SINAI at ImageCLEF 2009 wikipediaMM task

M.C. Díaz-Galiano, M.T. Martín-Valdivia, L.A. Ureña-López, J.M. Perea-Ortega
University of Jaén. Departamento de Informática
Grupo Sistemas Inteligentes de Acceso a la Información
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{mcdiaz,maite,laurena,jmperea}@ujaen.es

Abstract

This paper describes the first participation of the SINAI team in the CLEF 2009 wikipediaMM task. This year, we only want to establish a first contact with the task and the collections. Thus, we have generated a new collection expanding with WordNet terms in order to perform the information included in this collection. In addition, we have expanded de queries with WordNet too. We have used the LEMUR toolkit as the Information Retrieval system in our experiments.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Experimentation, Languages, Performance

Keywords

Query expansion, Document expansion, WordNet ontology, Information Retrieval

1 Introduction

This paper presents the first participation of the SINAI research group at the CLEF wikipediaMM task, although this is our fifth participation at the ImageCLEF campaign. We have experience in tasks that involve the retrieving of relevant images using the text associated to each image. In this field, we have obtained promising results combining some information retrieval systems and using several ontologies for the expansion of queries and the textual information of the collection [2, 3, 5].

Our main goal in this work is to study the use of the WordNet expansion technique over a collection and the queries. The integration of external resources to achieve a better information is a technique that has been used to improve the accuracy of the systems. For example, the integration of knowledge through the use of ontologies has been very successful in many systems. Specifically, WordNet[1] [9] has been used with success in many works related to information retrieval [11], image retrieval [1, 4], disambiguation [7, 6] and text categorization [8, 10].

The following section describes the collection and queries expansion using WordNet. In Section 3, we explain the experiments and obtained results. Finally, conclusions are presented in Section 4.

[1] http://wordnet.princeton.edu

2 Expansion with WordNet

WordNet is a large lexical database of English, developed in Princeton University. It groups nouns, verbs, adjectives and adverbs into sets of synonyms called *synsets* and records various semantic relations between these synsets. In our experiments, we have expanded collection and topic set with WordNet ontology using only the words includes in all the synset of each word. Next steps have been followed to expand the text:

- The text are tagged with a POS (*part of speech*) tagger.
- We obtains synset of nouns and verbs using WordNet ontology.
- The words of all synsets are included in a bag of words. In this step we add only the word, without POS or synset information.
- Repeated words in the bag have been removed.
- All words in the bag are included at the end of the original text.

3 Experiment Description and Results

The purpose of these experiments was to compare the performance of collection expansion when the query is expanded in the same way. We have used two textual collections and two sets of topics. The collections are:

- **NT**: It contains the text of *narrative* and *title* labels from original collection.
- **NTWn**: Constains narrative and title labels expanded with *WordNet*.

The sets of topics use only the title label of original topics to investigate if the expansion of a short text topic improves the system performance. The name of the topic sets are:

- **T**: Its contains the title label of original topics.
- **TWn**: The title label of original topics expanded with *WordNet*

The name of the experiment includes the name of the group, the name of the collection and the name of the topic set.

- **sinai_NT_T**: Collection include narrative and title and topics include only title.
- **sinai_NTWn_T**: Collection include narrative and title expanded with WordNet and topics include only title.
- **sinai_NT_TWn**: Collection include narrative and title and topics include title expanded with WordNet.
- **sinai_NTWn_TWn**: Collection include narrative and title expanded with WordNet and topics include title expanded with WordNet.

The dataset of the collection has been indexed using Lemur[2] IR system, by applying KL-divergence weighing function and using Pseudo-Relevance Feedback (PRF). Table 1 shows the main average precision (MAP) of our runs.

As we can see in Table 1 the query expansion does not improve the results although using only a expansion for the collection slightly achieve better MAP.

