are often available in different languages and parallel multilingual corpora. Some previous works have demonstrated the effectiveness of a discrete Particle Swarm Optimizer algorithm, named CLUDIPSO, for clustering monolingual corpora containing very short documents. In all the considered cases, CLUDIPSO outperformed different algorithms representative of the state-of-the-art in the area. This paper presents a preliminary study showing the performance of CLUDIPSO on parallel Spanish-English corpora. The idea is to analyze how this bilingual information can be incorporated in the CLUDIPSO algorithm and to what extent this information can improve the clustering results. In order to adapt CLUDIPSO to a bilingual environment, some alternatives are proposed and evaluated. The results were compared considering CLUDIPSO in both environments, bilingual and monolingual. The experimental work shows that bilingual information allows to obtain just comparable results to those obtained with monolingual corpora. More work is required to make an effective use of this kind of information.
Keywords: Clustering of Short Texts, Parallel Spanish-English Corpora, Particle Swarm Optimizer.

1 Introduction

Vast amounts of information are actually available on internet in documents such as news, academic works, web-repositories, etc. many of which are in a short-text format. Document clustering groups automatically a large set of documents into different clusters. In this context, the clustering of short-text corpora, is one of the most difficult tasks in natural language processing due to the low frequencies of terms in the documents.

In document clustering, the information about categories and correctly categorized documents is not provided in advance. An important consequence of this lack of information is that in realistic document clustering problems, results can not usually be evaluated with typical *external* measures like *F*-Measure and Entropy, because the correct categorizations specified by a human expert are not available. Therefore, the quality of the resulting groups is evaluated with respect to *structural* properties expressed in different *Internal Clustering Validity Measures* (ICVMs). Classical ICVMs used as cluster validity measures include the Dunn and Davies-Bouldin indexes, the *Global Silhouette* (GS) coefficient and, new graph-based measures such as the *Expected Density Measure* and the λ-Measure (see (Ingaramo et al., 2008) for detailed descriptions of these ICVMs).

The use of ICVMs has not been limited to the cluster evaluation stage. Different ICVMs have also been used as explicit *objective functions* that the clustering algorithm attempts to optimize *during* the grouping process. This approach has been adopted, for example, in CLUDIPSO, a discrete Particle Swarm Optimizer (PSO) which obtained in previous work (Ingaramo et al., 2009) interesting results on small short-text collections. This algorithm uses the unsupervised measure GS as objective function to be optimized.

CLUDIPSO, and other techniques to cluster short documents (see for example (Alexandrov, Gelbukh, and Rosso, 2005; Pinto, Benedí, and Rosso, 2007; He et al., 2007; Carullo, Binaghi, and Gallo, ; Hu et al., 2009)), have obtained good results with documents written in the same language, i.e. monolingual environments. However, nowadays, there are many linguistics resources which make available information written in different languages. These resources include for example, parallel and aligned parallel corpora which have been used in different applications such as extraction of word translation equivalents (Ribeiro and Lopes, 2000) and studies of lexical semantics (Sharoff, 2002), among others. However, little effort has been dedicated to analyze if this multilingual information could help to improve the results that classical text analysis methods like text categorization and clustering obtain on monolingual corpora.

In this work, information from parallel Spanish-English corpora is used in short-text clustering and two main research questions are addressed: 1) how this bilingual information can be incorporated in the CLUDIPSO algorithm, and 2) to what extent this information can improve the clustering results. The first aspect is addressed in Section 2.1, where some modifications are introduced in CLUDIPSO to incorporate information from parallel Spanish-English short-text corpora. The second one is analyzed in Section 3 which includes results of CLUDIPSO with Spanish and English documents (taken separately) and results with approaches that simultaneously consider documents written in both languages.

The remainder of the paper is organized as follows. Section 2 describes CLUDIPSO, the PSO-based algorithm under study and the proposed alternatives to bilingual document clustering. Section 3 describes some general features of the corpora used in the experiments, the experimental setup and the analysis of the results obtained from the empirical study. Finally, some general conclusions are drawn and present and future work is discussed in Section 4.

2 The CLUDIPSO Algorithm

CLUDIPSO (CLUstering with a DIscrete Particle Swarm Optimization), is based on a PSO (Eberhart and Kennedy, 1995) algorithm that operates on a population of par-

* This work has been done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQEI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, by MICINN as part of the Text-Enterprise 2.0 project (TIN2009-13391-C04-03) within the Plan I+D+i, and by UPV as part of the PAID-02-10 programme (grant. no. 2257).

ticles. Each particle, in the basic version of PSO, is a real numbers vector which represents a position in the search space defined by the variables corresponding to the problem to solve. The best position found so far for the swarm (*gbest*) and the best position reached by each particle (*pbest*) are recorded at each cycle (iteration of the algorithm). The particles evolve at each cycle using two updating formulas, one for velocity (Equation (1)) and another for position (Equation (2)).

$$v_{id} = w(v_{id} + \gamma_1(pb_{id} - par_{id}) + \gamma_2(pg_d - par_{id})) \quad (1)$$

$$par_{id} = par_{id} + v_{id} \quad (2)$$

where par_{id} is the value of the particle i at the dimension d, v_{id} is the velocity of particle i at the dimension d, w is the inertia factor (Eberhart and Shi, 1998) whose goal is to balance global exploration and local exploitation, γ_1 is the personal learning factor, and γ_2 the social learning factor, both multiplied by 2 different random numbers within the range [0, 1]. pb_{id} is the best position reached by the particle i and pg_d is the best position reached by any particle in the swarm.

In the discrete version CLUDIPSO, each valid clustering is represented with a particle. The particles are n-dimensional integer vectors, where n is the number of documents in the corpus. Since the task was modeled with a discrete approach, a new formula was developed for updating the positions (shown in Equation (3)).

$$par_{id} = pb_{id} \quad (3)$$

where par_{id} is the value of the particle i at the dimension d and pb_{id} is the best position reached by the particle i until that moment. This equation was introduced with the objective of accelerate the convergence velocity of the algorithm (principal incoming of discrete PSO models). It is important to note that in this approach the process of updating particles is not as direct as in the continuous case (basic PSO algorithm). In CLUDIPSO, the updating process is not carried out on all dimensions at each iteration. In order to determine which dimensions of a particle will be updated the following steps are performed: 1) all dimensions of the velocity vector are normalized in the [0, 1] range, according to the process proposed by Hu et al. (Hu, Eberhart, and Shi, 2003) for a discrete PSO version; 2) a random number $r \in [0, 1]$ is calculated; 3) all the dimensions (in the velocity vector) higher than r are selected in the position vector, and updated using the Equation (3).
A Dynamic mutation operator (Cagnina, Esquivel, and Gallard, 2004) is applied with a *pm*-probability calculated with the total number of iterations in the algorithm (*cycles*) and the current cycle number: $pm = max\_pm - \frac{max\_pm - min\_pm}{max\_cycle} * current\_cycle$. Where max_pm and min_pm are the maximum and minimum values that pm can take, max_cycle is the total number of cycles and the current cycle in the iterative process is $current\_cycle$. The mutation operation is applied if the particle is the same that its own *pbest*, as was suggest by (Hu, Eberhart, and Shi, 2003). The mutation operator swaps two random dimensions of the particle and in that way avoids premature convergence.

Global Silhouette (GS) Coefficient was used as an *objective function* $f(p_i)$, because gives a reasonable estimation of the quality of the obtained groups. The optimization of GS drives the entire CLUDIPSO process.
The GS measure combines two key aspects to determine the quality of a given clustering: *cohesion* and *separation*. Cohesion measures how closely related are the objects in a same cluster whereas separation quantifies how distinct (well-separated) a cluster from other clusters is. The GS coefficient of a clustering is the average cluster silhouette of all the obtained groups. The cluster silhouette of a cluster C also is an average silhouette coefficient but, in this case, of all objects belonging to C. Therefore, the fundamental component of this measure is the formula used for determining the silhouette coefficient of any arbitrary object i, that we will refer as $s(i)$ and that is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (4)$$

with $-1 \leq s(i) \leq 1$. The $a(i)$ value denotes the average dissimilarity of the object i to the remaining objects in its own cluster, and $b(i)$ is the average dissimilarity of the object i to all objects in the nearest cluster. From this formula it can be observed that negative values for this measure are undesirable and

that for this coefficient values as close to 1 as possible are desirable.

2.1 CLUDIPSO adaptations to bilingual contexts

The idea of using bilingual information in a clustering algorithm is motivated by similar reasons to those giving origin to *ensemble methods* (Dietterich, 2000): the combined use of information (about the same problem) coming from different sources, can be more effective than consider this information separately. In the context of our work, this intuitive idea consists in determining how the information obtained from the alignment of documents in parallel corpora can be used by a clustering algorithm instead of clustering the documents in the different languages separately.

CLUDIPSO allows to combine this kind of information in a relatively direct way: using the values that the evaluation function obtains with the documents in each language, and combining these values using different criteria. For example, the minimum (MIN), maximum (MAX) or average (AVG) value between the evaluation function's values obtained with the documents in each language could be used. Or simply taking the value that the evaluation function returns with documents in both languages, but alternating in each iteration the language used to evaluate this function. More formally:
Let D be a bilingual environment (parallel corpus) with Spanish-English documents. Then, each document $d_i \in D$ will be represented by an English text, $d_{i_{EN}}$ and the corresponding representation of d_i in Spanish language, $d_{i_{ES}}$. Let D_{EN} and D_{ES} be the documents in D in its English and Spanish representation respectively.

We will denote as $CLUDIPSO_{MULTI}$ the CLUDIPSO version that incorporates D_{EN} and D_{ES} information in the clustering process. $CLUDIPSO_{MULTI}$ uses the available bilingual information by adapting the CLUDIPSO *evaluation function step* (see section 2). The four alternatives that were considered to incorporate the bilingual information in the evaluation function are described below.

Let p be a particle representing a possible solution (clustering) and let $f(p_{en})$ and $f(p_{es})$ be the fitness values of p with respect to D_{EN} and D_{ES} respectively. Then, the fitness value for:

1. $CLUDIPSO_{MULTI-MAX}$ is defined as: $f(p) = max(f(p_{en}), f(p_{es}))$.

2. $CLUDIPSO_{MULTI-MIN}$ is defined as: $f(p) = min(f(p_{en}), f(p_{es}))$.

3. $CLUDIPSO_{MULTI-AVG}$ is defined as: $f(p) = \frac{f(p_{en})+f(p_{es})}{2}$.

4. $CLUDIPSO_{MULTI-ALT}$ in the iteration i is $f(p_{en})$ if i is odd and $f(p_{es})$ in other case.

Thus, for example, if the Silhouette Coefficient is used as evaluation function, $CLUDIPSO_{MULTI-AVG}$ will use as fitness value the average value obtained from the Silhouette value for the clustering p_i using the English documents in D and the Silhouette value for the same clustering, but using in this case the Spanish documents.

3 Experimental Setting and Analysis of Results

For the experimental work, two collections with different levels of complexity with respect to the size, length of documents and vocabulary overlapping were selected: SEPLN-CICLing and JRC-Acquis. Table 1 shows some general features of these corpora: corpus size (CS), number of categories and documents ($|\mathcal{C}|$ and $|\mathcal{D}|$ respectively), total number of terms in the collection ($|\mathcal{T}|$), vocabulary size ($|\mathcal{V}|$) and average number of terms per document ($\overline{\mathcal{T}}_d$).

The first one, SEPLN-CICLing, is a small collection based on CICLing-2002[1] scientific abstract corpus which has been intensively used in different works (Ingaramo et al., 2009; Ingaramo, Errecalde, and Rosso, 2010; Errecalde, Ingaramo, and Rosso, 2010). This corpus was enriched with bilingual abstracts (in Spanish and English) of the SEPLN[2].

The SEPLN-CICLing$_{EN}$ corpus was composed by the English abstracts of CICLing-2002 and SEPLN. The SEPLN-CICLing$_{ES}$ was obtained adding to the Spanish version of SEPLN abstracts the manual translation of the CICLing abstracts.

[1] http://www.cicling.org/2002/
[2] http://www.sepln.org/

JRC-Acquis refers to a sub-collection of the Acquis (Steinberger et al., 2006), a popular multilingual collection with legal documents and laws corresponding to different countries of the European Union. For this work, we selected 563 documents in the English and Spanish versions, denoted JRC-Acquis$_{EN}$ and JRC-Acquis$_{ES}$ respectively.

Corpora	CS	$\|\mathcal{C}\|$	$\|\mathcal{D}\|$
SEPLN-CICLing$_{EN}$	25	4	48
SEPLN-CICLing$_{ES}$	21	4	48
JRC-Acquis$_{EN}$	798	6	563
JRC-Acquis$_{ES}$	870	6	563
Corpora	$\|\mathcal{T}\|$	$\|\mathcal{V}\|$	$\hat{\mathcal{T}}_d$
SEPLN-CICLing$_{EN}$	3143	1169	65.48
SEPLN-CICLing$_{ES}$	2129	904	44.35
JRC-Acquis$_{EN}$	110887	7391	196.96
JRC-Acquis$_{ES}$	121953	7903	216.61

Table 1: Features of the collections used in the experimental work.

Due to the fact the gold standard is known for each of the two sub-collections, the quality of the results was evaluated by using the classical (external) F-measure. Each algorithm generated 50 independent runs per collection after performing 10,000 iterations (stopping condition). The reported results in Table 2, correspond to the minimum (F_{min}), maximum (F_{max}) and average (F_{avg}) F-measure values obtained by the different the algorithm of CLUDIPSO. The values highlighted in bold, indicate the best obtained results.

	SEPLN-CICLing		
Algorithm	F_{avg}	F_{min}	F_{max}
$CLUDIPSO-EN$	**0.75**	**0.63**	0.85
$CLUDIPSO-ES$	0.7	0.6	0.83
$CLUDIPSO_{MULTI-ALT}$	0.52	0.39	0.68
$CLUDIPSO_{MULTI-AVG}$	0.7	**0.63**	0.79
$CLUDIPSO_{MULTI-MAX}$	0.71	0.62	0.83
$CLUDIPSO_{MULTI-MIN}$	0.7	0.62	**0.87**
	JRC-Acquis		
Algorithm	F_{avg}	F_{min}	F_{max}
$CLUDIPSO-EN$	**0.29**	**0.26**	**0.33**
$CLUDIPSO-ES$	**0.29**	0.21	0.31
$CLUDIPSO_{MULTI-ALT}$	**0.29**	**0.26**	0.32
$CLUDIPSO_{MULTI-AVG}$	0.28	**0.26**	0.31
$CLUDIPSO_{MULTI-MAX}$	**0.29**	0.25	0.32
$CLUDIPSO_{MULTI-MIN}$	**0.29**	0.25	0.32

Table 2: F-measures values per collection.

