MIRACLE at NTCIR-7 MOAT: First Experiments on Multilingual Opinion Analysis

Julio Villena-Román[1,2] Sara Lana-Serrano[1,3], José C. González-Cristóbal[1,3]
[1]DAEDALUS, [2]Universidad Carlos III de Madrid, [3]Universidad Politécnica de Madrid
jvillena@daedalus.es, slana@diatel.upm.es, jgonzalez@dit.upm.es

Abstract

This paper describes the participation of MIRACLE research consortium at NTCIR-7 Multilingual Opinion Analysis Task, our first attempt on sentiment analysis and second on East Asian languages. We took part in the main mandatory opinionated sentence judgment subtask (to decide whether each sentence expresses an opinion or not) and the optional relevance and polarity judgment subtasks (to decide whether a given sentence is relevant to the given topic and also the polarity of the expressed opinion). Our approach combines a semantic language-dependent tagging of the terms of the sentence and the topic and three different ad-hoc classifiers that provide the specific annotation for each subtask, run in cascade. These models have been trained with the corpus provided in NTCIR-6 Opinion Analysis pilot task.

Keywords: *NTCIR, MIRACLE, Multilingual Opinion Analysis Task, semantic tagging, statistical approach.*

1 Introduction

MIRACLE team is a research consortium formed by research groups of three different universities in Madrid (UPM[1], UAM[2] and UC3M[3]) along with DAEDALUS[4], a small/medium size enterprise founded in 1998 as a spin-off of two of these groups and a leading company in the field of linguistic technologies in Spain. MIRACLE was born specifically to participate in CLEF (Cross Language Evaluation Forum) [1], the European homologue of NTCIR, in which we have taken part since 2003 and submitted experiments for all tasks, including bilingual, monolingual and cross lingual retrieval tasks, image, video, web and geographic information retrieval, question answering and interactive task.

This paper describes our participation at the Multilingual Opinion Analysis Task (MOAT) at NTCIR-7 [2]. MOAT poses different and interesting challenges in two research aspects: sentiment analysis (semantic opinion-related tagging and machine learning applied to natural language classification) and also work with East Asian languages.

On one hand, there exist different techniques focused on the sentiment analysis problem. Some of them are based on extracting the subjectivity of a given text by means of the application of semantic tagging and then performing a probabilistic classification with the (naïve) Bayes algorithm or similar. Other methods try to build supervised classification models based on different lexical and syntactic features that are present in the text.

On the other hand, East Asian languages have some factors and differential characteristics with respect to European languages which make them very appealing: a complex writing system made up of a mixture of scripts, a morphological structure which makes it hard to perform an accurate segmentation and conflation, lack of a standard orthography and/or the presence of numerous orthographic variants which force the use of cross-orthographic searching, and other miscellaneous technical requirements such as transcoding between multiple character sets and encodings and support for Unicode and input method editors [3].

The main idea behind our approach to MOAT is to research on statistical sentiment classification techniques based on semantic features extracted from the text. These techniques are in nature independent of the lexical and syntactical knowledge about the specific language. Our objective is to compare their performance when applied to corpus in both European and East Asian languages and draw some conclusions about the results.

We finally submitted runs for monolingual English and Japanese in (1) the main mandatory opinionated sentence judgment subtask that consists in deciding whether each sentence expresses an opinion or not, (2) the optional relevance sentence judgement subtask whose objective is to decide whether the sentences are relevant to the given topic or not, and also (3) the optional polarity judgment subtask that tries to establish the polarity (positive, neutral or negative) of the opinionated sentences.

In the following sections, we describe our approach and the system that was developed to carry out the experiments, comment about the evaluation results and, finally, present some conclusions and possible future lines of work.

2 System Description

Based on our previous experience on similar task-oriented evaluation campaigns, we designed a modular system composed of a set of small components that are easily combined in different setups and executed sequentially to produce the final results. This architecture offers the greatest flexibility to carry out a variety of experiments with different setups without much effort.

