Applying Light Natural Language Processing to Ad-Hoc Cross Language Information Retrieval

Christina Lioma Craig Macdonald Ben He Vassilis Plachouras Iadh Ounis
Department of Computing Science
University of Glasgow
G12 8QQ

Abstract

In the CLEF 2005 Ad-Hoc Track we experimented with language-specific morphosyntactic processing and light Natural Language Processing (NLP) for the retrieval of Bulgarian, French, Italian, English and Greek.

Categories and Subject Descriptors

[H.3 Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing – *Linguistic Processing*; H.3.3 Information Search and Retrieval – *Information Filtering, Retrieval Models.*

Keywords

Cross Language Information Retrieval, Morphological Analysis, Part-of-Speech Tagging.

1 Introduction

This paper is organised as follows. Section 2 presents an overview of the linguistic foundations of this work. Section 3 discusses our monolingual and bilingual runs. Section 4 concludes with a summary of the approaches tested and the extent of their success.

The driving force behind our participation in CLEF 2005 has been to explore the effect of diverse languages across a set Information Retrieval (IR) platform. It was anticipated that this effect would be considerable, not only in terms of technical implementation issues, but also in terms of language resources. We used the same retrieval platform as reported in CLEF 2004 [2], on top of which we added Natural Language Processing (NLP).

2 Linguistic Background

NLP is assumed to be essential in IR for morphologically rich languages. We tested the validity of this statement for Greek – English retrieval. Also, noun phrase extraction, a popular NLP application, has been tested for monolingual French and bilingual Italian – French retrieval, using our in-house Noun Phrase (NP) extractor. Other NLP applications used in the context of this work include light syntactic analysis, achieved by a probabilistic part-of-speech tagger, lemmatisation and morphological analysis [6, 7].

3 Monolingual and Bilingual Runs

The motivation behind our participation in CLEF 2005 was to examine the performance of a set IR platform across a span of dissimilar languages, and thus reveal the extent to which retrieval models and system tuning issues are accountable for IR performance on a per language basis. We used our existing retrieval platform, which accommodates a range of matching models and a strong query expansion baseline [2, 5].

For Bulgarian, the lack of language resources meant that the collection was simply stemmed and indexed, without any supplementary analysis. Stemming was realised using the Russian Snowball stemmer [4]. For the English – Bulgarian retrieval, the Skycode machine translation system was used [3]. Table 2 summarises these runs. Submitted runs are printed in boldface and optimal runs in italics.

		Title+Description MAP			*Title+Description+Narrative MAP*		
	Model	*BG*	*EN-BG*	*% mono*	*BG*	*EN-BG*	*% mono*
Query Expansion False	DLH	0.2211	**0.1290**	58.34%	0.2036	0.1316	64.64%
	PL2	0.2363	0.1294	54.76%	0.2203	**0.1344**	61.01%
Query Expansion True	DLH	0.2409	0.1534	63.68%	0.2277	**0.1668**	73.25%
	PL2	**0.2514**	0.1685	67.02%	**0.2412**	**0.1799**	74.58%

Table 2. Bulgarian and English – Bulgarian Mean Average Precision (MAP)

Table 2 reveals the modifying influence of translation on retrieval performance, which is even stronger for shorter topics. The overall performance of the matching models is highly correlated, with a *p*-value of 0.00048.

This delineates the need for additional language resources for Bulgarian, whose evidence would weigh more heavily on retrieval performance than that of simple stemming.

For Greek we used a POS tagger and morphological analyser developed by Xerox [8]. Closed class terms were rejected to reduce noise, while lemmas were automatically translated into English using Babelfish [1]. The performance of these runs is presented in Table 3. Submitted runs are printed in boldface and optimal runs in italics.

		Title+Description MAP			*Title+Description+Narrative MAP*		
	Model	*EN*	*GR-EN*	*% mono*	*EN*	*GR-EN*	*% mono*
Query Expansion False	InexpB2	0.4115	0.2724	66.20%	0.4303	**0.2295**	53.33%
	PL2	0.3634	0.2574	70.83%	0.4042	0.2126	52.60%
Query Expansion True	InexpB2	0.4307	**0.2935**	68.14%	0.4433	0.3117	70.31%
	PL2	0.3961	0.2488	62.81%	0.4347	0.2838	65.29%

Table 3. English and Greek – English Mean Average Precision (MAP)

Table 3 shows that translation has a considerable effect on retrieval performance. The overall scores for Greek – English retrieval are closer to their monolingual equivalents than the English – Bulgarian scores are to the monolingual Bulgarian scores. This underlines the auxiliary service rendered to the Greek topics by the morphological analysis and lemmatisation. Table 3 also reveals that light syntactic analysis and lemmatisation can assist retrieval just as well as stemming.