With the smaller SEPLN-CICLing collection it is observed that the version $CLUDIPSO-EN$ obtained the best F_{avg} value and $CLUDIPSO_{MULTI}$ the best F_{max} value. Similar results can be observed with the F_{min} values in both cases. It should be noted that $CLUDIPSO_{MULTI}$ slightly overcomes the algorithm $CLUDIPSO-ES$ but not the $CLUDIPSO-EN$ although both algorithms have a similar performance no matter the language used.

With respect to the obtained results with the larger collection JRC-Acquis, $CLUDIPSO_{MULTI}$ gets similar values to $CLUDIPSO-EN$ and a minimum improvement compared to $CLUDIPSO-ES$ (like in SEPLN-CICLing). It should be noted that $CLUDIPSO_{MULTI}$ improves F_{min} in all the cases, excluding $CLUDIPSO_{MULTI-ALT}$. Experiments carried out show an improvement related to $CLUDIPSO-ES$ but results are similar to $CLUDIPSO-EN$. However, in JRC-Acquis results needs to be improved for both languages.

4 Conclusions and Future Work

This work presents a preliminary study of performance of different versions of $CLUDIPSO_{MULTI}$, a novel bilingual PSO-based clustering algorithm. The results obtained by $CLUDIPSO_{MULTI}$ on Spanish-English corpora indicate that the approach is an alternative to solve bilingual clustering of small short-text corpora, although no significant improvement was obtained so far with respect to the monolingual PSO-based version CLUDIPSO. $CLUDIPSO_{MULTI}$ was also tested with a larger size collection and the performance was comparable to its predecessor monolingual CLUDIPSO, possibly the lack of improvement it is due to the limitations derived by a wide search space in large document collection.

Future works include text-enrichment of documents, combining both documents representations by using a term selection technique and also, including bilingual information into a novel on going version named $CLUDIPSO^{\star}$ considering newer mechanisms to incorporate the bilingual knowledge.

The proposed algorithm was tested with Spanish-English corpora although other bilingual corpora could be used in a future.

In order to tackle the problem of the size of the particle that $CLUDIPSO_{MULTI}$ suffers with collection such as JRC-Acquis,

we aim at investigating the possibility of dividing the particle in two: a part of the particle would deal with the representation in English and the other one with the Spanish representation.

References

Alexandrov, M., A. Gelbukh, and P. Rosso. 2005. An approach to clustering abstracts. In Andrés Montoyo, Rafael Muñoz, and Elisabeth Metais, editors, *Natural Language Processing and Information Systems*, volume 3513 of *LNCS*. Springer Berlin / Heidelberg, pages 1–10.

Cagnina, L., S. Esquivel, and R. Gallard. 2004. Particle swarm optimization for sequencing problems: a case study. In *Congress on Evolutionary Computation*, pages 536–541.

Carullo, M., E. Binaghi, and I. Gallo. An online document clustering technique for short web contents. *Pattern Recognition Letters*, 30.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *Int. Workshop on Multiple Classifier Systems*, pages 1–15. Springer-Verlag.

Eberhart, R. and J. Kennedy. 1995. A new optimizer using particle swarm theory. In *Proc. of the Sixth International Symposium on Micro Machine and Human Science, MHS'95*, pages 39–43, Nagoya, Japan.

Eberhart, R. and Y. Shi. 1998. A modified particle swarm optimizer. In *International Conference on Evolutionary Computation*. IEEE Service Center.

Errecalde, M., D. Ingaramo, and P. Rosso. 2010. Itsa*: an effective iterative method for short-text clustering tasks. In *Proc. of the 23rd Int. Conf. on Industrial Engineering and other Applications of Applied Intelligent Systems, IEA/AIE 2010*, pages 550–559, Berlin, Heidelberg. Springer-Verlag.

He, H., B. Chen, W. Xu, and J. Guo. 2007. Short text feature extraction and clustering for web topic mining. In *Proc. of the Third Int. Conf. on Semantics, Knowledge and Grid*, pages 382–385, Washington, DC, USA. IEEE Computer Society.

Hu, X., R. Eberhart, and Y. Shi. 2003. Swarm intelligence for permutation optimization: a case study on n-queens problem. In *Proc. of the IEEE Swarm Intelligence Symposium*, pages 243–246.

Hu, X., N. Sun, C. Zhang, and T. Chua. 2009. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proc. of the 18th ACM Conf. on Information and knowledge management*, pages 919–928, New York, NY, USA. ACM.

Ingaramo, D., M. Errecalde, L. Cagnina, and P. Rosso, 2009. *Computational Intelligence and Bioengineering*, chapter Particle Swarm Optimization for clustering short-text corpora, pages 3–19. IOS press.

Ingaramo, D., M. Errecalde, and P. Rosso. 2010. A general bio-inspired method to improve the short-text clustering task. In *Proc. of CICLing 2010*, LNCS 6008, pages 661–672. Springer-Verlag.

Ingaramo, D., David Pinto, P. Rosso, and M. Errecalde. 2008. Evaluation of internal validity measures in short-text corpora. In *Proc. of the International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2008*, volume 4919 of *Lecture Notes in Computer Science*, pages 555–567. Springer-Verlag.

Pinto, D., J. M. Benedí, and P. Rosso. 2007. Clustering narrow-domain short texts by using the Kullback-Leibler distance. In *Proc. of the International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2007*, volume 4394 of *Lecture Notes in Computer Science*, pages 611–622. Springer-Verlag.

Ribeiro, A. and G. Pereira Lopes. 2000. Extracting portuguese-spanish word translations from aligned parallel texts. *Procesamiento del lenguaje natural*, 26:73–80.

Sharoff, S. 2002. Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. In *Proc. of Language Resources and Evaluation Conference (LREC02)*, pages 447–452.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy.

Cross-language Semantic Relations between English and Portuguese*

Relaciones Semánticas entre los Idiomas Inglés y Portugués

Anabela Barreiro
L2F – INESC-ID
Rua Alves Redol nº 9, 1000-029
Lisboa, Portugal
anabela.barreiro@l2f.inesc-id.pt

Hugo Gonçalo Oliveira
CISUC, University of Coimbra, Pólo II
Pinhal de Marrocos 3030-290
Coimbra, Portugal
hroliv@dei.uc.pt

Resumen: Este artículo describe las relaciones semánticas conceptuales obtenidas de los recursos del sistema OpenLogos que fueron convertidos al formato NooJ. Estas relaciones están representadas simbólicamente en el léxico OpenLogos como un esquema taxonómico llamado abstracción semántico-sintáctica del lenguaje (SAL), que se utiliza para generar las relaciones jerárquicas de hiponimia e hiperonimia. El artículo también describe las relaciones acción-de, resultado-de, y sinonimia entre unidades multi-palabra y palabras sueltas, sobre todo donde existe una relación morfo-sintáctica y semántica entre las palabras de distintas categorías gramaticales. Las relaciones semánticas se generaron automáticamente a partir de la información lingüística asociada a cada entrada lexical en los diccionarios NooJ. Se desarrollaron gramáticas locales como mecanismo para leer esta información lingüística y generar las relaciones semánticas que se han utilizado en la producción de paráfrasis y en traducción automática. Los diccionarios y las gramáticas se pueden adaptar fácilmente a distintas lenguas y son útiles para diferentes tareas de procesamiento natural de la lengua, tanto monolingües como entre idiomas.
Palabras clave: relaciones semánticas, ontologías, diccionarios, gramáticas locales, relaciones entre idiomas

Abstract: This paper describes conceptual semantic relations obtained from OpenLogos resources converted into NooJ format. These relations were symbolically represented in the OpenLogos lexicon as a taxonomic scheme called semantico-syntactic abstraction language (SAL), used to generate hierarchical hyponymy and hypernymy relations. The paper also describes action-of, result-of, and synonymy relations between multiword units and single words, mostly where there is a morpho-syntactic and semantic relation between words of distinct parts-of-speech. The semantic relations were generated automatically, based on the linguistic information associated with each lexical entry in NooJ dictionaries. Local grammars were developed as a mechanism to read this linguistic information and generate the semantic relations, which have been used in paraphrasing and machine translation. Dictionaries and grammars can easily be adapted to distinct languages and are useful to various natural language processing monolingual or cross-language tasks.
Keywords: semantic relations, ontologies, dictionaries, local grammars, cross-language relations

1 Introduction

Lexical Semantics (Cruse, 1986) is the subfield of semantics that studies the words of a language and their meanings. It sees the lexicon as a finite list of lexical items (words or expressions) with a highly systematic structure that controls what words can mean. It can be seen as the bridge between a language and the knowledge expressed in that language (Sowa, 1999). The conceptual model of a language is structured around lexical items, their meaning (often referred as sense) and lexico-semantic relations held between

* Anabela Barreiro was partially supported by the UPV, award 1931, under the program Research Visits for Renowned Scientists (PAID-02-11). Hugo Gonçalo Oliveira is supported by the FCT scholarship grant SFRH/BD/44955/2008, co-funded by FSE.

the latter. To deal with the meaning of a language it is important to study these relations.

Semantic relations are crucial to understand and to structure the meaning of natural language. They are vital to communication overall, and highly employed in technical and specialized domains, where the most important content of texts is conveyed through the semantic relations between the terms that represent the domain's concepts, rather than by the meaning of the words alone (e.g., the semantic relations between BRCA1/*protein* and RNF53/*gene* in the biomedical field). Additionally, semantic relations are important for applications in the semantic web, mapping ontologies, text categorization, natural language understanding, etc., and a requisite for paraphrasing and machine translation, where words and expressions often must be substituted by semantic equivalents, such as synonyms between support verb constructions and single verbs (*make an operation* = *operate*; *say hello to* = *greet*), or other type of semantic alternates.

The most studied lexico-semantic relations are: (1) synonymy, when different lexical items have the same meaning (e.g. *car* synonym-of *automobile*); (2) homonymy, when lexical items have the same orthographic form but different meanings (e.g. *bank*, financial institution vs. *slope*); (3) hyponymy, when a lexical item is a subclass or a specific kind of another (e.g. *dog* hyponym-of *mammal*); and (4) meronymy, when a lexical item is a part, piece or member of another (e.g. *wheel* part-of *car*).

This paper describes the first attempt to extract cross-language semantic relations between English and Portuguese from the lexical resources of the OpenLogos machine translation system described by Scott (2003) and Barreiro et al. (2011). In combination with the former resources, new resources were created, namely derivational rules and grammars to recognize and generate morpho-syntactic and semantically related words and multiword units. Semantic relations, obtained by means of local grammars developed within NooJ linguistic environment (Silberztein, 2007), cover a larger number of items and can be extracted in a simple and easy way. This paper aims at showing how these resources combined can be used in cross-language tasks. Section 2 describes the state of the art in lexical semantics and automatic acquisition of distinct types of lexico-semantic relations. Section 3 presents the base linguistic resources used to attain semantic relations. Section 4 describes the relations of synonymy, hyponymy, action-of, and result-of. Section 5 presents the method for the extraction of the semantic relations. It describes, in particular, the morpho-syntactic and semantic relations established in the dictionary, how the grammars read this linguistic information, and how they use it to generate semantic pairs. This latter section also shows how to expand from monolingual to cross-language relations with minimal change in the local grammars. Section 6 presents some preliminary results. And finally, section 7 presents the conclusions and guidelines for future research work.

2 State of the Art

Dictionaries are probably the main source of lexico-semantic knowledge, as they are repositories of words, which include the description of several word senses. However, as definitions are written in natural language, dictionaries are not completely ready for being used as computational lexical resources.

Common representations of lexico-semantic knowledge, ready for being used in natural language processing tasks, include thesauri, taxonomies, as well as lexical ontologies or lexical knowledge bases. For example, the Roget Thesaurus (Roget, 1852) is one of the most well-known and complete thesaurus that is available in a machine readable format. Also, Princeton WordNet (Fellbaum, 1998) is a public domain lexical knowledge base, widely used in the natural language processing community. It is a handcrafted resource based on synsets, which are groups of synonymous words that may be seen as natural language concepts. Each synset has a gloss, which is similar to a dictionary definition, and several types of semantic relations between synsets are represented.

As the manual creation of lexical knowledge bases is typically an extensive and time-consuming task, there are several works where lexico-semantic relations are extracted automatically from text, and then used either to create new knowledge bases from scratch or to enrich existing knowledge bases. Due to their structure, dictionaries are an obvious

target for the extraction of lexico-semantic relations (see, for example, (Chodorow, Byrd, and Heidorn, 1985) or (Richardson, Dolan, and Vanderwende, 1998)). Corpora and the Web have as well been exploited in the automatic acquisition of several types of lexico-semantic relations, including hyponymy (Hearst, 1992), meronymy (Berland and Charniak, 1999), causal relations (Girju and Moldovan, 2002), as well as in the discovery of new concepts (Lin and Pantel, 2002).

For Portuguese, in the latest years, semantic relations have also been a subject of increasing research interest. Santos et al. (2010) provide a review of the existing Portuguese lexico-semantic resources. Briefly, there are two handcrafted wordnets for European Portuguese, namely WordNet.PT (Marrafa, 2002) and MWN.PT[1], and an electronic thesaurus for Brazilian Portuguese, TeP (Maziero et al., 2008). There have also been attempts to the automatic acquisition of semantic relations, including: hyponymy extraction from corpora (Freitas and Quental, 2007); the extraction of several relations from a dictionary and the creation of the lexical resource PAPEL (Gonçalo Oliveira, Santos, and Gomes, 2010); and Onto.PT (Gonçalo Oliveira and Gomes, 2010), an ongoing project on the automatic creation of a lexical ontology for Portuguese, where several textual resources (thesauri, dictionaries, encyclopedias) are being exploited in the automatic acquisition of lexico-semantic relations.

Still, to the best of our knowledge, no research has been published on the automatic generation of cross-language semantic relations by using a linguistic method to map syntactic and semantically related words. This method can be extended to the type of relations that set equivalence between a word and a multiword unit (e.g. *take a look* = *look*), with a relative clause (*that was corrected* = *corrected*), with complex compounds (*bottle made of plastic* = *plastic bottle*) or even with a more complex construction, such as a possessive construction or a passive, by exploiting the morpho-syntactic and semantic relations pairs described in the dictionaries. The method has the advantage of being systematic, expandable, holding an unlimited possibility to grow and improve in observance of natural language complexity and compliant to distinct languages and across languages. This is the novel aspect of the work presented in this paper in relation to the state of the art.

[1]See http://mwnpt.di.fc.ul.pt/

3 Resources

In this section, we will describe the English and Portuguese resources used to achieve cross-language semantic relations.

