BUAP-UPV TPIRS: A System for Document Indexing Reduction at WebCLEF

David Pinto[1,2], Héctor Jiménez-Salazar[2], Paolo Rosso[1], and Emilio Sanchis[1]

[1] Department of Information Systems and Computation,
UPV, Valencia 46022,
Camino de Vera s/n, Spain
{dpinto, prosso, esanchis}@dsic.upv.es
[2] Faculty of Computer Science, BUAP, Puebla 72570,
Ciudad Universitaria, Mexico
{davideduardopinto, hgimenezs}@gmail.com

Abstract. In this paper we present the results of BUAP/UPV universities in WebCLEF, a particular task of CLEF 2005. Particularly, we evaluate our information retrieval system at the bilingual "English to Spanish" task. Our system uses a term reduction process based on the Transition Point technique. Our results show that it is possible to reduce the number of terms to index, thereby improving the performance of our system. We evaluate different percentages of reduction over a subset of EuroGOV, in order to determine the best one. We observed that after reducing the 82.55% of the corpus, a Mean Reciprocal Rank of 0.0844 was obtained, compared with 0.0465 of such evaluation with full documents.

1 Introduction

High volume of information in Internet led to the development of novel techniques for managing of data, specially when we deal with information in multiple languages. There are sufficient example scenarios in which users may be interested in information which is in a language other than their own native language. A common language scenario is where a user has some comprehension ability for a given language but s/he is not sufficiently proficient to confidently specify a search request in that language. Thus, a search system that can deal with this problem should be of a high benefit. The World Wide Web (WWW) is a natural setting for cross-lingual information retrieval; the European Union is a typical example of a multilingual scenario, where multiple users have to deal with information published in at least 20 languages.

In order to reinforce research in this area, CLEF (Cross-Language Evaluation Forum) has been compiling a set of multi-lingual corpora and promoting the evaluation of multiple multi-lingual information retrieval systems for diverse kinds of data [5]. A particular task for the evaluation of such systems that deal with information on the web has been set up this year as a part of CLEF. This forum was named WebCLEF, and the best description of this particular task

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 873–879, 2006.

can be seen in [14]. In WebCLEF, three subtasks were defined within this year: mixed monolingual, multilingual, and bilingual English to Spanish.

This paper reports results on the evaluation of a Cross-Language Information Retrieval System (CLIRS) for the bilingual English to Spanish subtask of WebCLEF 2005. A document indexing reduction is proposed, in order to improve precision of CLIRS and to diminish the storing space on such systems. Our proposal is based on the use of the Transition Point (TP) technique, which is somehow a method that obtains important terms from a document. We evaluate different percentages of TP over a subset of EuroGOV corpus [13], and we observed that it is possible to improve precision results by reducing the number of terms for a given corpus.

The next section describes our information retrieval system in detail. Section 3 briefly introduces the corpus used in our experiments, and the results obtained after evaluation. Finally, a discussion of our experiments is presented.

2 Description of TPIRS

We used a boolean model with Jaccard similarity formula for our CLIRS. Our goal was to determine the behaviour of document indexing reduction in an information retrieval environment. In order to reduce the terms from every document treated, we applied a technique named Transition Point, which is described as follows.

2.1 The Transition Point Technique

The Transition Point (TP) is a frequency value that splits the vocabulary of a document into two sets of terms (low and high frequency). This technique is based on the Zipf Law of Word Ocurrences [18] and also on the refined studies of Booth [2], as well as of Urbizagástegui [17]. These studies are meant to demonstrate that mid-frequency terms are closely related to the conceptual content of a document. Therefore, it is possible to form the hypothesis that terms closer to TP can be used as indexes of a document. A typical formula used to obtain this value is given in equation 1:

$$TP = \frac{\sqrt{8 * I_1 + 1} - 1}{2}, \tag{1}$$

where I_1 represents the number of words with frequency equal to 1 [12] [17].

Alternatively, TP can be localized by identifying the lowest frequency (from the highest frequencies) that it is not repeated in each document; this characteristic comes from the properties of the Booth's law of low frequency words [2]. In our experiments we have used this approach.

Let us consider a frequency-sorted vocabulary of a document; i.e., $V_{TP} = [(t_1, f_1), ..., (t_n, f_n)]$, with $f_i \geq f_{i-1}$, then $TP = f_{i-1}$, iif $f_i = f_{i+1}$. The most important words are those that obtain the closest frequency values to TP, i.e.,

$$TP_{SET} = \{t_i | (t_i, f_i) \in V_{TP}, U_1 \leq f_i \leq U_2\}, \tag{2}$$

where U_1 is a lower threshold obtained by a given neighbourhood percentage of TP (NTP), thus, $U_1 = (1 - NTP) * TP$. U_2 is the upper threshold and it is calculated in a similar way ($U_2 = (1 + NTP) * TP$).