For French we used a variation of the monolingual French strategy tested in CLEF 2004 [2]. We opted for a less aggressive stemming approach, which targets mainly inflectional variants. A probabilistic POS tagger [6] provided a pellucid syntactic analysis of the topics. Noun phrases were extracted using our in-house NP extractor. For Italian – French, Italian noun phrases were extracted and translated separately into French, using Worldlingo [7]. The collective scores of the above runs are presented in Table 4. Submitted runs are printed in boldface and optimal runs in italics.

			Title+Description MAP			*Title+Description+Narrative MAP*		
		Model	*FR*	*IT-FR*	*% mono*	*FR*	*IT-FR*	*% mono*
POS NP True	*Query Expansion False*	DLH	0.3228	**0.2066**	64.00%	0.3371	0.2305	68.38%
		PL2	0.3092	0.2070	66.95%	**0.3206**	**0.2291**	71.46%
	Query Expansion True	DLH	**0.4017**	0.2731	67.99%	0.4198	0.3029	72.15%
		PL2	**0.3765**	0.2626	69.75%	0.3809	**0.2883**	75.69%
POS NP False	*Query Expansion False*	DLH	0.3007	0.1978	65.78%	0.3042	0.2184	71.79%
		PL2	0.2921	0.2028	69.43%	0.2976	0.2218	74.53%
	Query Expansion True	DLH	0.3530	0.2584	73.20%	0.3823	0.3015	78.86%
		PL2	0.3469	0.2566	73.97%	0.3606	0.2843	78.84%

Table 4. French and Italian – French Mean Average Precision (MAP)

Table 4 shows that POS analysis and NP extraction is associated with better retrieval performance, and appears to benefit monolingual retrieval more than it assists bilingual retrieval, as can be deduced by the fact that the difference between the monolingual and bilingual runs is higher when POS NP is used (29.58% on average), than when it is not (26.78% on average). This observation is indicative of the fact that even though light NLP can be of significant assistance to IR, it cannot counter the shortcomings of insufficient translation resources.

4 Conclusion

Our participation in the CLEF 2005 Ad-Hoc track for Bulgarian, English – Bulgarian, French, Italian – French and Greek – English retrieval was shown to be successful, with a difference from the Median Precision ranging between +1.135 (for Bulgarian) and +7.830 (for English – Greek). On a collective basis, poor or no language resources were at all times associated with consistently low retrieval performance. On an individual basis, lemmatisation was shown to be a satisfactory replacement of stemming for Greek, while noun phrase extraction was shown to benefit retrieval directly and consistently for French and Italian – French. We have shown that light morphosyntactic processing can assist IR for highly inflectional languages, and by doing so, we have carried our initial contention *a posse ad esse* successfully.

5 References

[1] Babelfish Machine Translation: http://babelfish.altavista.com/

[2] Lioma, C., He, B., Plachouras, V., and Ounis, I. "The University of Glasgow at CLEF 2004: French Monolingual Information Retrieval with Terrier". In *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, Volume Editors: C. Peters, P. D. Clough, G. F. J. Jones, J. Gonzalo, M. Kluck and B. Magnini, Lecture Notes in Computer Science, Springer-Verlag, 2005.

[3] Skycode Machine Translation: http://webtrance.skycode.com/online.asp

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

[5] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D. "Terrier Information Retrieval Platform". In *Proceedings of the 27th European Conference on Information Retrieval (ECIR 05)*, Santiago de Compostela, Spain, 2005. URL: http://ir.dcs.gla.ac.uk/terrier/

[6] TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[7] Worldlingo Machine Translation: http://www.worldlingo.com/

[8] Xerox Greek Language Analysis: http://www.xrce.xerox.com/competencies/content-analysis/demos/greek