Eng4NooJ and Port4NooJ (Barreiro, 2007) are sets of resources developed with the NooJ linguistic environment (Silberztein, 2007), aiming at the processing of the English and Portuguese languages. Both Eng4NooJ and Port4NooJ resources include lexica and grammars which are used for different tasks, including morphological and semantico-syntactic analysis, disambiguation, paraphrasing and translation. Both include a morphological system, contextual rules, different types of grammars (disambiguation, multiword units, etc.), and domain-specific dictionaries.

The Port4NooJ resources are publicly available[2] and, at the moment, are being used in tools such as Corpógrafo, a corpora tool (Maia and Sarmento, 2005; Sarmento et al., 2006; Maia and Matos, 2008), ParaMT, a paraphraser for machine translation (Barreiro, 2008a; Barreiro, 2008b), and eSPERTo[3], a system of paraphrasing for text editing and revision, currently being integrated in a cyber-school pedagogical program. Port4NooJ resources have not been reviewed, but they were made available to the Portuguese natural language processing (NLP) community because of their novelty aspects, which we hope are evocative for further pioneering research, including exploitation to other languages and cross-language tasks. The semantic relations included in the

[2]Port4NooJ can be found at the NooJ website under Portuguese module (http://www.nooj4nlp.net) and its resources are also available at Linguateca since October 2008 (http://www.linguateca.pt/Repositorio/Port4NooJ/).

[3]eSPERTo (in Portuguese, stands for Sistema de Parafraseamento para Edição e Revisão de Texto). It is a derivative of ReEscreve, proposed by Barreiro (2008a), and also described in (Barreiro and Cabral, 2009). The English version of eSPERTo is called SPIDER, standing for a System of Paraphrasing In Document Editing and Revision (formerly ReWriter). SPIDER uses Eng4NooJ resources and is described in (Barreiro, 2011).

Port4NooJ and Eng4NooJ resources resulted from the application of simple local grammars to the semantico-syntactic properties in the lexical entries and the use of derivational rules that link semantically related words of different parts-of-speech.

Eng4NooJ and Port4NooJ lexica were inherited from the OpenLogos system and enhanced with several new properties, which will be described in detail in Section 5.

The OpenLogos lexical entries are classified with more than 1,000 distinct categories, based on a taxonomy called SAL (*Semantico-syntactic Abstraction Language*)[4]. In the OpenLogos model, SAL is a meta-language that represents natural language, in effect, an ontology that represents things, ideas, relationships, dispositions, conditions, processes, etc., as well as the elements of grammar such as articles, prepositions, conjunctions, etc. In terms of natural language processing, the meta-language represents both syntax and semantics. SAL is an actual language, not a set of linguistic markers or primitives. This implies that natural language can be readily mapped to SAL. The granularity of the representational ontology is sufficient for translation purposes only, i.e., the ontology does not need to be especially fine-grained.

SAL elements are divided in a hierarchical scheme of supersets, sets and subsets, distributed by all parts-of-speech. SAL comprises 12 supersets for nouns: Concrete (CO), Mass (MA), Animate (AN), Place (PL), Information (IN), Abstract (AB), Process intransitive (PI), Process transitive (PT), Measure (ME), Time (TI), Aspective (AS), and Unknown (UN). For example, the concrete nouns superset consists of countable physical things, either man-made or natural, including parts of the human body. Concrete (count[5]) contain both sets and subsets. The principal sets of concrete nouns are functional things and agentive things. Other sets are: natural things (COnat); impulses/lights (COlight); marks/blemishes (COblem); edibles non-mass (COednm); edibles/color (COedcol); classifiers (COclass); amorphous (COamorph); and atomistic (COatom). For example, the set of natural things (COnat) includes subsets such as: minute flora (COflora) (e.g. *algae*, *spore*); plants (COplant) (e.g. *rose*, *weed*); trees (COtree) (e.g. *apple*, *willow*); trees/wood (COtrwd) (e.g. *oak*, *maple*); and miscellaneous natural things (COmnat) (e.g. *pebble*, *iceberg*).

The SAL meta-language is semantico-syntactic in nature, representing natural language at a second-order abstractions (common nouns are first-order abstractions). Syntax and semantics are seen as a continuum. This semantico-syntactic continuum is always taken into account when classifying each lexical entry within SAL. The classification was done through the years by trial and error. For example, when classifying elements into the functional (COfunc) or agentive (COagen) of the concrete noun superset, the following reasoning is taken into consideration: functional things tend to be passive, i.e. typically do not act of their own accord and generally require an agent to use them. Hence, they are more instrumental in nature. Agents typically do work in and of themselves. This distinction may sometimes seem arbitrary. For example, *hinge* is a fastener under functional things and clearly does work of itself, but is not coded as an agent. *Airplane*, on the other hand, obviously does require an agent and yet is coded under agentives as a vehicle. As a rule, agentives have a source of power or energy in themselves, while functionals do not. Parts of the human/animal body are also classified as concrete. Words like *heart*, *brain*, *digestive tract*, *stomach*, and organs in general are machines/systems under agentives. Words like *teeth*, *fingernail*, *toes*, *lips*, *tendons*, *ligaments*, *bones*, etc. belong to various subsets under functionals.

SAL categories contain domain-independent ontological (lexical-contextual) and semantico-syntactic relations (the same word form can be mapped to different concepts) are assigned to general language words or domain-specific terms. The general language dictionary contains many lexical entries which are broadly classified, which could be considered to pertain to a more specific domain. For example, the lexical entries

[4]The full description of the multiple SAL categories can be found at the Logos System Archives (http://logossystemarchives.homestead.com/) and all the resources (and descriptions) are downloadable from OpenLogos website at DFKI (http://logos-os.dfki.de/).

[5]Concrete nouns are always count nouns and, unless in the plural, generally cannot occur without a preceding article or quantifier. For example: *Computers are effective.* **Computer is effective.*

dog IS_HYPONYM_OF *animal* *cão* É_HIPÓNIMO_DE *animal*
dog IS_HYPONYM_OF *mammal* *cão* É_HIPÓNIMO_DE *mamífero*
dog IS_HYPONYM_OF *non-human being* *cão* É_HIPÓNIMO_DE *ser não humano*
dog IS_HYPONYM_OF *invertebrate* *cão* É_HIPÓNIMO_DE *ser vertebrado*
dog IS_HYPONYM_OF *animate being* *cão* É_HIPÓNIMO_DE *ser vivo/animado*

Table 2: Hyponymy relations for the noun *dog - cão*

for *HIV* (immunology), *manic-depressive disorder*, *bipolar disorder* (mental health) and *asthma* (pulmonology) are all classified under the superset Abstract and subset State (also for conditions and relationships). This subset corresponds to abstract nouns that describe something about a thing or person that is not inherent to its nature (e.g. *cancer*, *coma*, *circumstance*, *condition*, *disease*, *fatherhood*, *inequality*, *insolvency*, *loneliness*, *parity*, *poverty*, *status*). Being more extrinsic, these states, conditions or relationships could conceivably change without altering the nature of the thing or person. This is not a strict rule but is indicative of the difference between this subset and the properties/qualities/nature subset.

The information noun superset is comprised of nouns that denote data, information, or knowledge, which might be considered more specific to certain domains. But, this category also includes the medium on which the information is recorded, represented or communicated; i.e., spoken, written, dramatized, sung, etc. Table 1 presents a list of terms classified as Instructional/legal (INinst) under the information noun superset (IN).

4 Semantic Relations for English and Portuguese

Both in Eng4NooJ and Port4NooJ, each lexical entry is described with semantico-syntactic properties, which represent relations between words or expressions. These relations can be synonymy, hyponymy, action-of, result-of, process-of, made-of, property-of, member-of, among others. Table 2 illustrates several semantic relations for the concrete English and Portuguese nouns *dog* and *cão*, respectively. These relations were inferred from the SAL hierarchical categories.

A C T I O N	abolishment IS_ACTION_OF abolish aboliçao É AÇAO DE abolir
	abuse IS ACTION OF abuse abuso É AÇAO DE abusar
	happening IS ACTION OF happen acontecimento É AÇAO DE acontecer
	agreement IS ACTION OF agree acordo É AÇAO DE acordar
R E S U L T	lit IS RESULT OF light aceso É RESULTADO DE acender
	stu ed IS RESULT OF stu embalsamado É RESULTADO DE embalsamar
	rotten IS RESULT OF rotten podre É RESULTADO DE apodrecer
	interdicted IS RESULT OF interdict interditado É RESULTADO DE interditar

Table 3: Action-of and result-of semantic relations

In addition to the taxonomical classification inherited from OpenLogos, which allowed the establishment of hyponymy relations, both Eng4NooJ and Port4NooJ resources include regular derivational, morpho-syntactic and semantic relations, such as synonymy, action-of, and result-of. The morpho-syntactic and semantic relations are established between words of a different part-of-speech, as for example, between an adjective and its derived adverb (e.g. *quick* > *quickly* - *rápido* > *rapidamente*), between a noun and an adjective (e.g. *enthusiasm* > *enthusiastic* - *entusiasmo* > *entusiasmado*), or between a noun and an adverb (e.g. *imagination* > *imaginatively* = *with imagination* - *imaginação* > *imaginativamente* = *com imaginação*).

Table 3 illustrates action-of and result-of semantic relations. Action-of relations are established between a noun and a verb, where the noun is a morphological derivation of the verb. Result-of relations are established between an adjective and a verb, where the adjective is morphologically derived from the verb.

5 Methodology for the Extraction of Semantic Relations

In order to obtain hyponymy relations from the OpenLogos properties in Port4NooJ and Eng4NooJ dictionaries, we created a local grammar that matches on the SAL code and presents, as an output, one or more words from the description of that specific SAL code. For the examples in Table 1, the NooJ local grammar recognizes the property [SAL=ANmamm], standing for Ani-

intimacão,N+FLX=CANCÃO+ININST+EN=summons	garantia,N+FLX=CASA+ININST+EN=guarantee
arrendamento,N+FLX=ANO+ININST+EN=lease	garantia,N+FLX=CASA+ININST+EN=warranty
autorizacão,N+FLX=CANCÃO+ININST+EN=fiat	lei,N+FLX=CASA+ININST+EN=law
autorizacão,N+FLX=CANCÃO+ININST+EN=license	licenca,N+FLX=CASA+ININST+EN=license
autorizacão,N+FLX=CANCÃO+ININST+EN=permit	mandato,N+FLX=ANO+ININST+EN=mandate
autorizacão,N+FLX=CANCÃO+ININST+EN=warrant	moratoria,N+FLX=CASA+ININST+EN=moratorium
cânone,N+FLX=ANO+ININST+EN=canon	norma,N+FLX=CASA+ININST+EN=norm
clausula,N+FLX=CASA+ININST+EN=clause	norma,N+FLX=CASA+ININST+EN=standard
condicão,N+FLX=CANCÃO+ININST+EN=proviso	ordem,N+FLX=MARGEM+ININST+EN=order
contrato,N+FLX=ANO+ININST+EN=contract	ordem,N+FLX=MARGEM+ININST+EN=ordinance
credo,N+FLX=ANO+ININST+EN=credo	pacto,N+FLX=ANO+ININST+EN=pact
declaracão,N+FLX=CANCÃO+ININST+EN=affidavit	patente,N+FLX=CASA+ININST+EN=patent
decreto,N+FLX=ANO+ININST+EN=decree	renuncia,N+FLX=CASA+ININST+EN=waiver
diretiva,N+FLX=CASA+ININST+EN=guideline	testamento,N+FLX=ANO+ININST+EN=will
estatuto,N+FLX=ANO+ININST+EN=bylaw	tratado,N+FLX=ANO+ININST+EN=treaty
estatuto,N+FLX=ANO+ININST+EN=statute	veredicto,N+FLX=ANO+ININST+EN=veredict

Table 1: Sample of terms classified as Information + Instructional/legal (ININST)

mate, Mammal and retrieves, as its output, words that will be used as hypernyms of the words *dog* or *cão*, in English or Portuguese, respectively. These words are: animal, mammal, non-human being, invertebrate, animate being. If the description of the SAL category included more hypernyms, these could, of course, be easily added to the list of pairs of the semantic relation IS_HYPONYM_OF for *dog*/*cão*.

Table 4 shows distinct types of dictionary entries with implicit semantic information, namely the support verb construction that can be synonymous to a verb entry (*impressionar* = *causar impressão* – *impress* = *make an impression*; f*icar azedo* = *azedar* – *turn sour* = *sour*), the semantic relation between an adjective and a semantically related adverb (*aesthetic* – *aesthetically*), and the semantic relation between a noun and a semantically related adverb (*skepticism* – *skeptically*). These relations are established by means of grammar rules. We have focused on the most regular rules, which are the ones that allow transformation of part-of-speech through the process of derivation.

In the examples illustrated in Table 4, the properties in bold correspond to the derivational rule and inflectional paradigm. Accordingly, DRV=NDRV01:CANÇÃO is a dictionary property that calls the rule to derive (through the process of nominalization) the predicate noun *impressão* (*impression*) from the verb *impressionar* (*impress*) and assigns it the inflectional paradigm CANÇÃO (the noun *impressão* inflects in the same way as the noun *canção*; i.e., following the same process and using the same morphemes to form the plural, etc.); DRV=ADRV00:ALTO is a dictionary property that calls the rule to derive the predicate adjective *azedo* (*sour*) from the verb *azedar* (*sour*) and assigns it the inflectional paradigm ALTO (the adjective *azedo* inflects like the adjective *alto*). DRV=AVDRV03 is a dictionary property that calls the rule to derive the adverb aesthetically from the adjective aesthetic; and, finally, DRV=NAVDRV02 is a dictionary property that calls the rule to derive the adverb skeptically from the noun skepticism. The lexical entries for the verbs *impressionar* (*impress*), *adaptar* (*adapt*), *azedar* (*sour*), have the property VSUP, that is, the description of the support verb that occurs with the predicate nouns *impressão* (*impression*), *adaptação* (*adapt*) and with the predicate adjective *azedo* (*sour*), which derive from the corresponding cited verbs. The combination of the description in the properties VSUP and DRV allows the semantic association between these verbs and their equivalent support verb constructions, namely *fazer*/*causar impressão* (*make*/*cause impression*), *fazer adaptação* (*make adaptation*), and *ficar azedo* (*turn sour*).