The expasion algorithm has several problems. The query expansion includes all synsets and add noise to the query. For example, the query number 102 is *'build site'* and the word *site* has

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

Experiment	Collection	Topic	MAP
sinai_NT_T	NT	T	0.1538
sinai_NTWn_T	NTWn	T	**0.1566**
sinai_NT_TWn	NT	TWn	0.1275
sinai_NTWn_TWn	NTWn	TWn	0.0998

Table 1: MAP values of SINAI experiments

synsets with the multi-words *'internet site'* and *'web site'*. These multi-words are included in the expansion of the query but its have not relation with the query. Other problem with the expansion methodology is the eliminition of repeated words. The multi-words *'internet site'* and *'web site'* contain the word *site*. Therefore, the expansion includes the word *site* several time, because the algorithm does not compare each word in the multi-word. The algorithm handles each multi-word like a simple word.

4 Conclusions

In this paper we describe our first participation at ImageCLEFwiki. We have experimented with different kind of expansion using the external resource WordNet. However, the obtained results show that it is necessary to continue investigating the expansion methodology.

Thus, our next goal will be to improve the expansion by applying some more techniques. For example, it will be interesting to prove a word disambiguation procedure before to incorporate the synset. In addition, further filtering of words included in multiword expressions could achieve better results.

Acknowledgements

This work has been supported by the Regional Government of Andalucía (Spain) under excellence project GeOasis (P08-41999), the Spanish Government under project Text-Mess TIMOM (TIN2006-15265-C06-03) and the local project RFC/PP2008/UJA-08- 16-14.

References

[1] Miyoung Cho, Chang Choi, and PanKoo Kim. Image Retrieval and Classification Through Conceptualization Based on WordNet. In De-Shuang Huang, Laurent Heutte, and Marco Loog, editors, *ICIC (3)*, volume 2 of *Communications in Computer and Information Science*, pages 751–759. Springer, 2007.

[2] M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia, A. Montejo-Ráez, and L.A. Ureña-López. Integrating MeSH Ontology to Improve Medical Information Retrieval. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *CLEF*, volume 5152 of *Lecture Notes in Computer Science*, pages 601–606. Springer, 2008.

[3] M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia, L.A. Ureña-López, and A. Montejo-Ráez. Query Expansion on Medical Image Retrieval: MeSH vs. UMLS. In Carol Peters et al., editor, *CLEF 2008*, volume 5706 of *Lecture Notes in Computer Science*, pages 732–735. Springer, 2009.

[4] M.C. Díaz-Galiano, J.M. Perea-Ortega, M.T. Martín-Valdivia, A. Montejo-Ráez, and L.A. Ure na López. SINAI at TRECVID 2007. In Paul Over, editor, *Proceedings of TRECVid 2007*, 2007.

[5] M.A. García-Cumbreras, M.C. Díaz-Galiano, M.T. Martín-Valdivia, and L.A. Ureña-López. SINAI at ImageCLEFPhoto 2008. In Carol Peters, editor, *Proceedings of CLEF 2008*, 2008.

[6] M. García-Vega, M.A. García-Cumbreras, M.T. Martín-Valdivia, and L.A. Ure na López. The University of Jaén Word Sense Disambiguation System. In *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, July 2004*. Association for Computational Linguistics (ACL), 2004.

[7] Francisco Joao Pinto, Antonio Farina Martinez, and Carme Fernandez Perez-Sanjulian. Joining automatic query expansion based on thesaurus and word sense disambiguation using WordNet. *IJCAT*, 33(4):271–279, 2008.

[8] M.T. Martín-Valdivia, L.A. Ureña López, and M. García Vega. The learning vector quantization algorithm applied to automatic text classification tasks. *Neural Networks*, 20(6):748–756, 2007.

[9] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database. *Journal of Lexicography*, 3(4):235–244, 1990.

[10] J.M. Perea-Ortega, M.T. Martín-Valdivia, A. Montejo-Ráez, and M.C. Díaz-Galiano. Categorización de textos biomédicos usando UMLS. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 40:121–127, 2008.

[11] Ray Richardson and Alan F. Smeaton. Using WordNet in a Knowledge-Based Approach to Information Retrieval. In *Proc. of the BCS-IRSG Colloquium*, number CA-0395, 1995.