Figure 1 shows the logical architecture of the system, which is composed of three different functional modules.

[1] Universidad Politécnica de Madrid <www.upm.es>
[2] Universidad Autónoma de Madrid <www.uam.es>
[3] Universidad Carlos III de Madrid <www.uc3m.es>
[4] DAEDALUS <www.daedalus.es>

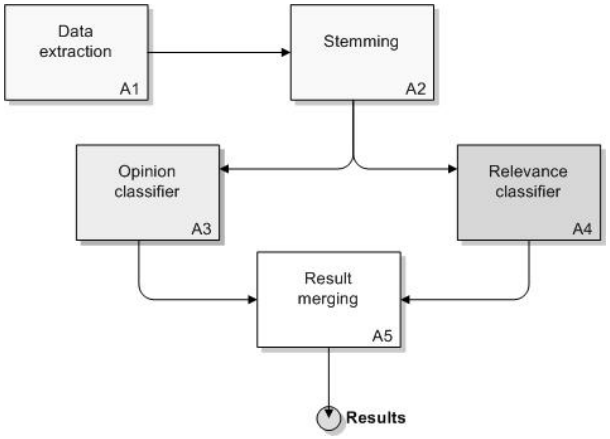


Figure 1: Overview of the system.

The first module (A1, A2) performs the preprocessing and linguistic analysis of each sentence of the text. It is in turn composed of two blocks. The first block extracts the information associated to each document that is pertinent to each topic: identifier, title and content of each sentence. The second block filters out the stopwords and performs the stemming of the resulting words. Freeling [4] and Mecab [5] have been used for stemming English and Japanese texts, respectively.

The second module (A3, A4) is the core of the system and contains a supervised classification model and algorithm associated to each of the previously described subtasks. The corpus of the NTCIR-6 Opinion Analysis pilot task [6] has been used for training each model.

The third module (A5) is in charge of collecting the outputs of each classification block, and, based on these values, determining the most adequate final global value.

As shown in the system architecture, we have wrongly considered that a given sentence would be judged as relevant or not-relevant independently of the fact that it was opinionated or non-opinionated, the same criteria as in NTCIR-6 Opinion Analysis pilot task, without noticing that the specification for the present task stated that only opinionated sentences had to be annotated for relevance.

2.1 Opinion classifier

The opinion classification subsystem is in charge of deciding whether a sentence expresses an opinion or not, and, for those sentences that in fact express an opinion, annotating their polarity, according to three degrees: positive, neutral or negative.

The approach to annotate the global polarity of a given sentence is based on a weighted semantic addition of the individual polarities that have been identified for each fragment (opinion unit) that is part of the whole sentence (the average value). For this purpose, a Semantic Knowledge Base (SKB) has been developed.

The SKB contains a set of terms (specifically, nouns, adjectives and verbs) annotated with their semantic orientation (positive/negative polarity and normal/strong degree). These resources are specific for the domain and language of each subtask, in this case, news articles and English and Japanese language.

The SKB has been obtained from the linguistic resources provided by General Inquirer [7], originally in English. For the English side, 4,543 terms were directly extracted from the "Positive", "Negative", "Strong" and "Weak" categories of Harvard IV-4 dictionary (included in the General Enquirer). For Japanese, English terms have been machine-translated using JMDict [8]. This dictionary returns a set of fields, but only two of them are selected: "keb" (words or short phrases in Japanese that are written using at least one kanji, optionally with kana characters) and "reb" (content restricted to kana). In both cases, the valid characters are kanji, kana, related characters such as chouon and kurikaeshi, and in exceptional cases, letters from other alphabets. The final knowledge base for Japanese contains a set of over 55,000 terms.

This major difference in number of terms with respect to English may indicate a wrong selection of Japanese semantic terms that may be a possible explanation for the difference in the performance for each language. Table 1 shows the term distribution for each semantic category and language.

Table 1: Semantic Knowledge Base (SKB).