Table 5 shows the transformational rules to associate morpho-syntactic and semantically related words of different parts-of-speech, extracted individually from the Eng4NooJ and Port4NooJ rule databases. Rules are indexed according to different types of transformation. NDRV transforms verbs into nouns, ADRV transforms verbs into adjectives, and AVDRV transforms adjectives into nouns. The rules of each type are numbered. For example, NDRV04 is the rule number 04 that transforms a verb into a noun. The slash (/) after each ending in-

```
impressionar,V+FLX=FALAR+SAL=PVPCpleasetype+EN=impress+VSUP=fazer+VSUP=causar+DRV=NDRV01:CANCÃO
adaptar,V+FLX=FALAR+Aux=1+INOP57+Subset=132+EN=adapt+VSUP=fazer+DRV=NDRV00:CANCÃO
azedar,V+FLX=LIMPAR+Aux=1+OBJTRundif98+Subset=740+EN=sour+VSUP=ficar+DRV=ADRV00:ALTO
aesthetic,AFLX=NATURAL+SAL=AVstate+PT=estetico+DRV=AVDRV03
skepticism,N+FLX=BOOK+SAL=ABcause+PT=cepticismo+DRV=NAVDRV02
```

Table 4: General language dictionary entries with implicit semantic relations

troduces the part-of-speech of the derived word. The plus sign (+) introduces information about a specific noun or adjective. For example, Npred and Apred stand for predicate noun and predicate adjective, respectively. The capital letters between the less-than and the greater-than signs (<, >) correspond to commands. The command <B> means "backspace one character and add the string that follows the command, assigning it a new part-of-speech". The command <B2> means "delete the last two characters of the word from which the new word derives and add the string that follows the command", and so on and so forth. The strings that follow a command are the endings of the new generated words (e.g. *-ion* for the noun *acceleration*, *-tically* for the adverb *realistically*, etc.). The command <E> means that no character needs to be deleted. The command <A> means "delete the acute accent in the word from which the new word derives".

Eng4NooJ and Port4NooJ grammars are the devices used to recognize words or expressions and generate new ones, paraphrase or translate them. For example, the grammar in Figure 1, is used to recognize adverbial compounds in Portuguese and transform them into equivalent single adverbs. This grammar transforms multiword adverbs such as *de (um) modo rápido* (*in a fast/quick way*) into single adverbs such as *rapidamente* (*quickly*). This type of transformation is allowed by operations like the one represented in the first path of the graph. The box calls a new graph to recognize the strings *de (um) modo*, *de (uma) forma/maneira* (*in a (ADJ) way*), which make up the multiword adverbial. The output $A_ADV retrieves the adverb that is linked to the adjective $A. The adjective is transformed in the equivalent adverb by means of the derivational rules. The same grammar also recognizes multiword adverbs whose head is a noun, such as *por acidente* (*by accident*) or *com entusiasmo* (*with enthusiasm*), following the second and third paths.

The grammar in Figure 1 is monolingual,

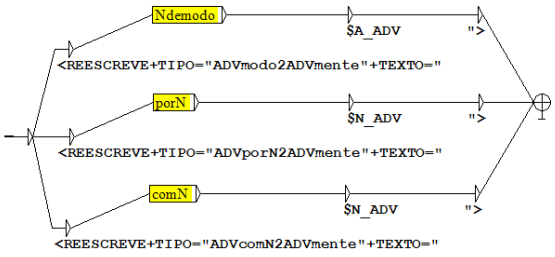
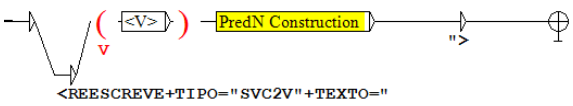


Figure 1: Grammar to recognize multiword adverbials in Portuguese and transform them into equivalent single adverbs

because there is no specification of the output for a different language. However, both Eng4NooJ and Port4NooJ resources contain Portuguese and English transfers for each lexical entry, i.e., they are in fact bilingual dictionaries. As a result, any grammar used to obtain monolingual transformations can be reused to generate bilingual (or multilingual) transformations. That is, the same grammar can be used to retrieve the output in English or in any other language (separately or together) as long as the words of that language are in the bilingual or multilingual dictionary and there are rules associated to the relevant dictionary properties. This means that, the grammar can generate translations from one to many languages, i.e., it can be used to create cross-language semantic relations. For monolingual transformations, no output language is specified. For bilingual or cross-language transformations, the parameter for the specification of the output language needs to be added. The parameter $EN for English, $IT for Italian, $SP for Spanish, etc. specifies the retrieval of the output in one of these languages or in all of them simultaneously. Similarly, the grammar presented in Figure 2, can be used for cross-language semantic relations. This grammar matches on a support verb construction of the type [Predicate Noun Construction] (*dar um abraço (a) – give a hug (to)*) (in the figure represented in a box that calls a sub-graph) and paraphrases it into a single verb (*abraçar – hug*).

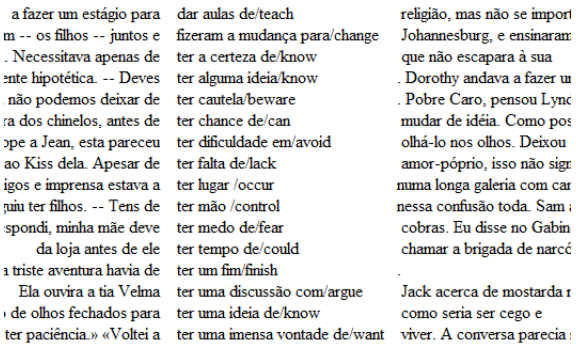
Eng4NooJ	Port4NooJ
NDRV04 = <B>ion/Npred	NDRV02 = <B>nca/N+Npred
e.g. *accelerate* > *acceleration*	e.g. *mudar* > *mudança*
ADRV02 = <B>icable/ADJ	ADRV02 = <B2>o/A+Apred
e.g. *apply* > *applicable*	e.g. *azedar* > *azedo*
AVDRV01 = <E>ly/ADV	AVDRV00 = <B>zmente/ADV
e.g. *frequent* > *frequently*	e.g. *veloz* > *velozmente*
AVDRV04 = <B>tically/ADV	AVDRV05 = <A> <B>amente/ADV
e.g. *realism* > *realistically*	e.g. *rápido* > *rapidamente*

Table 5: Rules to transform morpho-syntactic and semantically related words of different parts-of-speech



<REESCREVE+TIPO="SVC2V"+TEXTO="

Figure 2: Grammar to generate cross-language relations between Portuguese support verb constructions and equivalent English single verbs



a fazer um estágio para	dar aulas de/teach	religião, mas não se import
m -- os filhos -- juntos e	fizeram a mudança para/change	Johannesburg, e ensinaram
. Necessitava apenas de	ter a certeza de/know	que não escapara à sua
ente hipotética. -- Deves	ter alguma ideia/know	. Dorothy andava a fazer u
não podemos deixar de	ter cautela/beware	. Pobre Caro, pensou Lync
ra dos chinelos, antes de	ter chance de/can	mudar de idéia. Como pos
ope a Jean, esta pareceu	ter dificuldade em/avoid	olhá-lo nos olhos. Deixou
ao Kiss dela. Apesar de	ter falta de/lack	amor-próprio, isso não sign
igos e imprensa estava a	ter lugar /occur	numa longa galeria com car
guiu ter filhos. -- Tens de	ter mão /control	nessa confusão toda. Sam
spondi, minha mãe deve	ter medo de/fear	cobras. Eu disse no Gabin
da loja antes de ele	ter tempo de/could	chamar a brigada de narcó
a triste aventura havia de	ter um fim/finish	.
Ela ouvira a tia Velma	ter uma discussão com/argue	Jack acerca de mostarda r
de olhos fechados para	ter uma ideia de/know	como seria ser cego e
ter paciência.» «Voltei a	ter uma imensa vontade de/want	viver. A conversa parecia

Figure 3: Cross-language relations between Portuguese support verb constructions and equivalent English single verbs

Figure 3 illustrates the output of a grammar that generates cross-language semantic relations between Portuguese support verb constructions and English single verbs. At present, the semantic relations included in Eng4NooJ and Port4NooJ are mostly used to generate paraphrases and integrated in the paraphrasing tools SPIDER and eSPERTo. However, cross-language relations such as those illustrated in Figure 3 can be used directly in machine translation and are fuelling the ParaMT bilingual paraphrasing tool. At the current stage of development, ParaMT translates mostly multiword units, performing well in the translation of Portuguese support verb constructions into English verbs, and vice-versa, the linguistic phenomena most researched when applying the current methodology.

Relation	Quantity
Hyponymy	14,963
Synonymy	10,395
between nouns	5,367
between verbs	20
between adjectives	34
between adverbs	5,014
Action-of	3,773
Result-of	283

Table 6: Relations in Port4NooJ v.2.0

6 *Preliminary Results*

In theory, the exploitation of the lexicon in combination with SAL allows the establishment of numerous relations between words and expressions. For the current paper, we focused only on a few of those relations which cover a larger number of items and could be extracted in a simple and easy way. The result of extraction for Portuguese (not yet reviewed) is publicly available[6]. Currently, Port4NooJ contains more than 30,000 morpho-syntactic relations between semantically related elements. Table 6 presents some preliminary results, which do not refer to paraphrasing capabilities, but simply to relations between lexical items. The total results for paraphrasing are significantly higher. Local grammars, applied to information (properties) described in the dictionary, enable the recognition and analysis of expressions such as *de (um) modo rápido*, *de (uma) forma/maneira rápida* (*in a fast/quick way*) (which could be considered as relations between an adjective and an adverb, but which were not counted), and also inflected forms such as *dar uns passeios* (*go for some walks*), etc.

Port4NooJ contains approximately 600 derivational rules, most of them transforming verbs into predicate nouns (587). 119 of

[6]See http://www.linguateca.pt/Repositorio/Port4NooJ/relacoes_semanticas_explicitas/

these rules are productive, covering nominalizations. 486 rules correspond to verb relations between verbs and autonomous predicate nouns. At this point in the research, rules were only superficially evaluated.

7 Conclusions and Future Research

This paper presented semantic relations, namely domain-independent semantico-syntactic and ontological relations, suitable for paraphrasing and cross-language tasks, including machine translation. We have demonstrated that given the appropriate linguistic resources, the generation of semantic relations can become very systematic. Any grammar to generate monolingual semantic relations can be reused to generate cross-language relations, rules can be standardized and often re-used across close languages, etc. Even though the methodology adopted was applied to the OpenLogos resources, it is compliant with the exploitation of other lexical resources with semantic relations, for any language besides English and Portuguese, studied in this research.

Future work would gather and combine open source available semantic resources, enhance properties on the existing resources, and enlarge the linguistic phenomena coverage.

References

Barreiro, Anabela. 2007. Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation. In Xavier Blanco, Max Silberztein, Xavier Blanco, and Max Silberztein, editors, *Proceedings of the 2007 International NooJ Conference*, pages 19–47. Cambridge Scholars Publishing, June 7-9.

Barreiro, Anabela. 2008a. *Make it simple with paraphrases. Automated paraphrasing for authoring aids and machine translation.* Ph.D. thesis, Universidade do Porto, Portugal.

Barreiro, Anabela. 2008b. ParaMT: A paraphraser for machine translation. In *Proceedings of Computational Processing of the Portuguese Language, 8th International Conference (PROPOR 2008)*, volume 5190 of *LNCS*, pages 202–211, Aveiro, Portugal. Springer.

Barreiro, Anabela. 2011. SPIDER: a System for Paraphrasing In Document Editing and Revision - applicability in machine translation pre-editing. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11, pages 365–376, Berlin, Heidelberg. Springer.

Barreiro, Anabela and Luís Miguel Cabral. 2009. ReEscreve: a translator-friendly multi-purpose paraphrasing software tool. In Marie-Josée Goulet, Christiane Melançon, Alain Désilets, and Elliott Macklovitch, editors, *Proceedings of the Workshop Beyond Translation Memories: New Tools for Translators, The Twelfth Machine Translation Summit*, pages 1–8.

Barreiro, Anabela, Bernard Scott, Walter Kasper, and Bernd Kiefer. 2011. Openlogos machine translation: philosophy, model, resources and customization. *Machine Translation*, 25(2):107–126.

Berland, M. and E. Charniak. 1999. Finding parts in very large corpora. In *Proceedoings of 37th annual meeting of the ACL on Computational Linguistics*, pages 57–64, Morristown, NJ, USA. Association for Computational Linguistics.

Chodorow, Martin S., Roy J. Byrd, and George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of 23rd annual meeting on Association for Computational Linguistics*, pages 299–304, Morristown, NJ, USA. ACL Press.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication).* The MIT Press.

Freitas, Cláudia and Violeta Quental. 2007. Subsídios para a elaboração automática de taxonomias. In *XXVII Congresso da SBC - V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 1585–1594.

Girju, Roxana and Dan Moldovan. 2002. Text mining for causal relations. In Susan M. Haller and Gene Simmons, editors, *Proc. 15th Intl. Florida Artificial*

Intelligence Research Society Conference (FLAIRS), pages 360–364.

Gonçalo Oliveira, Hugo and Paulo Gomes. 2010. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*. IOS Press.

Gonçalo Oliveira, Hugo, Diana Santos, and Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93, May.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. 14th Conf. on Computational Linguistics*, pages 539–545, Morristown, NJ, USA. ACL Press.

Lin, Dekang and Patrick Pantel. 2002. Concept discovery from text. In *Proceedings of 19th International Conference on Computational Linguistics (COLING)*, pages 577–583.

Maia, Belinda and Sérgio Matos. 2008. Corpógrafo v4: tools for researchers and teacher using comparable corpora. In *Proceedings of LREC 2008 Workshop on Comparable Corpora*, pages 79–82, Marrakech, Morocco. ELRA.

Maia, Belinda and Luís Sarmento. 2005. The corpógrafo - an experiment in designing a research and study environment for comparable corpora compilation and terminology extraction. In *Proceedings of eCoLoRe / MeLLANGE Workshop, Resources and Tools for e-Learning in Translation and Localisation*, pages 45–48, Leeds University, UK, March 21-23. Center for Translation Studies.

Marrafa, Palmira. 2002. Portuguese Wordnet: general architecture and internal semantic relations. *DELTA*, 18:131–146.

Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo, and Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 390–392.

Richardson, Stephen D., William B. Dolan, and Lucy Vanderwende. 1998. Mindnet: Acquiring and structuring semantic information from text. In *Proceedings 17th International Conference on Computational Linguistics (COLING)*, pages 1098–1102.

Roget, P. M. 1852. *Roget's Thesaurus of English words and phrases.* Available from Project Gutemberg, Illinois Benedectine College, Lisle IL (USA).

Santos, Diana, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes, and Rosário Silva. 2010. Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. In A. M. Brito, F. Silva, J. Veloso, and A. Fiéis, editors, *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística.* APL, pages 681–700.

Sarmento, Luís, Belinda Maia, Diana Santos, Ana Pinto, and Luís Cabral. 2006. Corpógrafo v3: From terminological aid to semi-automatic knowledge engine. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC 2006, pages 1502–1505. ELRA.

Scott, Bernard. 2003. The logos model: An historical perspective. *Machine Translation*, 18:1–72, March.

