FATS: A Framework for Annotation of Travel Blogs Based on Subjectivity

Inmaculada Álvarez de Mon y Rego
Ingenieria Técnica de Telecomunicación. UPM
Lingüistica aplicada a la ciencia y la tecnología
Madrid, Spain
ialvarez@euitt.upm.es

Liliana Ibeth Barbosa Santillán
University of Guadalajara
Tecnologías de la Información
Guadalajara, México
ibarbosa@cucea.udg.mx

***Abstract*—This paper describes a framework for annotation on travel blogs based on subjectivity (FATS). The framework has the capability to auto-annotate -sentence by sentence- sections from blogs (posts) about travelling in the Spanish language. FATS is used in this experiment to annotate components from travel blogs in order to create a corpus of 300 annotated posts. Each subjective element in a sentence is annotated as positive or negative as appropriate. Currently correct annotations add up to about 95 per cent in our subset of the travel domain. By means of an iterative process of annotation we can create a subjectively annotated domain specific corpus.**

Keywords-Annotation; Subjectivity; Blogosphera; Spanish Language.

I. INTRODUCTION

The Social Web [1] has enabled humans to express and share their opinions and concerns to the World. One of the tools to share data and information within the net is a Blog. According to the WordPress [2] online dictionary, *a blog, or weblog,* is an online journal, diary, or serial published by a person or group of people.

Currently, there are several platforms such as Blogger APPis of Google [3] or APPis of MyBlogLog of Yahoo [4] that offer blogs classified according to several taxonomies including classes such as personal, business, non-profits, politics, etc [2]. Travel blogs belong to the personal class.

In this work, the authors take into consideration the following premises:

- According to [2] the type of blog, those studied belong to the personal class.
- The Blogs of study belong to the travel domain in Spanish language.
- The writing process used by bloggers to express their ideas depends on age, context, time, culture and geographic place.
- Despite the availability of good practices on the web to preserve the quality of writing, for instance FS250062 [5], most bloggers express their opinions in very heterogeneous ways.
- The blogs have several components including title, banner, tagboard, links, archives and posts.
- The posts of each blog keep their chronological order.
- Some of the elements of posts are positive or negative or both.

Considering all this, the scientific hypothesis of our research is to auto-annotate the bloggers' expressions based on their subjectivity with at least 90% recall and precision for posts. The methodology used in this research is top-down in contrast to the Folksonomies methodology [6] where the main aim is to annotate collaboratively the social web.

This research deals with corpora in two dimensions: linguistic and technical. The linguistic dimension refers to the selection of sentences and their elements of the posts. The technical dimension encodes, builds the meta-model for annotation and annotates the posts.

Linguistic dimension:

- The sentence selection is based on finding blogs with two main components: 1) the title, and 2) the first post.

Technical dimension:

- The encoding is based on the standard ANSI/NISO Z39.19-2005 [7] to represent parts of a sentence and, in addition, to add the mark P for positive and N for negative.
- Annotation is based on bracketing conventions of segments according to the recommendations for the morphosyntactic annotation of corpora in tagging from lexical data (EAGLES96) [8].
- The meta-model for annotation is represented by eight patterns proposed in this research (see Section IV).

In order to analyze the bloggers' expressions, a subset of an ad-hoc corpus collected by [9] is used. This corpus consists of 10 thousand words extracted from Spanish blogs, sampled from a comprehensive range of travel blogs.

The aim of this work is to automatically recognize the positive or negative elements of the posts and annotate them.

The remainder of the paper is organized as follows. Section II briefly discusses the related work. In Sections III and IV, the detailed functionalities of FATS are presented. Section V briefly evaluates the performance of the framework proposed. Finally, Section VI concludes this paper.

II. RELATED WORK

The related work dealt with takes into account three topics: subjectivity, lexicons and annotation.

 ISBN: 978-1-61208-162-5

A. Subjectivity

According to Wiebe [10], subjectivity is "the linguistic expression of somebody's opinions, sentiments, emotions, evaluations, beliefs, and speculations".

Research work on subjectivity includes: Kanayama and Nasukawa [11] who built some domain-dependent polarity lexicons for Japanese language; Andreevskaia and Bergler [12] proposed various methods for learning subjectivity from WordNet. Esuli and Sebastiani [13] proposed a method for identifying both, the subjectivity and prior polarity of a word, also using WordNet. Wiebe and Mihalcea [14] were able to automatically identify whether a particular word was subjective or not, using a computer program. Kim and Hovy [15] used small seed sets and WordNet to identify sets of subjective adjectives and verbs, and Kobayashi et al. [16] identified domain-dependent sets of subjective expressions. This work is based on the subjectivity classification for the previous research work.

B. Lexicons

According to [17], a general definition of a lexicon is "the vocabulary of a language that contains all the words or LEXEMES in the language". A more specific one suitable for our domain dependent research is "Word stores that are primarily consulted for the reason of information retrieval are referred to as "dictionaries". By contrast, word-stores that constitute a component within a natural language processing system are called a lexicon [18], the LEXICON is understood broadly as a finite list of stored forms and the possibilities for combining them".