Category	Description	Tag	English terms	Japanese terms
Positive	Positive term	P	1,158	13,111
Negative	Negative term	N	1,289	5,816
Positive Strong	Positive terms implying strength	P+	443	358
Positive Weak	Positive terms implying weakness	P-	17	13,005
Negative Strong	Negative terms implying strength	N+	273	3,839
Negative Weak	Negative terms implying weakness	N-	419	4,878

Figure 2 shows the architecture of the opinion classifier. First, the text content of each opinion unit that forms each sentence is parsed as explained before (stemming and stopword removal), and then the resulting terms are semantically tagged using the SKB. This semantic tagging was used for topic expansion in medical image retrieval with very good results [9].

Then, as defined in the Vector Space Model, a weighted vector is built for each opinion unit, representing the term frequency of the main opinionated features (i.e., terms in the SKB) in the given sentence.

Finally, two lazy classifiers (based on the k-Nearest Neighbour algorithm) are built using the feature matrix composed of the vectors of the sentences in the corpus provided in the NTCIR-6 Opinion Analysis Pilot task, one classifier for deciding whether the sentence is opinionated or not, and another classifier for annotating the opinion polarity. If a given sentence is positively classified as opinionated by the first classifier, then the other classifier is executed to extract the polarity of the opinion expressed in the sentence.

These feature vectors had been already applied to perform automatic annotation of medical images [10] [11] and also the combination of several classifiers in cascade for speech transcript topic detection and classification [12].

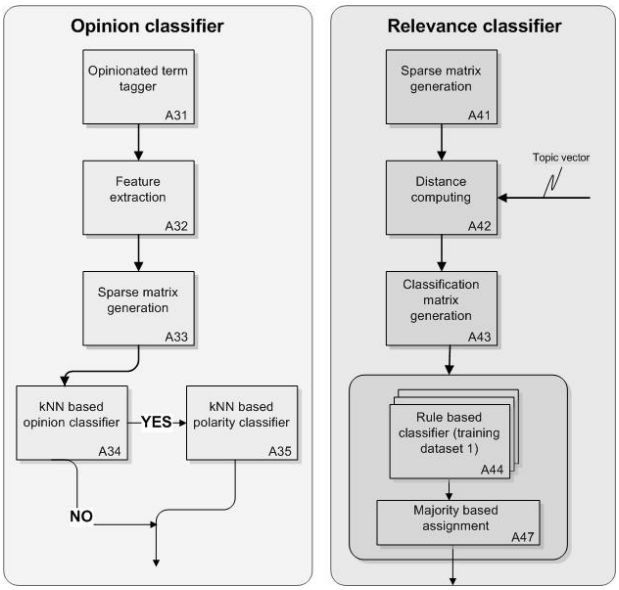


Figure 2: Classifier Architecture.

2.2 Relevance classifier

Our approach to solve the relevance judgment problem, i.e., to determine whether a sentence is relevant to the topic or not, is based on the semantic distance between the vectors that represent them. If the distance is lower than a given threshold, the sentence is positively classified as relevant (and, on the contrary, as non-relevant). A supervised classification model is used to find this optimum discrimination threshold for the semantic similarity between topic and the given fragment.

The contents of both each sentence and the title of topic are first stemmed and filtered as before. Then a new weighted vector is built for each sentence and topic, representing the term frequency of the main most significant terms. After this, the distance (the degree of semantic similarity) between both vectors is calculated. Several metrics were studied, but finally the cosine distance was selected as slightly better results were achieved. As a result of this process, a new vector is built, having as features the value of this distance between the sentence and the topic and the ratio between the number of terms in the sentence and the number of terms in the topic. Finally this feature vector is processed by a rule-based classifier, trained with the NTCIR-6 corpus, to decide whether the sentence is relevant or not with respect to the given topic. The classifier is based on a tree chart algorithm using those two features.

Table 2 shows an excerpt of the rule set obtained from the supervised training on the English NTCIR-6 corpus with the data corresponding to assessor 1. In the table, "normTerms" means the number of terms, normalized over the number of terms in the topic title, and "similarity" represents the degree of similarity between the sentence and the topic.