We have used the TP technique in different areas of Natural Language Processing (NLP) like: clustering of short texts [7], categorization of texts [9], keyphrases extraction [10] [16], summarization [3], and weighting models for information retrieval systems [4]. Thus, we believe that there exist enough evidence to use this technique as a terms reduction process.

2.2 Information Retrieval Model

Our information retrieval is based on the Boolean Model, and, in order to rank documents retrieved, we used the Jaccard's similarity function, applied to both, the query and every document of the corpus used. Previously, each document was preprocessed and its index terms were selected (the preprocessing phase is described in section 3.1). For this purpose, several values of a neighbourhood of TP were used as thresholds, as equation 2 indicates.

3 Evaluation

3.1 Corpus

We used a subset of the EuroGOV corpus for our evaluation. This subset was composed by a set of Spanish Internet pages, originally obtained from European government-related sites.

In order to construct this corpus, for every page compiled in the EuroGOV corpus, we determine its language by using TexCat [15], a language identification program widely used. We construct our evaluation corpus with those documents identified as Spanish language.

The preprocessing process consisted of the elimination of punctuation symbols, Spanish stopwords, numbers, html tags, script codes and cascading style sheets codes.

For the evaluation of this corpus, a set of 134 queries was composed and refined, in order to provide gramatically correct "English" queries. Supervised queries (queries and related webpages) were created by the participants in the WebCLEF task, and the particular case of the queries were later reviewed and in some cases corrected in their English translation by the NLP Group at UNED. Queries were distributed in the following way: 67 homepages and 67 named page findings.

We applied a preprocessing phase to this set of queries. First, we used an online translation system [1] in order to translate every query from English to Spanish. After that, an elimination of punctuation symbols, spanish stopwords and numbers was done.

We did not apply a rigorous method of translation, due to the fact that our main goal in our first participation in WebCLEF was to determine the quality of terms reduction in our CLIRS.

[1] http://www.freetranslation.com

3.2 Indexing Reduction

In order to determine the behaviour of document indexing reduction on CLIRS, we submitted to the contest, a set of five runs, which are described as follows.

First Run: → Full: This run used "Full documents" as evaluation corpus, and conformed the baseline for our experiments.

Second Run: → TP10: This run used an evaluation corpus composed by the reduction of every document, using the TP technique with a neighbourhood of 10% around TP.

Third Run: → TP20: This run used an evaluation corpus composed by the reduction of every document, using the TP technique with a neighbourhood of 20% around TP.

Fourth Run: → TP40: This run used an evaluation corpus composed by the reduction of every document, using the TP technique with a neighbourhood of 40% around TP.

Fifth Run: → TP60: This run used an evaluation corpus composed by the reduction of every document, using the TP technique with a neighbourhood of 60% around TP.

Table 1 shows the size of every evaluation corpus used, as well as the percentage of reduction obtained for each one. As can be seen, the TP technique obtained a big percentage of reduction (between 75 and 89%), which also implies a reduction in time for the indexing process in a CLIRS.

Table 1. Evaluation corpora

Corpus	Size (Kb)	% of Reduction
Full	117,345	0%
TP10	12,616	89.25%
TP20	19,660	83.25%
TP40	20,477	82.55%
TP60	28,903	75.37%

3.3 Results

Table 2 shows the results for every run submitted. The first column indicates the name of each run. The last column shows the Mean Reciprocal Rank (MRR) obtained for each run. Additionally, the average success at different number of documents retrieved is shown; by instance, the second column indicates the average success of the CLIRS at the first answer. The "TP20" approach, obtained fewer than 50 results, and therefore, it average success at 50 was not calculated.

As can be seen, an important improvement was gained by using an evaluation corpus obtained with a neighbourhood of 40% of TP. We were hoping to obtain comparable results with the "Full" run, but as can be seen, the "TP40" run received double the score of the "Full" run when evaluated using MRR.

Table 2. Evaluation results

Corpus	Average Success at 1	5	10	20	50	Mean Reciprocal Rank
Full	0.0224	0.0672	0.1119	0.1418	0.1866	0.0465
TP10	0.0224	0.0373	0.0672	0.0821	0.1119	0.0331
TP20	0.0299	0.0448	0.0672	0.1045	–	0.0446
TP40	**0.0597**	0.0970	0.1119	0.1418	**0.2164**	**0.0844**
TP60	0.0522	**0.1045**	**0.1269**	**0.1642**	0.2090	0.0771

Three teams participated at the bilingual "English to Spanish" subtask at WebCLEF. Every team submitted at least one run [1,11,8]. A comparison among the results obtained by each team can be seen in Table 3. Our second place in this contest can be dramatically improved by applying a better translation process and by using a better representation model for our information retrieval system.