Silberztein, Max. 2007. An alternative approach to tagging. In *Proceedings of Natural Language Processing and Information Systems, 12th International Conference on Applications of Natural Language to Information Systems (NLDB 2007)*, volume 4592 of *LNCS*, pages 1–11, Paris, France, June 27-29. Springer.

Sowa, John. 1999. *Knowledge Representation: Logical, Philosophical and Computational Foundations.* Thomson Learning, New York, NY, USA.

Generación semiautomática de recursos de Opinion Mining para el gallego a partir del portugués y el español

Semiautomatic generation of Opinion Mining resources for Galician from Portuguese and Spanish resources

Paulo Malvar Fernández
Departamento de Ingeniería Lingüística,
imaxin|**software**
Salgueiriños de abaixo nº11 L6, 15891,
Santiago de Compostela
A Coruña
paulomalvar@imaxin.com

José Ramom Pichel Campos
Departamento de Ingeniería Lingüística,
imaxin|**software**
Salgueiriños de abaixo nº11 L6, 15891,
Santiago de Compostela
A Coruña
jramompichel@imaxin.com

Resumen: A pesar del crecimiento experimentado en los últimos años en el ámbito del Procesamiento del Lenguaje Natural (PLN), investigadores y desarrolladores que pretenden llevar a cabo desarrollos para lenguas diferentes del inglés aún se encuentran con el problema de que los recursos y aplicaciones necesarios son escasos, cuando no inexistentes. En este trabajo proponemos una metodología semiautomática para generar recursos para una aplicación de Opinion Mining para el gallego aprovechando recursos del español y utilizando el portugués como lengua-puente que, por su proximidad, asegura una alta tasa de transferencia léxica con relación al gallego.
Palabras clave: Opinion Mining, Generación Semiautomática, Recursos, Español, Gallego, Portugués

Abstract: In spite of the growth experienced in recent years in the field of Natural Language Processing (NLP), researchers and developers who intend to carry out developments for languages other than English still have to face the old problem that needed resources and applications are scarce, if not nonexistent. In this paper we propose a semiautomatic method to generate resources for an Opinion Mining application for Galician. For this we drew from Spanish resources and used Portuguese as a bridge-language that, due to its linguistic proximity, ensures a high lexical transfer rate with Galician.
Keywords: Opinion Mining, Semiautomatic Generation, Resources, Spanish, Galician, Portuguese

1 Introducción

El crecimiento experimentado en los últimos años en el ámbito del Procesamiento del Lenguaje Natural (PLN), no sólo desde el punto de vista de investigación académica, sino también desde el punto de vista de desarrollo de aplicaciones y soluciones comerciales, se apoya en el trabajo realizado durante los últimos 60 años, desde que se comenzaron a desarrollar los primeros traductores automáticos en el contexto de la Guerra Fría.

Es precisamente por esta investigación y desarrollo previo ya realizado durante décadas que hoy en día es posible contar con numerosos y diversos corpora, así como innumerables herramientas.

Sin embargo, el problema con el que se encuentran investigadores y desarrolladores que pretenden llevar a cabo aplicaciones de PLN para lenguas diferentes del inglés es que este tipo de recursos y aplicaciones son escasos, cuando no inexistentes. Así por ejemplo, si dentro del ámbito del Opinion Mining en inglés es posible contar con corpora anotados con

información acerca de la orientación de las opiniones, en español, portugués o gallego este tipo de recursos es prácticamente inexistente.

Frente a esta escasez de recursos, ideamos una solución para aprovechar la relación de proximidad entre gallego con el español y de la especialmente próxima relación entre gallego y portugués, para semiautomáticamente generar los recursos necesarios para una aplicación de Opinion Mining basada en Machine Learning.

2 *Recursos disponibles*

Demostración empírica de la abismal distancia que existe en términos de investigación y desarrollo de soluciones de Opinion Mining entre el inglés y otras lenguas, como el español y el portugués, es la amplia diferencia en número de recursos disponibles. Así, para inglés existen actualmente numerosos vocabularios y corpora disponibles para descarga desarrollados para esta lengua[1], que han sido generados a partir de inúmeras investigaciones como (Blitzer, J. et al, 2007), (Ding, X. et al, 2008) y (Pang, B. et all, 2002).

Por el contrario, para el español sólo tenemos constancia de un único corpus, "Spanish Movie Reviews"[2], generado dentro de la investigación desarrollada en (Cruz, F. et al, 2008), y para gallego y portugués nos resultó imposible localizar ningún corpus y/o vocabulario precompilado.

Para hacer frente a la ausencia total de recursos para el gallego, en imaxin|software ideamos una solución para la generación rápida de recursos aprovechando la especialmente estrecha relación entre el gallego y el portugués así como la relación de proximidad entre gallego y español.

3 *Metodología propuesta*

Para el desarrollo de una aplicación de Opinion Mining basada en Machine Learning para el gallego necesitábamos un corpus etiquetado con información acerca de la orientación de las opiniones y un vocabulario controlado en el cual también se incluyese información de este tipo acerca de los adjetivos, sustantivos, verbos y adverbios en él contenidos.

Para el desarrollo de la aplicación análoga para español contamos con:

a) El corpus "Spanish Movie Reviews", compuesto de un total de 3875 críticas de cine anotadas con la puntuación que sus autores asignaron la película de la cual versa cada crítica. Del total de 3875 documentos, 351 tienen asociada una estrella de puntuación, 923 dos estrellas, 1253 tres estrellas, 887 cuatro estrellas y 461 cinco estrellas.

b) Un vocabulario controlado, derivado mediante la aplicación del algoritmo explicado en (Turney, P.D., 2002) y completado por traducción automática de las formas contenidas en el General Inquirer[3].

Para la generación de recursos análogos para el gallego se aplicó el siguiente flujo trabajo:

1- Traducción de español a gallego del corpus "Spanish Movie Reviews" y del vocabulario controlado de español. Para este paso se optó por el sistema de traducción Opentrad, en cuyo desarrollo imaxin|**software** ha colaborado, (Loinaz, I. et at, 2006).

2- Traducción de español a portugués de las palabras desconocidas por el par es-gl de Opentrad. Para esta tarea se optó por Google Translate[4] que en nuestra opinión, subjetiva a todos los efectos, tiene, para este par de lenguas (es-pt), mayor cobertura léxica que Opentrad aunque a costa de una mucho menor corrección gramatical.

3- La lista de palabras traducida a portugués obtenida en el paso anterior fue, en un tercer paso, traducida al gallego utilizando Opentrad pt-gl.

[1] Dentro del ámbito de un proyecto llamado "Web Mining, Text Mining, and Sentiment Analysis", Bing Liu ha puesto a disposición de la comunidad un corpus de críticas de usuarios de tiendas on-line que puede ser descargado desde http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. John Blitzer también ha puesto a disposición de la comunidad un corpus llamado "Multi-Domain Sentiment Dataset" que puede ser descargado desde http://www.cs.jhu.edu/~mdredze/datasets/sentiment/. Finalmente, mencionar también la contribución de Bo Pang y Lillian Lee, que han puesto a disposición de la comunidad un corpus de críticas de cine que puede ser descargado desde http://www.cs.cornell.edu/people/pabo/movie-review-data/.

[2] Este corpus puede ser descargado desde http://www.lsi.us.es/~fermin/corpusCine.zip

[3] Dentro del ámbito de los vocabularios o lexicones etiquetados con información sobre la orientación sentimental, el más famoso y utilizado es General Inquirer, que puede ser descargado desde http://www.wjh.harvard.edu/~inquirer/.

[4] http://translate.google.com/#es|pt|

4- En un cuarto paso se detectaron las palabras desconocidas para el par Opentrad pt-gl, las cuales se transliteraron de portugués a gallego utilizando un script de transliteración llamado port2gal[5]. El hecho de que portugués y gallego pertenezcan, tal y como afirman (Coseriu, E., 1987) y (Cunha, C. y Cintra, L., 2002), a un mismo conjunto dialectal gallego-portugués, asegura una alta tasa transferencia léxica entre ambas variantes apenas modificando su forma superficial, esto es su ortografía, tal y como demuestra (Malvar, P. Et al, 2010).

5- Para la depuración de errores contenidos en la lista final de palabras obtenidos tras los sucesivos pasos explicados, se procedió a una corrección manual de dicha lista que finalmente se utilizó para corregir el corpus generado en el primer paso.

Mediante este flujo de trabajo finalmente se obtuvo, en primer lugar, un corpus de críticas de cine en gallego compuesto, al igual que en el caso del español de 3875 documentos clasificados según el ranking de estrellas asociadas por los usuarios responsables de dichas críticas. En segundo lugar se obtuvo un vocabulario controlado compuesto de un total de 5448 palabras, de las cuales 2293 fueron clasificadas como positivas y 3155 palabras clasificadas como negativas.

4 *Configuración del algoritmo*

El tipo de estrategia que se adoptó para el desarrollo de este proyecto estuvo condicionada por fuertes restricciones en relación a los recursos que imaxin|**software**, como PYME, podía invertir.

Además, como es bien sabido, las soluciones basadas en *Machine Learning* ofrecen resultados aceptables en un muy corto espacio de tiempo. Por lo tanto, se optó por una estrategia basada en *Machine Learning*, e, inspirados en los resultados obtenidos en (Pang, B. et al, 2002), se escogió Support Vector Machines (SVM) como algoritmo a utilizar para el entrenamiento de un módulo de *Opinion Mining*.

Para la implementación del módulo de SVM se utilizó la versión 2.90 de libSVM, (Fan, R.E. et al, 2005), en cuya configuración estándar sólo se modificó el tipo de kernel, pasando del estándar RBF kernel a un POLYNOMIAL kernel.

Para la conversión de los textos en vectores de clasificación se utilizaron las siguientes *features*:

1- La presencia de palabras en los textos de entrenamiento que estuviesen contenidas en nuestro vocabulario controlado de términos positivos, codificados con valor 1, y negativos, codificados con valor -1.

2- En cuanto al resto de palabras no contenidas en las listas de términos positivos o negativos, se optó por la codificación con valor 2 para aquellas palabras del conjunto del corpus presentes también en un determinado texto; y la codificación con valor 3 para aquellas palabras del conjunto del corpus no presentes en un determinado texto.

3- Por último, en los vectores de clasificación se incluyeron dos coordenadas adicionales: el total de palabras positivas y el total de palabras negativas detectadas.

5 *Resultados*

Dado el muy reciente auge del *Opinion Mining* como rama de investigación dentro del PLN, hoy en día aún no existe ni para español ni para gallego ningún *gold standard* con el cual comparar nuestro sistema de clasificación de sentimientos para determinar su rendimiento. Por esta razón, optamos por crear nosotros mismos un pequeño corpus de pruebas que construimos extrayendo al azar textos clasificados como críticas positivas o negativas por los usuarios de diversos sitios web. Los sitios web de los cuales se extrajeron los textos fueron: Google Maps[6], booking.com y la tienda de aplicaciones App Store[7] de Apple. Los dominios a los que pertenecen los textos extraídos son los siguiente: 10 textos (5 positivos y 5 negativos) de críticas de hoteles de Santiago de Compostela y Madrid, 10 textos (5 positivos y 5 negativos) de críticas de restaurantes de Santiago de Compostela; y 10 textos (5 positivos y 5 negativos) de críticas de

[5] port2gal es un simple script de Perl que fue inicialmente desarrollado por Alberto García (de la empresa Igalia) y que posteriormente fue mejorado por Pablo Gamallo (Departamento de Lengua Española de la Universidad de Santiago de Compostela). Este script simplemente convierte la ortografía do portugués europeo a la ortografía actual del gallego. port2gal está disponible bajo GPL en http://gramatica.usc.es/~gamallo/port2gal.htm.

[6] http://maps.google.com/

[7] http://itunes.apple.com/es/

aplicaciones disponibles en la App Store de Apple.

Resulta evidente la disparidad entre estos dominios y el dominio de la crítica cinematográfica, al que pertenecen los textos de entrenamiento del clasificador. Sin embargo, en **imaxin**|software queremos aplicar estos modelos de clasificación de opiniones a ámbitos que no se encuentran dentro del dominio de la crítica cinematográfica. Por lo tanto, pensamos que los resultados obtenidos para los textos de evaluación, sin ser en modo alguno concluyentes, podrían ser un indicador de la aplicabilidad a diversos dominios de los modelos de clasificación aprendidos.

Los textos escogidos estaban escritos en español y fueron traducidos a gallego manualmente por los autores de este trabajo. De esta manera, nos es posible realizar una comparativa directa entre los resultados en español y gallego, pues se trata de los mismos textos simplemente escritos en una u otra lengua.

En la tabla 1 se presentan los resultados obtenidos por el motor de clasificación para español. Y en la tabla 2 se presentan los resultados obtenidos por el motor de clasificación para gallego.

	Precisión	Cobertura
Positivos	0.79	0.73
Negativos	0.75	0.80
Global	0.77	0.77

Tabla 1: Resultados del clasificador SVM para español

	Precisión	Cobertura
Positivos	0.72	0.87
Negativos	0.83	0.67
Global	0.78	0.77

Tabla 2: Resultados del clasificador SVM para gallego

5.1 Discusión de los resultados

Como se puede apreciar en las tablas 1 y 2 los resultados son muy similares para gallego y español. La diferencia más significativa entre ambos es la mayor tendencia que tiene el motor de gallego para clasificar como positivos los textos, como sugiere su cobertura del 87% y su precisión del 72%), y la mayor tendencia del motor de español para clasificar los textos como negativos, como se aprecia por su cobertura del 80% y su precisión del 75%.

En cualquier caso, la clasificación de textos positivos y negativos no baja de una precisión del 70% y la cobertura sólo en el caso de los textos negativos para gallego se encuentra ligeramente por debajo del 67%.

Sin embargo, es necesario tener en cuenta que los textos que han servido para el entrenamiento de los clasificadores tanto para gallego como para español pertenecen al dominio de la crítica cinematográfica informal, el cual es muy diferente de los dominios representados en los textos de evaluación (que recordemos pertenecen al dominio hotelero, hostelero y tecnológico). Este es un factor que, a buen seguro, juega en contra de la precisión de ambos clasificadores. Aún así, como demuestran los resultados globales, que se encuentran tanto para la precisión como para la cobertura ligeramente por debajo del 80%, el desempeño global de ambos motores de clasificación es, en nuestra opinión, muy satisfactorio.