Table 2: Excerpt of the rule set for relevance classification.

```
normTerms <= 0,8333 [Mode: N] (support:707)
    similarity <= 0,0514 [Mode: N] => N (support:680; 0,938%)
    similarity > 0,0514 [Mode: N] => N (support:27; 0,704%)
normTerms > 0,8333 and normTerms <= 1,4000 [Mode: N] (support:791)
    similarity <= 0 [Mode: N] => N (support:714; 0,849%)
    similarity > 0 [Mode: N] => N (support:77; 0,61%)
normTerms > 1,400 and normTerms <= 2,2500 [Mode: N] (support:1.358)
    similarity <= 0,0514 [Mode: N] => N (support:1.186; 0,718%)
    similarity > 0,0514 [Mode: Y] => N (support:172; 0,5%)
normTerms > 2,2500 and normTerms <= 3,2500 [Mode: N] (support:1.553)
    similarity <= 0,0514 [Mode: N] => N (support:1.348; 0,589%)
    similarity > 0,0514 [Mode: Y] => Y (support:205; 0,537%)
normTerms > 3,2500 and normTerms <= 3,8333 [Mode: Y] (support:730)
    similarity <= 0,0514 [Mode: N] => N (support:623; 0,512%)
    similarity > 0,0514 [Mode: Y] => Y (support:107; 0,673%)
normTerms > 3,8333 and normTerms <= 5,7500 [Mode: Y] (support:1.421)
    similarity <= 0 [Mode: Y] => Y (support:1.100; 0,584%)
    similarity > 0 [Mode: Y] => Y (support:321; 0,698%)
normTerms > 5,7500 [Mode: Y] (support:812)
    similarity <= 0 [Mode: Y] => Y (support:643; 0,675%)
    similarity > 0 [Mode: Y] => Y (support:169; 0,817%)
```

3 Evaluation

Despite several implementation problems (mainly issues related to character encoding conversion and Unicode management, and a slow execution speed of the classifiers), we were finally able to submit one run for each language. Tables 3 and 4 show the official results provided by the task organizers for our runs for English and Japanese, respectively.

Results for polarity judgement in English have been omitted because values were poor and thus not significant. We think that this result was due to a bug in the implementation of the polarity classifier. However, this issue must be further investigated.

Table 3: Results for English language.

		Precision	Recall	F-Measure
Opinionated	Lenient	0.595	0.012	0.023
	Strict	0.286	0.012	0.022
Relevance	Lenient	0.374	0.319	0.344
	Strict	0.085	0.304	0.133

Table 4: Results for Japanese language.

		Precision	Recall	F-Measure
Opinionated	Lenient	0.316	0.090	0.140
	Strict	0.241	0.094	0.135
Relevance	Lenient	0.455	0.082	0.138
	Strict	0.223	0.082	0.120
Polarity	Lenient	0.247	0.018	0.034
	Strict	0.240	0.017	0.031

As previously told, the comparison of the results for the relevance judgment task with respect to other participants is not possible, as we have wrongly considered all sentences in the corpus instead of only those sentences that had been previously classified as opinionated. However, a script was provided by the organizers of the task in Japanese to perform the evaluation considering all sentences in the corpus for the relevance calculation. Table 5 shows this evaluation, which, as expected, achieves higher values for recall and thus for F-Measure.

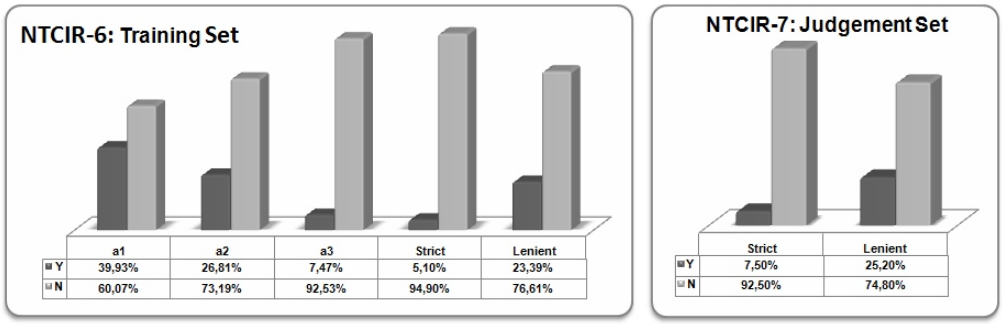
Table 5: Evaluation of relevance, for Japanese, considering all sentences.