Table 3. All teams results

Team Name	Average Success at 1	5	10	20	50	Mean Reciprocal Rank over 134 Topics
UNED	**0.0821**	**0.1045**	**0.1194**	0.1343	0.2090	**0.0930**
BUAP/UPV	0.0597	0.0970	0.1119	**0.1418**	**0.2164**	0.0844
ALICANTE	0.0299	0.0522	0.0597	0.0746	0.0970	0.0395

4 Conclusions

We have proposed an index reduction method for a cross-lingual information retrieval system. Our proposal is based on the transition point technique.

After submitting five runs at the bilingual English to Spanish subtask of WebCLEF, we observed that it is possible to reduce terms in the documents that conform the corpus of a CLIRS, not only by reducing the time needed for indexing but also by improving the precision of the results obtained by CLIRS.

Our method is linear in computational time, and therefore it can be used in practical tasks. Until now, results obtained in terms of MRR are very low, but findings show that by applying better techniques of English to Spanish translation of queries, results can be dramatically improved [6].

We were concerned with the impact of indexing reduction on CLIRS, and in the future we hope to improve other components of our CLIRS, for instance, the use of vector space model, in order to improve the MRR.

The TP technique has shown an effective use on diverse areas of NLP, and its best features for NLP, are mainly two: a high content of semantic information and the sparseness that can be obtained on vectors for document representation on models based on the vector space model. On the other hand, its language independence allows to use this technique in CLIRS, that is the matter of WebCLEF.

Acknowledgments

This work was partially supported by BUAP-VIEP 3/G/ING/05, R2D2 (CICYTTIC 2003-07158-C04-03), ICT EU-India (ALA/95/23/2003/077-054), and Generalitat Valenciana Grant (CTESIN/2005/12). We would like also thank the CLEF 2005 organising committe (This work is a revised version of the paper "TPIRS: A System for Document Indexing Reduction at WebCLEF").

References

1. J. Artiles, V. Peinado, A. Peñas, F. Verdejo: *UNED at WebCLEF 2005*, Extended abstract in Working notes of CLEF'05, Viena, 2005.
2. A. Booth: *A Law of Ocurrences for Words of Low Frequency*, Information and control, 1967.
3. C. Bueno, D. Pinto, H. Jimenez: *El párrafo virtual en la generación de extractos*, Research on Computing Science Journal, ISSN 1665-9899, 2005.
4. R. Cabrera, D. Pinto, H. Jimenez, D. Vilariño: *Una nueva ponderación para el modelo de espacio vectorial de recuperación de información*, Research on Computing Science Journal, ISSN 1665-9899, 2005.
5. CLEF 2005: *Cross-Language Evaluation Forum*, http://www.clef-campaign.org/, 2005.
6. W. B. Croft: *Language Modeling for Information Retrieval*, The Information Retrieval Series, Vol. 13, Lafferty, John (Eds.), 2003.
7. H. Jimenez, D. Pinto, P. Rosso: *Selección de Términos No Supervisada para Agrupamiento de Resúmenes*, In proceedings of Workshop on Human Language, ENC05, 2005.
8. T. Martínez, E. Noguera, R. Muñoz, F. Llopis: *Web Track for CLEF2005 at ALICANTE UNIVERSTITY*, Extended abstract in Working notes of CLEF'05, Viena, 2005.
9. E. Moyotl, H. Jimenez: *An Analysis on Frequency of Terms for Text Categorization*, Proceedings of XX Conference of Spanish Natural Language Processing Society (SEPLN-04), 2004.
10. D. Pinto, F. Pérez: *Una Técnica para la Identificación de Términos Multipalabra*, In Proceedings of 2nd. National Conference on Computer Science, Mexico, 2004.
11. D. Pinto, H. Jiménez-Salazar, P. Rosso, E. Sanchis: *TPIRS: A System for Document Indexing Reduction on WebCLEF*, Extended abstract in Working notes of CLEF'05, Viena, 2005.
12. B. Reyes-Aguirre, E. Moyotl-Hernández & H. Jiménez-Salazar: *Reducción de Términos Indice Usando el Punto de Transición*, In proceedings of Facultad de Ciencias de Computación XX Anniversary Conferences, BUAP, 2003.
13. B. Sigurbjörnsson, J. Kamps, and M. de Rijke: *EuroGOV: Engineering a Multilingual Web Corpus*, In Proceedings of CLEF 2005, 2005.
14. B. Sigurbjörnsson, J. Kamps, and M. de Rijke: *WebCLEF 2005: Cross-Lingual Web Retrieval*, In Proceedings of CLEF 2005, 2005.
15. TextCat: *Language identification tool*, http://odur.let.rug.nl/ vannord/TextCat/, 2005.

16. M. Tovar, M. Carrillo, D. Pinto, H. Jimenez: *Combining Keyword Identification Techniques*, Research on Computing Science Journal, ISSN 1665-9899, 2005.
17. R. Urbizagástegui: *Las posibilidades de la Ley de Zipf en la indización automática*, Research report of the California Riverside University, 1999.
18. G. K. Zipf: *Human Behavior and the Principle of Least-Effort*, Addison-Wesley, Cambridge MA, 1949.