Text Reduction-Enrichment at WebCLEF*

[(1)]Franco Rojas, [(1)]Héctor Jiménez-Salazar & [(1,2)]David Pinto
[1]Faculty of Computer Science, BUAP, Mexico
[2]Department of Information Systems and Computation, UPV, Spain
{frlb99, hgimenezs, davideduardopinto}@gmail.com

Abstract

In this paper we are reporting the results obtained after submitting one run to the Mixed Monolingual task of WebCLEF 2006. We have used a text reduction process based on the selection of mid-frequency terms. Although our approach enhances precision, it must be improved in recall by an enrichment process based on the addition of high co-ocurrence terms. We have seen that a improvement of 40% in the corpus used last year in the BiEnEs was obtained. But we also observed that low Mean Reciprocal Rank (MRR) values were obtained compared with those of the mixed monolingual task of WebCLEF 2005. We consider that our low MRR is derived of a bad preprocessing phase, but we must investigate this issue in detail.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Text reduction, Text enrichment, Mixed-Monolingual

1 Introduction

The big explosion of information published in Internet led us to develop 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 different language 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 several languages.

The Cross-Language Evaluation Forum (CLEF) has gathered a multi-lingual corpus and promotes the evaluation of cross-lingual information retrieval systems for different types of data [2]. WebCLEF is a particular task for the evaluation of such systems that deals with information on the Web [6].

*This work was partially supported by FCC-BUAP and the BUAP-701 PROMEP/103.5/05/1536 grant.

Nowadays, WebCLEF have defined one task for the evaluation of search engines: the Mixed Monolingual. Thus, in this paper we are reporting the results obtained after the submission of one run to this task.

We have used a text reduction and enrichment process and, therefore, we organized this document in three sections. The next section describes the components of our search engine. In Section 3.3 the evaluation results are presented, and finally a discussion of findings are given.

2 Description of the search engine

We used a boolean model with Jaccard similarity formula for our system. 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 text into two sets of terms (low and high frequency). This technique is based on the Zipf Law of Word Ocurrences [9] and also on the refined studies of Booth [1], as well as of Urbizagástegui [8]. 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: $TP = (\sqrt{8 * I_1 + 1} - 1)/2$, where I_1 represents the number of words with frequency equal to 1; see [4] [8].

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 [1]. 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.,

$$TP_{SET} = \{t_i | (t_i, f_i) \in V_{TP}, U_1 \leq f_i \leq U_2\}, \quad (1)$$

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 ($U_2 = (1 + NTP) * TP$). Either in WebCLEF-2005 and in the current competition, we have used $NTP = 0.4$, considering that the TP technique is language independent.

2.2 Term Enrichment

Certainly TP reduction may increase precision, but furthermore it decreases recall. Due to this fact, we enriched the selected terms by obtaining new terms, those with similar characteristics to the initial ones. Specifically, given a text T, with selected terms TP_{SET}, y is a new term if it co-occurs with some $x \in TP_{SET}$, i.e.,

$$TP'_{SET} = TP_{SET} \cup \{y | x \in TP_{SET} \wedge (fr(xy) > 1 \vee fr(yx) > 1)\}. \quad (2)$$

Considering the text length, we only selected a window of size 1 around each term of S, and a minimum frequency of two for each bigram was required as condition to include new terms.

2.3 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). As we will see in Section 3.3 we represent each text using the selection given by equation 1, additionally, after reduction, we carried out an enrichment process based on the identification of related terms to those selected, Eq. 2.

3 Evaluation

3.1 Corpus

We used the EuroGOV corpus provided by the WebCLEF forum which is better described in [5], but we indexed only 20 domains: DE, AT, BE, DK, SI, ES, EE, IE, IT, SK, LU, MT, NL, LV, PT, FR, CY, GR, HU, and UK (we did not indexed the following domains: EU, RU, FI, PL, SE, CZ, LT). Due to this fact, only 1470 from 1939 topics were evaluated, which is approximately a 75,81% of the total of topics. Although we presented in Section 3.3 the MRR over 1939 topics, 469 topics related with the not indexed domains were not evaluated.

The preprocessing phase of the EuroGOV corpus was carried out by writing two scripts for obtaining the terms to be indexed from each document. The first script uses regular expressions for excluding all the information which is enclosed by the characters $<$ and $>$. Although this script obtains very good results, it is very slow and therefore we decided to used it only with three domains of the EuroGOV collection, namely Spanish (ES), French (FR), and German (DE).

On the other hand, we wrote a script based in the *html* syntax for obtaining all the terms considered interesting for indexing, i.e., those different than script codes (javascript, vbscript, style cascade sheet, etc), *html* codes, etc. This script speeded up our indexing process but it did not took into account that some web pages are incorrectly written and, therefore, we missed important information from those documents.

For every page compiled in the EuroGOV corpus, we also determine its language by using TexCat [7], a language identification program widely used. We construct our evaluation corpus with those documents identified as a language of the above list.