		Precision	Recall	F-Measure
Relevance	Latent	0.136	0.864	0.234

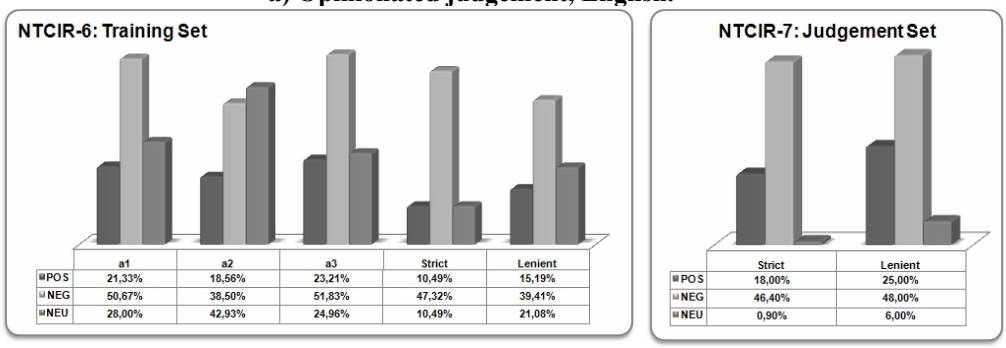
Table 6 shows the confusion matrixes for the official NTCIR-7 MOAT results and our own runs and allows to compare the class distribution for each judgment subtask between the official results and our own results.

Table 6: Confusion matrixes between NTCIR-7 MOAT test collection and MIRACLE runs.

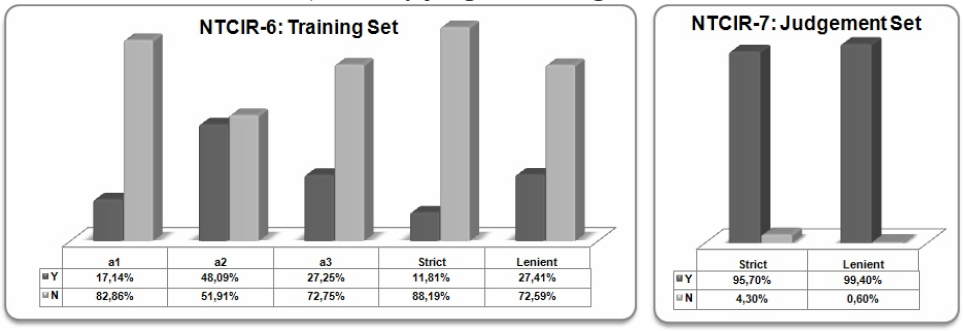
		Opinionated		Polarity			Relevance	
		Y	N	POS	NEG	NEU	Y	N
English	**MIRACLE Run**	0.97%	99.03%	0.00%	16.67%	83.33%	27.03%	72.97%
	Latent NTCIR-7	25.20%	74.80%	25.00%	48.00%	6.00%	99.40%	0.60%
Japanese	**MIRACLE Run**	7.76%	92.24%	10.69%	68.76%	20.55%	78.40%	21.60%
	Latent NTCIR-7	28.90%	71.10%	5.50%	15.30%	79.20%	43.20%	56.80%



a) Opinionated judgement, English.



b) Polarity judgement, English.



c) Relevance judgement, English.

Figure 3: Comparison of training (NTCIR-6) and test (NTCIR-7) data for different subtasks, English.