Por otro lado, y en concreto para el clasificador de gallego, existe otro factor que, en nuestra opinión, es responsable de cierta degradación de los resultados. Este factor es la naturaleza del gallego contenido en los textos que han servido como corpus de entrenamiento. Así, si bien para el español los textos fueron originalmente escritos en esta lengua, en el caso del gallego los textos han sido obtenidos de manera artificial, esto es, mediante un proceso semiautomático de traducción y transliteración. Por lo tanto, podríamos afirmar que mientras para el español contamos con textos naturales, para el gallego contamos con textos escritos en "pseudo-lengua". De cualquier manera, y a la luz de los resultados obtenidos, el clasificador de gallego tiene un desempeño comparable al clasificador de español.

6 Conclusiones

En este artículo hemos mostrado una metodología de conversión a gallego de fuentes de recursos disponibles en español y portugués necesarios para el entrenamiento de un motor SVM de clasificación de opiniones.

La metodología propuesta combina la traducción automática de español a gallego y de portugués a gallego, la expansión de

vocabularios mediante tesauros y la transliteración de palabras de portugués a gallego.

Los resultados obtenidos, que rondan el 80% de cobertura y precisión, son comparables a los de herramientas similares disponibles en otras lenguas.

Sin duda, queda demostrado que la metodología propuesta para la obtención de recursos para gallego ha sido un éxito. En nuestra opinión, esta metodología es perfectamente extrapolable a otras lenguas que guardan lazos especialmente estrechos con variedades lingüísticas desarrolladas en términos de recursos de Procesamiento del Lenguaje Natural.

Bibliografía

Blitzer, J., Drezde, M., Pereira, F.. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Annual Meeting-Association For Computational Linguistics, vol. 45 (1), pp. 440--448 (2007)

Coseriu, E.: El gallego en la historia y en la actualidad. In Actas do II Congresso Internacional da Língua Galego-Portuguesa, pp. 793-800 (1987)

Cunha, C., Cintra, L.: Nova Gramática do Português Comtemporâneo. Edições João Sá da Costa, Lisboa (2002)

Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In Proceedings of the international conference on Web search and web data mining, pp. 231--240 (2008)

Fan, R.E., Chen, P.H., Lin, C.J.: Working set selection using second order information for training SVM. Journal of Machine Learning Research, vol 6, pp. 1889--1918 (2005)

L. Cruz, F., Troyano, J.A., Enríquez, F., Ortega, J.: Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. In Procesamiento del Lenguaje Natural,vol. 41, pp. 73--80 (2008)

Loinaz, I., Aranrtzabal, I., Forcada, M.L., Gómez Guinovart, X., Padró, Ll., Pichel Campos, J.R., Waliño, J.: OpenTrad: Traducción automática de código abierto para las lenguas del Estado español. Procesamiento del Lenguaje Natural, vol. 27, pp 357--360 (2006)

Malvar, P., Pichel Campos, J.R., Senra, Ó., Gamallo, P., García, A.: Vencendo a escassez de recursos computacionais. Carvalho: Tradutor Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português. In Linguamática, vol 2, n. 2, pp. 31--38 (2010)

Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, pp. 79--86 (2002)

Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417--424 (2002)

Language dominance prediction in Spanish-English bilingual children using syntactic information: a first approximation*

Predicción de lenguaje dominante en niños bilingües en Español e Inglés usando información sintáctica: una primera aproximación

Gabriela Ramirez-de-la-Rosa
Thamar Solorio
University of Alabama at Birmingham
gabyrr,solorio@cis.uab.edu

Manuel Montes-y-Gómez
University of Alabama at Birmingham
INAOE México,
mmontesg@inaoep.mx

Yang Liu
The University of Texas at Dallas
yangl@hlt.utdallas.edu

Aquiles Iglesias
Temple University
iglesias@temple.edu

Lisa Bedore
The University of Texas at Austin
lbedore@mail.utexas.edu

Elizabeth Peña
The University of Texas at Austin
lizp@mail.utexas.edu

Resumen: Este artículo presenta los resultados de un estudio preliminar donde se usa información sintáctica para predecir el lenguaje dominante en niños bilingues en Español e Inglés. Nuestro enfoque usa una bolsa de reglas gramaticales sintácticas extraídas de narraciones en Inglés y Español. Medimos la exactitud de la predicción de categorizar niños dentro de 3 clases: Español-dominante, Inglés-dominante y Bilingüe balanceado. Los resultados son competitvos con trabajos previos que utilizan un conjunto de características mucho más grande y diverso. Este artículo presenta los beneficios potenciales de agregar un análisis sintáctico más profundo para modelar el lenguaje de niños, incluso en el caso de tener muestras con mezcla de idiomas.
Palabras clave: Lenguaje dominante, Reglas gramaticales sintácticas.

Abstract: This paper presents results on a preliminary study using syntactic information to predict language dominance in Spanish-English bilingual children. Our approach uses a bag of syntactic grammar rules taken from narratives in English and Spanish. We then measure prediction accuracy of categorizing children into Spanish-dominant, English-dominant, and Balanced Bilingual. The results are competitive to previous work using a much larger and diverse set of features with shallow syntactic analysis. This paper shows the potential benefit of adding a deeper syntactic analysis for modeling language in young children, even in the case of having mixed language samples.
Keywords: Language dominance, Syntactic grammar rules.

1. Introduction

In the field of communication disorders, the analysis of spontaneous language samples is a common practice to determine language status of children. Typically, this involves a very expensive process of manually coding and analyzing these samples to find patterns that are known to be good clinical markers. For the analysis of language from monolingual children, especially English-speaking children, there is a vast amount and breath of research that supports the use of these clinical markers. However, for bilingual popu-

* This research was partially supported by the National Science Foundation under grants 1018124 and 1017190, and by NIH NIDCD R01 grant DC007439. This work was also supported in part by the UPV, award 1932, under the program Research Visits for Renowned Scientists (PAID-02-11) and by the European Commission as part of the WIQ-EI project (project no. 269180) within the FP7 People Programme.

lations the literature is not as extensive, although it is steadily growing. One task considered critical by clinical researchers when analyzing language from bilingual children is identification of language dominance. That is, in order to make final recommendations or diagnosis, it has been found to be critical to know which language, if any of the two, is more developed in the child. Recent research in communication disorders presents two approaches for determining language dominance in bilingual children, one based on measures of language exposure (Bedore et al., 2010) and the other one based on measures of language productivity (Paradis et al., 2003), although the former seems to be more widely accepted. However, determination of language required ask to parents and teachers the amount of input and output of children over a period of time, typically a week; since the children are not monitored 100 % of the time.

Previous work by Solorio et al. (2011) from the Natural Language Processing (NLP) community has looked at a corpus driven approach for this problem of determining language dominance. They framed this problem as a text classification task, where the classes are the three potential language dominance categories: English dominant (ED), Spanish dominant (SD), and balanced bilingual (BB), and they extracted a large variety of features from the language samples to train a machine learning classifier. In this paper we follow the idea of using a machine learning algorithm, but the set of features we explore here are purely syntactic, and were not explored in the work mentioned above. Our results show that deeper syntactic information carries rich relevant content for the task of determining the language dominance of Spanish-English bilingual children. We extract features from the parse trees generated by off-the-shelf syntactic parsers for English and Spanish. Then we train a learning algorithm using the set of syntactic rules found in each transcript as features. We call this a bag of rules (BOR) approach. The accuracy results obtained by our simple syntactic based features are higher than several of the features presented in previous work. We speculate that combining this information with that in Solorio et al.'s paper can lead to even higher accuracies.

2. Related Work

Previous work has used NLP techniques to help in the areas of communication disorders. In Gabani et al. (2009), in order to predict language impairment in monolingual English and Spanish-English bilingual children, they used six sets of features to build a computational model: language productivity, morphosyntactic skills, vocabulary knowledge, speech fluency, perplexities from LMs and standard scores. In this previous work the best result reported was around 60 % of F-measure. In a more recent work, an addition of 3 sets of features to previous features was proposed. In particular, demographic information, syntactic complexity, and POS n-grams, were included to predict the dominant language in bilingual children (Solorio et al., 2011). This more recent work added some syntactic information as features but only at the level of part of speech tags. The best result obtained in this work was 72 % of accuracy.

On the other hand, NLP techniques have also been explored in the detection of mild cognitive impairment (Roark, Mitchell, and Hollingshead, 2007), where features such as Yngve and Frazier scores, together with features derived from automated parse trees are explored in that work to model syntactic complexity. Similar features are used in the classification of language samples as belonging to children with autism, language impairment, or none of the above (Prud'hommeaux et al., 2011).

The last two approaches inspired us to explore the use of information generated by automatically parsing the language samples. The features, as they are proposed here, have not been used in previous work. In this sense, the novelty of our study is the use of a representation analogous to bag of words that used syntactic patterns as extracted from parse trees. The next section describes our proposed method in more detail.

3. Proposed Approach

The goal of the task is the prediction of language dominance of a child into one of three core categories: BB (balanced bilingual), ED (English dominant), and SD (Spanish dominant). Since we want to streamline the process of language analysis as much as possible, we restrict the feature set to features that can be automatically extracted from the trans-

cripts. Moreover, since previous work for automated language dominance prediction has not explored the use of parse trees, or features derived from parse trees, we study in this work their contribution to developing an accurate model for this task. We expect that children at similar stages of language acquisition will have mastered a similar set of grammatical constructions and that this can be exploited by a learning algorithm. An interesting twist in this classification task is the fact of having information, language samples, in each of the two languages. While it is widely accepted that in a bilingual population is important to assess language ability on both languages, it is less clear how to do this in a machine learning scenario. Here, we explore different ways to combine the observed samples in both languages.

The idea of this study is very simple. It consists of the following steps:

1. **Automatically parsing the transcripts.** In this step we generate a set of parse trees for each transcript using trained monolingual parsers. Because we lack gold standard parse trees of bilingual child language, we are assuming that a parser trained on mostly adult language will not have a major negative effect in our proposed solution. However, it should be noted here that the noise from the parse trees is not only coming from the differences between adult language constructs and those from children, but also from the mixed language input. As explained in the following section, children are prompted to elicit the language samples in one target language, but frequently these children code switched between their two languages. Our assumption is that the parser will make consistent decisions when unexpected tokens appear during analysis, and thus the noise from those elements will by systematically added to both, training and testing data and this will not have a major effect on classification accuracy into language dominance. But we do recognize that if careful analysis will be performed on the parse trees, then adaptation of the parsers, to both child language, and mixed language input, might be needed.

2. **Finding rules**. Using every parse tree for a transcript, we find each rule of the form of $\alpha \rightarrow \beta$, where α is the root of a subtree and β is the set of children in that particular subtree. Because we are more interested in grammatical structure than in the actual vocabulary, we only add to the list those rules not involving a lexicon entry.

3. **Creating the representation of transcripts**. Once we gather the lexicon of grammar rules fired in the training set, we used them as features to represent each transcript. This representation is analogous to BOW (bag of words), but instead of words we have rules, thus we refer to this representation as BOR (bag of rules). We also use standard Boolean weights for the rules. The intuition is that it is enough to observe a syntactic construct once to assume the child masters that construction.

4. **Training a model for language dominance prediction.** Each transcript in the training set is transformed into a BOR vector. Then we use a standard machine learning algorithm to train a model. We assume then, that this problem of language dominance prediction can be cast as a classification problem.

5. **Classifying a child**. To classify the language dominance of a new child, we transform the transcript to a vector of n dimensions, where n is the number of elements in the BOR, and the value of each dimension is either presence (1) or absence (0) of the specific rule. Then we can use the trained model generated in the previous step to make a prediction for the new sample.

In the following section we describe the data set used to evaluate our proposed representation.

4. Data

The data set used in this paper contains transcripts gathered as part of an on-going longitudinal study of language impairment in bilingual Spanish-English speaking children (Peña et al., 2006). The children in this study were enrolled in kindergarten with a mean age of about 6 years and 1 month. A total of 180 children participated in this study, however, we only worked with 52 bilingual children

since the data for the rest of the children was not available for analysis at this point. Table 1 shows the distribution of our data.

Category	Children
Balanced Bilingual (BB)	19
English Dominant (ED)	11
Spanish Dominant (SD)	22

Table 1: Distribution of our dataset into the three categories

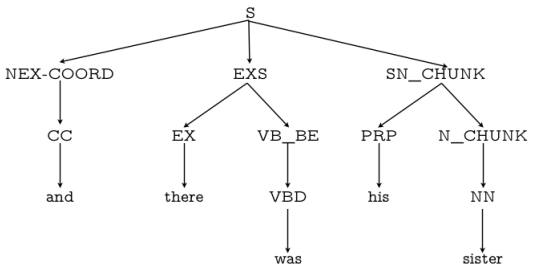
The transcripts were gathered following standard procedures for collection of spontaneous language samples in the field of communication disorders. For each child in the sample, four transcripts of story narratives were collected, two in each language. Children are shown a wordless picture book and are asked to narrate the story behind the book. The story narratives are based on Mayer's wordless picture books. The books used for English were *A boy, A dog, and a frog* (Mayer, 1967) and *Frog, where are you?* (Mayer, 1969b). The books used for Spanish were *Frog on his own* (Mayer, 1973) and *Frog goes to dinner* (Mayer, 1969a).

5. Experimental Setting

For extracting the parse trees we used FreeLing[1]. This parser comes with trained models for English and Spanish. The output of FreeLing is a set of parse trees. We break down the parse trees into grammar rules by traversing each tree in a breath first fashion. We only add rules to the BOR vector that are composed of a root and its immediate children. In Table 2 we show an example of a parse tree generate by FreeLing and the rules we extracted from it. Once we have the BORs we use them as features to represent the test transcripts. The value assigned to each rule in the vector is a boolean weight, $w_{i,j}$, one if the rule i appears in the transcript j, and zero otherwise.

As we mentioned in the previous section, we have 4 transcripts per child, but since our data set is small and we are using a corpus driven approach, we decided to duplicate the number of instances by separating the 4 sets of transcripts per child into 2 pairs. We realize that we are reducing by half how much

[1]FreeLing is available in the website: http://nlp.lsi.upc.edu/freeling



S → NEX-COORD EXS SN-CHUNK
NEX-COORD → CC
EXS → EX VB-BE
VB-BE → VBD
SN-CHUNK → PRP N-CHUNK
N-CHUNK → NN

Table 2: Parse tree generated by FreeLing for the sentence *and there was his sister* in one of the transcripts from our dataset and the rules we extracted from it

information we observe per child to train our model and to test prediction accuracy. However in this case we believe it is more important to have more data samples to both train and evaluate. Moreover, clinicians and clinical researchers use one transcript per language for the most part, so this is also aligned with current practices. Despite this separation of transcripts per story, we were careful to put in the same partition (training or test) all transcripts of the same child. That way we avoid confounding the ultimate goal of the task.