Another preprocessing problem consisted in the charset codification, which leads to a even more difficult analysis. Although the EuroGOV corpus is given in UTF-8, the documents that made up this corpus does not neccesarily keep this charset. We have seen that for some domains, the charset codification is given in the *html* metadata tag, but also we found that this codification could be wrong, perhaps because it was filled without the supervision of the creator of that page, who may be does not know anything, and evenmore does not matter about charsets codifications. We consider it as the most difficult problem in the preprocessing process.

Finally, we eliminated stopwords for each language (except for Greek language) and punctuation symbols. The same process was applied to the queries.

For the evaluation of this corpus, a set of queries was provided by WebCLEF-2006.

3.2 Indexing reduction

After our first participation in WebCLEF [3], we carried out more experiments using only those documents in Spanish language from the EuroGOV corpus. We observed that a value of $NTP = 0.4$ using the reduction process shown in the Equation 1 was adequated. Therefore, in this test we carried out one run with that value. Moreover, this run took the evaluation corpus composed by the reduction of every text, using TP technique with a neighbourhood of 40% around TP, an enriched this set of terms using related terms as described by Equation 2.

Table 1 shows the size of every evaluation corpus used; the vocabulary composed by representation of all texts, $|TP'_{SET}|$, as well as the percentage of reduction obtained by each one with respect to the original text. As we can see, the TP technique obtained a percentage of reduction lower than 5%, which also implies a reduction in time for the indexing process in a search engine.

Table 1: Vocabulary size and percentage of reduction.

Domain	DE	AT	BE	DK	SI	ES	EE	IE	IT	SK
Size (KB)	2,588	2,317	6,796	1,189	6,729	16,271	4,838	2,632	11,913	14,668
% of Reduction	4.7	2.8	2.0	2.1	2.9	1.5	2.8	4.0	1.6	2.5
Domain	LU	MT	NL	LV	PT	FR	CY	GR	HU	UK
Size (KB)	3,212	4,817	20,324	21,213	9,134	22,083	18,814	340	10,440	14,239
% of Reduction	0.8	4.3	2.3	2.2	2.4	4.2	3.5	2.6	1.2	3.9

3.3 Results

Table 2 shows the results for the run submitted. The first and second column indicates the number of topics evaluated and the test type. The last column shows the Mean Reciprocal Rank (MRR) obtained for each test. Additionally, the average success at different number of documents retrieved is shown; for instance, the second column indicates the average success of our search engine at the first answer.

Table 2: Evaluation results

		Average Success at					
#Topics	**Test**	**1**	**5**	**10**	**20**	**50**	**Mean Reciprocal Rank**
1939	All	0.0093	0.0217	0.0294	0.0371	0.0464	0.0157
1620	Auto	0.0025	0.0049	0.0086	0.0117	0.0160	0.0040
319	Man	0.0439	0.1066	0.1348	0.1661	0.2006	0.0750
810	A. bi.	0.0037	0.0062	0.0099	0.0123	0.0148	0.0049
124	M. new	0.0323	0.0968	0.1129	0.1613	0.2339	0.0657
810	A. uni.	0.0012	0.0037	0.0074	0.0111	0.0173	0.0031
195	M. old	0.0513	0.1128	0.1487	0.1692	0.1795	0.0810

4 Conclusions

We have used an index reduction method for our search engine that includes an enrichment step. Our proposal is based on the transition point technique which allows to obtain mid-frequency terms from every document to be indexed. Our method is linear in computational time and, therefore, it can be used in a wide spectrum of practical tasks.

After submitting our run we observed enhancement if we compare the results obtained with those of the BiEnEs task in WebCLEF 2005. By using the enrichment, more than 40% on MRR was achieved. However, using the Vector Space Model similar results to boolean model were obtained.

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 multilingual environments.

References

[1] A. Booth: *A Law of Ocurrences for Words of Low Frequency*, Information and control, 1967.

[2] CLEF 2005: *Cross-Language Evaluation Forum*, http://www.clef-campaign.org/, 2005.

[3] 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.

[4] 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.

[5] B. Sigurbjörnsson, J. Kamps, and M. de Rijke: *EuroGOV: Engineering a Multilingual Web Corpus*, In Proceedings of CLEF 2005, 2005.

[6] B. Sigurbjörnsson, J. Kamps, and M. de Rijke: *WebCLEF 2005: Cross-Lingual Web Retrieval*, In Proceedings of CLEF 2005, 2005.

[7] TextCat: *Language identification tool*, http://odur.let.rug.nl/ vannord/TextCat/, 2005.

[8] R. Urbizagástegui: *Las posibilidades de la Ley de Zipf en la indización automática*, Research report of the California Riverside University, 1999.

[9] G. K. Zipf: *Human Behavior and the Principle of Least-Effort*, Addison-Wesley, Cambridge MA, 1949.