It can be easily observed that our runs generate a high percentage of "No" classes for both opinionated and relevance judgments, i.e., tend to answer "Non-opinionated" and "Non-relevant" for the majority of the sentences. This fact is derived from the unbalanced training corpus, as shown in the uneven distribution of each class in the training and test datasets for those subtasks for English language, depicted in Figure 3 (a, b and c). This is especially noticeable in the case of relevance judgment: in the training data, only 27.41% of the examples belong to the "Yes" class, whereas in the judgment set, 99.40% of the instances are "Yes". In other words, "No" classes are predominant in the training data, so the classifiers are unable to get enough knowledge to model "Yes" classes.

Regarding the polarity, although the distribution of the different classes (POS, NEU and NEG) is not even, it could be reasonable expected that the model showed a clear tendency towards "NEG" class (the predominant one). However, the most frequent class in our runs is "NEU". This result is due to the fact that the default rule of the polarity classifier, if faced with ambiguous situations or not present in the knowledge base (the algorithm is the memory-based kNN), was to return a neutral polarity. This was another bug in our implementation, due to the fact that the Japanese language run was developed first and, in this case, the predominant class is "NEU".

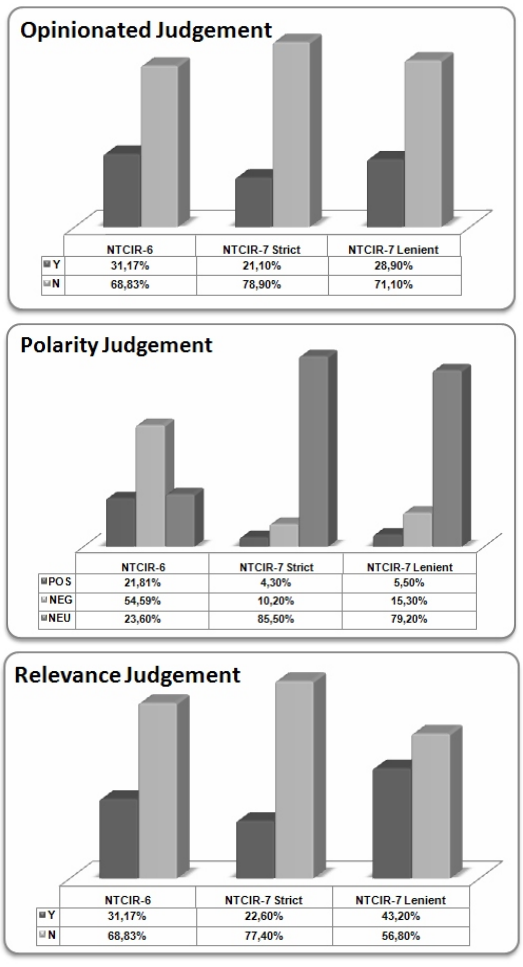


Figure 4: Comparison of training (NTCIR-6) and test (NTCIR-7) data, Japanese

We also observed some problems due to the lack of a disambiguation process for a considerable number of semantically-tagged terms. For example, the English verb "to increase" was tagged both "Positive" (increase happiness) and "Negative" (increase unemployment), which may mean added noise to the classification step.

Results of the opinionated judgment for English language are better than those for Japanese. This can be explained because the Semantic Knowledge Base for Japanese contains a remarkably higher number of terms, and thus more terms are semantically tagged. The reason is that the translation process from English adds additional terms (synonyms or related words) apart from the literal word-by-word translation, that inherit the same semantic tags (Positive, Negative, Strong and/or Weak) as the original English term, which introduce ambiguity (noise) in the classification step.

Figure 4 shows the same comparison of training and test corpus for each subtask, for Japanese. Class distribution is also uneven, as in English.

4 Conclusions and Future Work

As this was our first participation in a task focused on sentiment analysis, our main effort was mainly dedicated to study and learn the basics of the techniques and the distinctive semantic features of the languages involved and also to build the entire necessary linguistic infrastructure to be able to submit our experiments in time.