To decide the language dominance of a particular child or instance we consider 2 transcripts, thus $I = \{T_1 \cup T_2\}$. Because we have 4 transcripts per child, we consider the following options for combining the transcripts:

- One in English and one in Spanish
- Both in the same language (English or Spanish)

These two combinations are selected to answer one question: what is more helpful for analyzing language ability in bilingual children, using information from two languages, or more input in a single language? We already know the answer to this question from the point of view of communication disorders, and we speculate that in this case as well the most beneficial scenario will be when using

information from both languages. But it is interesting to explore if this pattern will hold when using a machine learning algorithm to predict language dominance.

To evaluate the performance of our method we used 5x2 cross fold validation, following recommendations in (Dietterich, 1998) for small sample sets. This means, we did 5 replications of 2-fold cross validation, in each repetition the available data was randomly partitioned into two equal-sized sets. In all our experiments we used the Weka (Witten and Frank, 1999) implementation of the machine learning algorithms.

6. Experimental Results

In our first experiment we wanted to determine whether by taking into account language samples only in one language is possible learn to distinguish between the three categories. However, to provide a fair comparison to that of using samples from each language, we took the two samples in the same language from each child. Thus we have two scenarios in this experiment: English-English and Spanish-Spanish. Table 3 shows the accuracy using five of the most common classification methods used in NLP problems: Naive Bayes, Support Vector Machines, C4.5, and k-Nearest Neighbors with $k = 1$ and $k = 5$.

	NB	**SVM**	**C4.5**	**1-NN**	**5-NN**
Eng.	45.9	**49.62**	43.7	45.2	45.9
Spa.	**58.5**	55.6	48.1	44.4	45.9

Table 3: Accuracy of BOR representation over 5 classification methods: Naive Bayes (NB), Support Vector Machine (SVM), Decision tree C4.5 and k-Nearest Neighbors with $k = 1$ (1-NN) and $k = 5$ (5-NN) using transcripts in one language: English (Eng.) or Spanish (Spa.)

The results shown are rather poor, but are comparable to results reported in (Solorio et al., 2011) on the same data set when using individual sets of features even though they are using information on both languages. Their reported accuracy ranges from 40 %, when using only demographic information, to 72 %, when using different metrics of syntactic complexity. However, direct comparisons are not possible since they used a leave one out cross validation setting.

Now we want to show that our hypothesis of combining information from both languages is better than looking only at one language. In this setting we used two transcripts per child, one for English and one for Spanish. Table 4 shows the results of this setting over the same 5 classification methods used in the previous experiment. The results improve accuracy by up to 10 % in relation to the first experiment.

	NB	**SVM**	**C4.5**	**1-NN**	**5-NN**
Eng. & Spa.	63.3	**67.8**	49.3	55.6	57.0

Table 4: Accuracy of BOR representation over 5 classification methods: Naïve Bayes (NB), Support Vector Machines (SVM), C4.5, and k-Nearest Neighbors with $k = 1$ (1-NN) and $k = 5$ (5-NN). Using transcripts in both languages: English and Spanish

As we mentioned in related work, the closer work that predicted language dominance and used the same datasets of transcripts (Solorio et al., 2011) shows an accuracy of 72 %. However, they used 9 types of features measuring different dimensions of language combined with some demographic information, and the only type of syntactic information used in that work was at the level of POS n-grams. In this paper we used only the syntactic information extracted from parsing the transcripts in a BOR representation. While our results are a little bit below previous results, they are still relevant in that they show how this syntactic information is valuable, and can outperform other feature types from previous work, including speech fluency measures, language productivity measures, demographic information, morphosyntactic features, speaking rate, and n-grams of POS. We believe that combining this BOR representation with those features used in (Solorio et al., 2011) can boost accuracy further.

7. Conclusions and Future Work

We proposed a representation based on bag of rules from parse trees for the problem of predicting language dominance in Spanish-English children. Our results show that combining information from transcripts in both languages yields the best results. This study also shows that syntactic information is important for language analysis, even though

there could be a considerable amount of noise in the parse trees from having mixed language, as well as child language.

The results obtained are comparable to the recent work looking at the same problem, but different from them we only look at one dimension of language. We only extract features derived from syntactic trees, while previous work looks at vocabulary, language production, fluency, and measures of readability, among others. We predict that adding this dimension to previous work will help achieve higher prediction accuracy.

As future work we want to explore other syntactic information that can also be extracted from the parse trees to build a more robust language model that can improve the results achieved so far. Other things we are working on include the use of different weighting schemes for the rules, such as TF-IDF, and entropy of the grammar rules.

References

Bedore, Lisa M., Peña, Elizabeth D., Gillam, Ron B., and Tsunghan Ho. 2010. Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of Communication Disorders*, 43(6):498–510, Nov-Dec.

Dietterich, Thomas. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924.

Gabani, Keyur, Melissa Sherman, Thamar Solorio, Yang Liu, Lisa M. Bedore, and Elizabeth D. Peña. 2009. A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT) 2009*, pages 46–55, Boulder, Colorado, June. Association for Computational Linguistics.

Mayer, Mercer. 1967. *A boy, a dog, and a frog.* Dial Press, New York, NY.

Mayer, Mercer. 1969a. *Frog goes to dinner.* Dial Press, New York, NY.

Mayer, Mercer. 1969b. *Frog, where are you?* Dial Press, New York, NY.

Mayer, Mercer. 1973. *Frog on his own.* Dial Press, New York, NY.

Paradis, Johanne, Martha Crago, Fred Genesee, and Mabel Rice. 2003. French-English bilingual children with SLI: How do they compare with their monolingual peers? *Journal of Speech, Language, and Hearing Research*, 46:113–127.

Peña, Elizabeth D., Lisa M. Bedore, Ronald B. Gillam, and Thomas Bohman. 2006. Diagnostic markers of language impairment in bilingual children. Grant awarded by the NIDCD, NIH.

Prud'hommeaux, Emily T., Brian Roark, Lois M. Black, and Jan van Santen. 2011. Classification of atypical language in autism. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–96, Portland, Oregon, USA, June. Association for Computational Linguistics.

Roark, Brian, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *BioNLP 2007: Biological, translational, and clinical language processing*, pages 1–8, Prague, June. ACL.

Solorio, Thamar, Melissa Sherman, Yang Liu, Lisa Bedore, Elizabeth Peña, and Aquiles Iglesias. 2011. Analyzing language samples of Spanish-English bilingual children for the automated prediction of language dominance. *Natural Language Engineering*, 17:367–395.

Witten, Ian. H. and Eibe. Frank. 1999. *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, San Francisco, CA.

Recursos y métodos de sustitución léxica en las variantes dialectales en euskera

Resources and methods for lexical substitution between Basque dialects

Larraitz Uria
IKER (UMR5478)
IKERBASQUE
larraitz.uria@ehu.es

Mans Hulden
University of Helsinki
Language Technology
mans.hulden@helsinki.fi

Izaskun Etxeberria
IXA taldea
UPV-EHU
izaskun.etxeberria@ehu.es

Iñaki Alegria
IXA taldea
UPV-EHU
i.alegria@ehu.es

Resumen: La coexistencia de cinco idiomas oficiales en la Península Ibérica (euskera, catalán, gallego, portugués y español) nos lleva a buscar la colaboración para compartir los recursos desarrollados en los diferentes idiomas de la región. Sin embargo, dentro de un mismo idioma se puede dar la coexistencia de más de un dialecto y así ocurre con el euskera. Las herramientas desarrolladas para este idioma se han centrado básicamente en el euskera unificado o estándar, de modo que no funcionan correctamente con los dialectos, que son numerosos. Este trabajo se enmarca dentro de la idea de buscar la forma de establecer semiautomáticamente una relación entre el euskera estándar y sus variantes dialectales. Esto permitiría aplicar las herramientas automáticas disponibles a los textos anteriores a la unificación del idioma, pudiendo explotar de forma automática la gran riqueza lingüística que aportan.
Palabras clave: Morfología computacional, reglas fonológicas, programación lógica inductiva, dialectos.

Abstract: The coexistence of five languages with official status in the Iberian Peninsula (Basque, Catalan, Galician, Portuguese, and Spanish), has prompted collaborative efforts to share and cross-develop resources and materials for these languages of the region. However, it is not the case that comprehension boundaries only exist between each of these five languages; dialectal variation is also present, and in the case of Basque, for example, many written resources are only available in dialectal (or pre-standardization) form. At the same time, all the computational tools developed for Basque are based on the standard language ("Batua"), and will not work correctly with other dialects, of which there are many. In this work we attempt to semiautomatically deduce relationships between the standard Basque and dialectal variants. Such an effort provides an opportunity to apply existing tools to texts issued before a unified standard Basque was developed, and so take advantage of a rich source of linguistic information.
Keywords: Computational morphology, phonological rules, inductive logic programming, dialects.

1. Introducción

En el área de la morfología computacional existe una línea de investigación abierta en relación a la forma de combinar las aproximaciones lingüísticas y las basadas en aprendizaje automático. Los métodos basados en aprendizaje automático (Goldsmith, 2001) pueden ser interesantes cuando se requiere un desarrollo rápido y se cuenta con pocos recursos o no se dispone de expertos en el idioma a tratar. Pero si se quiere un analizador que compagine cobertura y precisión, la mejor opción es una descripción basada en un léxico y un conjunto de paradigmas y reglas fonológicas especificados por expertos. Las descripciones basadas en tecnologías de estados finitos son las más populares para este fin (Beesley y Karttunen, 2002).

El desarrollo de las bibliotecas digitales y de la lingüística basada en corpus impli-

ca a menudo el tratamiento de las variantes dialectales y/o diacrónicas del idioma, pero no resulta viable tener que realizar una nueva especificación por cada variante a tratar. Así pues, el objetivo de nuestras investigaciones es inferir la morfología de las variantes, o la equivalencia entre variantes y formas estándar del euskera a partir de un pequeño corpus paralelo variante/estándar, un corpus de la variante y un analizador o reconocedor del estándar.

En el trabajo que presentamos tratamos de inferir métodos de sustitución léxica entre variantes y formas estándar del euskera basándonos en la morfología. Concretamente, nuestros primeros experimentos se centran en el dialecto labortano y el objetivo es la sustitución léxica de las formas propias del dialecto por las correspondientes del euskera estándar. La tarea clave, en una primera fase al menos, es la inferencia de las reglas fonológicas a partir de pares variante-estándar. En este artículo describimos los recursos básicos con los que contamos en nuestra investigación, así como los métodos que estamos experimentando para inferir las reglas.

Aunque los resultados obtenidos en los primeros experimentos son alentadores, todavía deben ser ampliados y mejorados antes de poder integrarlos en herramientas computacionales efectivas.

Las técnicas que describimos son, en su mayor parte, independientes del idioma y además, es de suponer que con cierta adaptación pueden ser aplicadas a otras variantes o registros del idioma (por ejemplo, idioma más informal: email, SMS...).

2. Trabajos relacionados

El problema general de aprendizaje supervisado de las variantes dialectales ha sido discutido en la literatura en varias áreas: fonología computacional, morfología, aprendizaje automático...

Por ejemplo, (Kestemont, Daelemans, y Pauw, 2010) presentan un sistema independiente del idioma que puede "aprender" variaciones intra-lema. El sistema se utiliza para producir una lematización coherente de textos en holandés antiguo sobre un corpus de literatura medieval (Corpus-Gysseling), que contiene manuscritos de fecha anterior al año 1300.

(Koskenniemi, 1991), por su parte, ofrece un esbozo de un procedimiento de inferencia de reglas fonológicas de dos niveles pero sin llegar a automatizarlo.

En un trabajo anterior, (Johnson, 1984) presenta un "procedimiento de inferencia" para el aprendizaje de reglas fonológicas a partir de datos, lo que puede ser considerado un trabajo precursor del algoritmo ILP (*Inductive Logic Programming*) que proponemos entre nuestros métodos.

3. Recursos lingüísticos

Para el aprendizaje o inferencia y para la evaluación se necesitan recursos que deben ser almacenados, testeados y, en su caso, etiquetados. La idea de este trabajo es usar métodos no supervisados o con un mínimo de supervisión, ya que ése es el escenario realista para generar aplicaciones en el área.

De momento vamos a probar distintas técnicas en el contexto de las variaciones dialectales en euskera, pero intentando que los métodos sean, en la medida de lo posible, independientes del idioma.

Para llevar a cabo nuestros experimentos en esta investigación, contamos con tres corpus de origen y características diferentes:

- Corpus de transcripciones en labortano
- Corpus de la Biblia en euskera estándar y labortano
- Corpus de transcripciones en diversos dialectos

3.1. Corpus de transcripciones en labortano

Por una parte, contamos con un corpus paralelo construido en el centro de investigación IKER (UMR5478) de Bayona (Francia) dentro del proyecto TSABL[1]. El objetivo de este proyecto es el estudio de la variación sintáctica de los dialectos del País Vasco al norte de los Pirineos (*Iparralde*). Para ello, se ha creado la aplicación BASYQUE[2], en la que se recogen datos y ejemplos de variantes dialectales que provienen de tres fuentes de información: cuestionarios específicos, vídeos de testimonios grabados en otros proyectos y textos literarios.

Una de las principales razones que nos ha llevado a utilizar los datos recogidos en

[1] *Towards a Syntactic Atlas of the Basque Language*: http://www.iker.cnrs.fr/-tsabl-towards-a-syntactic-atlas-of-.html?lang=fr

[2] http://ixa2.si.ehu.es/atlas2/index.php?lang=eu

BASYQUE es la posibilidad que nos ofrece de crear corpus paralelos. Los cuestionarios y testimonios grabados se transcriben y junto a cada ejemplo o frase dialectal también se especifica la forma estándar que le corresponde. En el caso de los textos literarios escritos en dialecto, también se indica la forma estándar que corresponde a cada frase. Estos corpus paralelos labortano-estándar son los que vamos a utilizar en los experimentos de sustitución léxica.

La aplicación BASYQUE pretende abarcar todos los dialectos y subdialectos de *Iparralde* y para ello la recopilación de los datos se extiende a todo el territorio. Para los experimentos, en cambio, en esta primera fase nos centramos en el dialecto labortano, por lo que hemos empleado los ejemplos y los textos que provienen de las zonas donde se habla dicho dialecto. Y de momento hemos utilizado los ejemplos recogidos mediante los cuestionarios y los textos literarios, ya que las grabaciones de video no están transcritas todavía. Cabe reseñar que dichos corpus están siendo actualizados y ampliados dentro del mencionado proyecto, de modo que los datos presentados en la Tabla 1 corresponden al corpus de transcripciones labortano-estándar disponible en el momento de realizar los experimentos.