The results of our runs are disappointing as compared to the results provided by the task organizers about the other participants. However, starting from scratch is always difficult, so there are many aspects that we could not address in our runs this year, due to evident limitations of computing resources, time and expertise.

We are interested in going on with this line of research and continue our participation in future editions of MOAT.

For this purpose, more effort has to be invested in improving the preprocessing step (better stemmers and stopword lists) and also the Semantic Knowledge Base used for tagging, specifically resources for Japanese. The translation from English to Japanese produces many Japanese terms that are semantically-related to the original English term, and this fact seems to cause trouble in the classification process. Thus, a filtering process should be considered.

In addition, there is obviously a large space for improvement in the classification modules. Classifiers must be completely redesigned, improving the training dataset and taking into account the uneven distribution of the output classes. Also, some semantic disambiguation techniques could also be applied, perhaps using similar techniques as for automatic topic semantic expansion [13] in a general-purpose information retrieval context.

Acknowledgments

This work has been partially supported by the Spanish R+D National Plan, by means of the project BRAVO (Multilingual and Multimodal Answers Advanced Search – Information Retrieval), TIN2007-67407-C03-

03 and by Madrid R+D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

References

[1] Cross Language Evaluation Forum (CLEF). <www.clef-campaign.org>

[2] Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen and Noriko Kando. *Overview of Multilingual Opinion Analysis Task at NTCIR-7*, National Institute of Informatics, October 24, 2008.

[3] Halpern, Jack. *The Challenges of Intelligent Japanese Searching*. The CJK Dictionary Institute, Inc. <http://www.cjk.org/cjk/joa/joapaper.htm>.

[4] FreeLing. <http://garraf.epsevg.upc.es/freeling>

[5] Mecab. *Yet Another Part-of-Speech and Morphological Analyzer*. <http://chasen.org/~taku/software/mecab>

[6]Y. Seki, D. K. Evans, L. W. Ku, H. H. Chen, N. Kando, and C. Y. Lin. *Overview of Opinion Analysis Pilot Task at NTCIR-6*. In Proc. of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross- Lingual Information Access, pages 265–278, NII, Japan, May 2007.

[7] General Inquirer. <http://www.wjh.harvard.edu/~inquirer>

[8] JMDict: Japanese Multi-lingual Dictionary. <www.csse.monash.edu.au/~jwb/jmdict.html>

[9] S. Lana-Serrano, J. Villena-Román and J.C. González-Cristóbal. *MIRACLE at ImageCLEFmed 2008: Evaluating Strategies for Automatic Topic Expansion*. Working Notes of the 2008 CLEF Workshop. Aarhus, Denmark, September 2008.

[10]S. Lana-Serrano, J. Villena-Román, J.C. González-Cristóbal and J.M. Goñi-Menoyo. *MIRACLE at ImageCLEFannot 2008: Classification of Image Features for Medical Image Annotation*. Working Notes of the 2008 CLEF Workshop. Aarhus, Denmark, September 2008.

[11]S. Lana-Serrano, J. Villena-Román, J.C. González-Cristóbal and J.M. Goñi-Menoyo. *MIRACLE at ImageCLEFannot 2007: Machine Learning Experiments on Medical Image Annotation*. Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers. Carol Peters et al (Eds.). Lecture Notes in Computer Science, ISBN: 978-3-540-85759-4. Vol. 5152: 597-600, 2008.

[12]J. Villena-Román and S. Lana-Serrano. *MIRACLE at VideoCLEF 2008: Classification of Multilingual Speech Transcripts*. Working Notes of the 2008 CLEF Workshop. Aarhus, Denmark, September 2008.

[13]J.L. Martínez-Fernández, A.M. García-Serrano, J. Villena-Román and P. Martínez. *Expanding Queries Through Word Sense Disambiguation*. Evaluation of Multilingual and Multi-modal Information Retrieval 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers. Carol Peters et al (Eds.). Lecture Notes in Computer Science, ISBN: 978-3-540-74998-1.Vol. 4730: 613-616, 2007.