	Corpus	**80 %**	**20 %**
N°frases	2.117	1.694	423
N°palabras	12.150	9.734	2.417
Palabras dif.	3.874	3.327	1.243
Pares filtrados	3.610	3.108	1.172
Pares idénticos	2.532	2.200	871
Pares diferentes	1.078	908	301

Tabla 1: Datos correspondientes al corpus labortano-estándar utilizado en los experimentos realizados hasta el momento. La primera columna corresponde al corpus completo. El 80 % ha sido utilizado en la fase de aprendizaje y el 20 % restante en la fase de test.

En la Tabla 2 se presentan varios ejemplos de frases con el fin de que se vea el tipo de diferencias que se pueden encontrar entre el dialecto y el estándar, así como la correspondencia palabra a palabra con que se cuenta en dicho corpus.

Éste es el corpus en el que hemos centrado nuestros primeros experimentos y con el que hemos obtenido los resultados que presentamos en el apartado 5.

Dialecto labortano vs **Euskera estándar**
Leihoa ***estea*** *erreusitu du.*
Leihoa ***ixtea*** *erreusitu du.*
Eni galdegin ***daut*** *100 euro.*
Eni galdegin ***dit*** *100 euro.*
Ez gero uste izan ***nexkatxa guziek*** *tu egiten* ***dautatela****.*
Ez gero uste izan ***neskatxa guztiek*** *tu egiten* ***didatela****.*

Tabla 2: Varios ejemplos de frases en el corpus paralelo labortano-estándar.

3.2. Corpus de la Biblia

Otra fuente de información básica para nuestro trabajo es la Biblia, que está publicada en euskera estándar y también en dialecto labortano, lo que nos proporciona un corpus paralelo bastante mayor que el anterior. La versión de la Biblia en euskera estándar ha sido editada dos veces, en 1994 y en 2004 respectivamente, y existe una versión electrónica en la web (http://www.biblija.net). En cuanto a la versión en dialecto labortano, se trata de una adaptación de la versión estándar realizada por Marcel Etcehandy y publicada en 2007, y dispone también de una versión electrónica (http://amarauna.org/biblia/). Debido a problemas de formato, de momento sólo hemos alineado 9 libros (elegidos al azar) con las características que se reflejan en la Tabla 3.

N°de libros total	76
N°de libros disponible	66
Palabras totales en euskera estándar	545.700
Palabras diferentes	38.069
Libros alineados	9
Palabras totales en libros alineados	104.967
Palabras diferentes en libros alineados	15.007

Tabla 3: Datos correspondientes al corpus de la Biblia y a los libros alineados hasta la fecha.

Este corpus, al ser de mayor tamaño, nos va a permitir realizar experimentos con distintos tamaños de corpus paralelo, y así conseguir estimar correlaciones entre tamaños de

corpus paralelo y calidad de la inferencia, pero todavía no tenemos resultados que mostrar sobre este aspecto ya que estamos en la fase de preparación y obtención de información de este corpus. Por otro lado, a diferencia del corpus descrito en 3.1, en el corpus de la Biblia no hay transcripción palabra a palabra tal y como se puede observar en el pequeño ejemplo[3] que se presenta a continuación, por lo que la obtención del diccionario de palabras equivalentes se prevé más complicada.

- Dialecto labortano:

  > *"**Errana dauzut: ukan indar eta kuraia.** Ez **ikara**, ez **izi**, ni, Jauna, zure Jainkoa, zurekin **izanen** bainaiz **joanen ziren toki guzietan**".*

- Euskera estándar:

  > *"**Kementsu eta adoretsu izateko esan dizut.** Ez **ikaratu**, ez **kikildu**, ni, Jauna, zure Jainkoa, zurekin **izango** bainaiz **zure ibilera guztietan**".*

3.3. Corpus de transcripciones en diversos dialectos

Existen varios proyectos en el País Vasco (Ahotsak.com[4] o EKE.org[5], por ejemplo) que tienen como objetivo recoger el habla tradicional de cada zona, es decir, recopilar y difundir testimonios orales de vasco-parlantes. En ambos proyectos se graban y se recogen conversaciones y/o testimonios de personas que se expresan en su propio dialecto.

Nosotros hemos creado una red de colaboración con Ahotsak.com para poder recopilar y ayudar a transcribir corpus paralelos de variantes dialectales relacionadas con la forma estándar, ya que el objetivo de Ahotsak.com es ir transcribiendo gran parte de los testimonios grabados. Hasta ahora, cuentan con 5.204 pasajes (1.462.555 palabras) transcritos en las formas dialectales. Sin embargo, para facilitar la búsqueda se quiere relacionar cada forma dialectal con su correspondiente estándar, y para hacerlo de forma (semi)automática nos queremos valer de las técnicas que estamos desarrollando y que describimos posteriormente.

Las características de este corpus son en parte equiparables a las del primer corpus descrito, pero con dos diferencias reseñables:

- recoge gran variedad de dialectos, ya que ciertas formas van cambiando casi de pueblo a pueblo (véase el mapa en http://ahotsak.com/herriak/mapa/)
- de momento sólo disponemos de la transcripción de las formas dialectales y queremos obtener de forma (semi)automática las correspondientes formas estándar. Una parte de la investigación que hacemos es determinar el mínimo de trabajo manual (para relacionar las formas estándar con las dialectales) necesario para obtener unos buenos resultados después en la posterior sustitución léxica.

4. Métodos

Nuestra primera aproximación se va a basar en obtener pares de palabras variante/estándar a partir de un corpus paralelo (que quisiéramos minimizar). Para ello reutilizamos lo que hemos llamado métodos básicos. Posteriormente inferiremos reglas fonológicas mediantes dos métodos.

4.1. Métodos básicos

De cara a obtener pares de palabras equivalentes a partir de corpus paralelos vamos a utilizar dos programas: *lexdiff* y Giza++.

El primero, *lexdiff*, ha sido diseñado y utilizado para la migración automática de textos entre diferentes ortografías del portugués (Almeida, Santos, y Simoes, 2010), debido al cambio de norma que se produjo en ese idioma. Este programa trata de identificar la equivalencia de palabras a partir de frases paralelas. Funciona muy bien cuando los textos son equivalentes palabra por palabra, y es por ello que lo hemos utilizado en los experimentos realizados hasta ahora con el corpus de transcripciones labortano-estándar.

Adicionalmente, *lexdiff* también calcula los cambios de ngramas y sus frecuencias, obteniendo resultados de este tipo: `76 ait ->at`; `39 dautz ->diz`; lo que indica que el ngrama `ait` ha cambiado a `at` 76 veces en el corpus y que `dautz` ha cambiado 39 veces a `diz`.

Estos resultados pueden expresar cambios (morfo)fonológicos regulares entre los textos,

[3] El ejemplo corresponde al versículo 9 del capítulo 1 del libro de Josué.

[4] http://www.ahotsak.com/

[5] http://www.eke.org/

y han sido explotados en el primero de los métodos de inferencia que presentamos a continuación.

Giza++[6] es una conocida herramienta para inferir diccionarios, con probabilidades de traducción, a partir de corpus paralelos. Lo queremos comparar con *lexdiff* dado que el corpus de la Biblia con el que contamos es un corpus paralelo divergente y de mayor tamaño, pero todavía no podemos presentar resultados sobre dicha comparación.

4.2. Métodos de inferencia

Estamos experimentando con dos métodos de inferencia:

1. Inferencia de reglas fonológicas basada en substrings
2. Inferencia usando programación lógica inductiva sobre pares de palabras equivalentes

El método *baseline* consiste en aprender las equivalencias de pares diferentes en el corpus de aprendizaje (corpus paralelo) y sustituirlas en el de test, suponiendo que si no se ha aprendido la forma estándar correspondiente a la variante es la propia variante. Este método tiene como resultado buena precisión y baja cobertura. Los dos métodos que proponemos parten de una lista de equivalencia de palabras o de substrings obtenida por las herramientas básicas y tratan de inferir reglas fonológicas de reemplazamiento que puedan ser compiladas por *xfst* de Xerox (Beesley y Karttunen, 2002) o *foma* (software libre, (Hulden, 2009)).

4.2.1. Inferencia de reglas fonológicas basada en substrings.

En principio se basa en los cambios de ngramas que obtiene *lexdiff*. Hay varias formas de transformar esa salida de *lexdiff* en reglas de reemplazamiento que se compilan a transductores finitos. Estamos teniendo en cuenta los siguientes factores:

- Limitar los cambios a tener en cuenta a aquellos que tienen un mínimo de frecuencia (por ejemplo, dos o tres). Si aumentamos el mínimo mejoraremos la precisión, pero perderemos cobertura.
- Limitar el número de reglas que pueden ser aplicadas a la misma palabra. Por ejemplo, la correspondencia *agerkuntza/agerpena* puede expresarse mediante dos reglas: `rkun ->rpen` y `ntza ->na`, pero permitir varios cambios puede producir ruido innecesario y bajar la precisión.
- La forma de aplicar las reglas: secuencialmente o paralelamente.
- Hacer que los cambios sean de longitud mínima y condicionados por el contexto.

4.2.2. Inferencia usando programación lógica inductiva.

El segundo método consiste en los siguientes pasos:

1. Alinear los pares de palabras letra por letra usando la mínima distancia de edición.
2. Extraer un conjunto de reglas fonológicas.
3. Por cada regla, buscar contraejemplos.
4. Buscar la restricción de contexto mínima que resuelva los contraejemplos.

Por ejemplo, si tenemos los pares *emaiten/ematen* e *igorri/igorri*, en el primer paso se detecta el cambio *i/0*, que en el paso dos se convierte en la regla `i ->0`. Pero ese cambio no se puede aplicar con *igorri*, por lo que la regla se transforma para evitar que sea aplicada. Este método tiene la ventaja de explotar las formas que son idénticas en el dialecto y en el estándar.

5. Resultados y trabajos futuros

Hemos centrado los experimentos en el corpus descrito en el apartado 3.1 con el fin de testear y evaluar los métodos descritos en el apartado 4. Los primeros resultados nos muestran una mejora respecto al método *baseline*, pero todavía deben ser mejorados para utilizarlos en herramientas computacionales efectivas.

La Tabla 4 muestra los resultados obtenidos. Dichos resultados corresponden tanto al método *baseline*, como a los mejores resultados obtenidos con cada una de las propuestas de inferencia de reglas descritas y se expresan en términos de precisión (*precision*), cobertura (*recall*) y la medida-F (*F-score*), que es la combinación de ambas. En los tres casos, el proceso de aprendizaje se ha llevado a cabo

[6] http://code.google.com/p/giza-pp/

con el 80 % del corpus, y el test, cuyos resultados son los que see muestran en la Tabla 4, se ha realizado sobre el 20 % restante.

Aunque no se presentan más que los mejores resultados obtenidos con cada método, el número de experimentos realizados con ambos métodos ha sido numeroso, sobre todo con el método de inferencia de reglas basada en substrings, debido a los diferentes factores que se pueden tener en cuenta para inferir las reglas fonológicas. Dichos experimentos nos muestran que:

- Disminuir la mínima frecuencia exigida a un cambio para obtener una regla fonológica a partir de él, aumenta notablemente la cobertura, pero también hace que disminuya la precisión, con lo que el resultado en términos de F-score apenas mejora.
- La aplicación de más de una regla en una palabra no parece aportar incrementos importantes en la mejora de los resultados.
- El modo de aplicación, secuencial o paralelo, de las reglas (cuando se aplica más de una regla en la misma palabra) presenta resultados muy similares, aunque algo mejores si la aplicación es paralela.
- Por último, minimizar la longitud de los cambios y hacer que sean condicionados por el contexto, obtiene claramente mejores resultados.

En los primeros experimentos con este método de inferencia, ya pudimos comprobar que la aplicación exclusivamente de las reglas fonológicas no mejoraba los resultados del método *baseline*, debido a que la precisión era excesivamente baja (para cada término a sustituir, el número de candidatos era a menudo elevado). Ello nos llevó a aplicar un post-filtro al proceso, basado en la frecuencia de los candidatos en euskera estándar[7]. El filtro aplicado es muy simple: si hay más de un candidato se elige el más frecuente, pero a pesar de su simplicidad se mejoran los resultados y se consigue superar el *baseline* tal y como se puede ver en los resultados presentados en la Tabla 4.

[7]La frecuencia de cada término la hemos obtenido de un corpus de un diario de noticias editado en euskera.

	Precision	**Recall**	**F-score**
Baseline	95,62	43,52	59,82
Método 1	75,10	60,13	66,79
Método 2	85,02	58,47	69,29

Tabla 4: Mejores resultados (en términos de *F-score*) obtenidos con ambos métodos de inferencia en los experimentos realizados con el corpus de transcripciones labortano-estándar.

Con respecto al segundo método de inferencia, basado en programación lógica inductiva, los resultados obtenidos han sido mejores, y además, con este método no es necesaria la aplicación del filtro posterior. El motivo fundamental es que este método no sólo utiliza la información de los pares diferentes, sino también la de los pares iguales en el dialecto y en el estándar.

Se puede consultar información más detallada tanto de los métodos propuestos como de la evaluación realizada en (Hulden et al., 2011).

Todavía nos queda mucho trabajo por realizar en el campo de esta investigación. La aplicacion de los métodos descritos al corpus de la Biblia nos va a permitir precisar hasta qué punto es determinante que la transcripción entre dialecto y estándar sea palabra a palabra, y qué tamaño de corpus es necesario para obtener resultados que indiquen que es posible conseguir herramientas automáticas de sustitución léxica.

Además, creemos que los métodos utilizados deben ser combinados con otros que infieran relaciones entre lemas y morfemas (variantes y formas estándar), variantes de paradigmas y que contrasten esas inferencias con corpus de variantes (sin que sean corpus paralelos) más amplios.

Bibliografía

Almeida, J. J, A. Santos, y A. Simoes. 2010. Bigorna–a toolkit for orthography migration challenges. En *Seventh International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta.*

Beesley, K. R y L. Karttunen. 2002. Finite-state morphology: Xerox tools and techniques. *Studies in Natural Language Processing. Cambridge University Press.*

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198.

Hulden, M., I. Alegria, I. Etxeberria, y M. Maritxalar. 2011. An unsupervised method for learning morphology of variants from the standard morphology and a little parallel corpus. En *(EMNLP workshop) Dialects-2011 — First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*.

Hulden, Mans. 2009. Foma: a finite-state compiler and library. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, páginas 29–32, Athens, Greece. Association for Computational Linguistics.

Johnson, Mark. 1984. A discovery procedure for certain phonological rules. En *Proceedings of the 10th international conference on Computational linguistics*, COLING '84, páginas 344–347. Association for Computational Linguistics.

Kestemont, M., W. Daelemans, y G. De Pauw. 2010. Weigh your words—memory-based lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25(3):287–301.

Koskenniemi, K. 1991. A discovery procedure for two-level phonology. *Computational Lexicology and Lexicography: A Special Issue Dedicated to Bernard Quemada*, páginas 451–446.