Previous research has focused on the creation of lexicons in English such as: Higashinaka [19] who used a set of dialogues to build her own lexicon. Lexicons are also available as linguistic resources in the Internet, some examples are SentiWordnet [20], NTU Sentiment Dictionary [21], Opinion Finders subjectivity Lexicon [22], etc. However our research relies on a controlled vocabulary that tries to eliminate noise by providing a list of preferred and non-preferred terms and a domain-semantic structure in Spanish. Therefore, the lexicon built by [9] was taken as reference. The process for creating this lexicon was the following: a) key terms used in valorative sentences are extracted and b) words or groups of words with positive or negative sentiment are selected for the lexicon. The final classes were nouns, adjectives, diminutives, prefixes, verbs, adverbs, interjections, and idioms taken from the language sample.

C. Annotation

Although there are many Spanish Corpora involving grammatical analysis or annotation of Spanish texts (e.g., Atwell [23]) and a significant number of software libraries developed at universities to annotate texts (Exmaralda [24] and MMax 2 [25] tools). However, no research has been found on annotating each component of a post -as positive or negative- as appropriate. The closest results were found to be by [23], [24], [25] and their results evaluated only part of speech tagging. Our proposal is based on the subjectively annotation of posts of blogs supported by a reference-subjectivity lexicon.

The Corpus of blogs used for this research consists of 10 thousand words of Spanish written blogs, sampled from a comprehensive range of blog within the travel domain. For this research annotation is the process of attaching subjectivity information to the posts which can be opinions, evaluations, emotions and beliefs. It consists of two main steps: identifying elements on the post, and attaching polarity information to these elements.

The process for annotation in our research is aligned with the automatic annotation of three linguistic levels: morphosyntactic, syntactic and semantic.

- At the morphosyntactic level, a word is divided into its root and suffixes.
- At the syntactic level, the lexico-syntactic representation of nouns, adjectives, verbs and adverbs (positive or negative or both) -all of them based on the lexicon proposed by [9]- are mapped according to the eight patterns proposed in Section IV.
- At the semantic level, the eight patterns (see Section IV) link the relationships between each word in each sentence of the posts.

III. Framework of FATS

The general architecture of FATS consists in a series of components performing sequential transformations on an input blog. The architecture is structured into four layers: Data Source, Matching Components, Analysis Components and Result Components, as shown in Fig. 1.

- *Data Source:* this layer contains the basic elements of blogs (see Fig. 2) required by the FATS.
- *Matching Components*: this second layer contains the main matching engineering functionality carried out through the analysis of components. Additionally, other components such as selector, split up, match, patterns and blogger-opinion are shown.
- *Analysis Components:* this middle layer contains the main linguistic levels of analysis in the blog: morphosyntactic, syntactic and semantic as explained in the introduction section.
- *Result Components*: the last layer contains the posts subjectively annotated of the blogs.

The relationships among the different components of the framework of FATS are explained below:

First of all, blogs are collected in digitized documents from the WWW throughout Heritrix [26] as an example the authors made a job limited by scope and frontier. Next, they are taken to the matching components where the blogs are structured and modelled on separated components such as

 ISBN: 978-1-61208-162-5

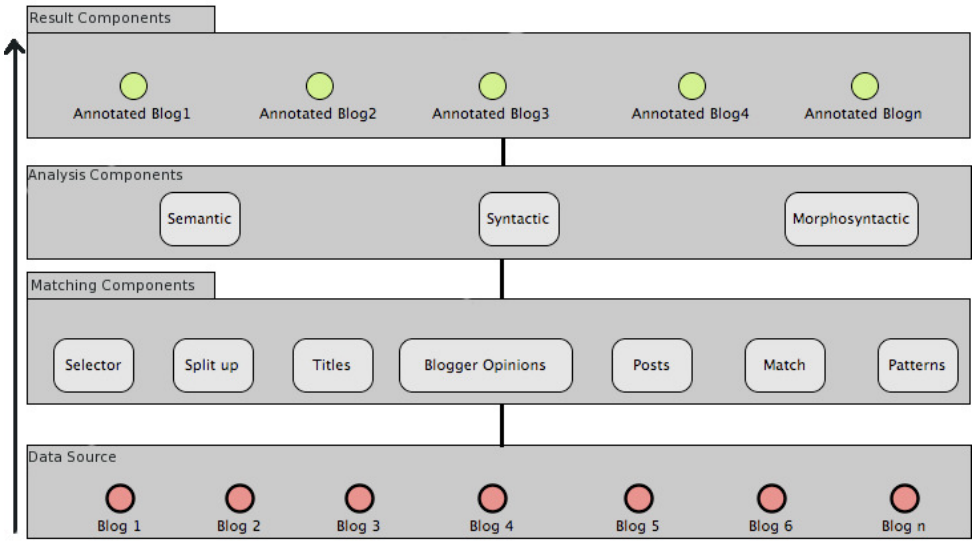

Figure 1. The general architecture of FATS.

titles and posts. Then, the selector component takes the first post and matches it with the related patterns (see Fig. 3); we do not match with the titles yet. In this stage, the analysis is conducted by the syntactic, morphosyntactic and semantic components. The syntactic analysis is done by mapping onto the lexico-syntactic representation of nouns, adjectives, verbs and adverbs (positive or negative or both) all of them based on the lexicon proposed by [9]. In the next stage, the morphological component processes a word into its smallest meaningful components, or morphemes. This is done by dividing a word into its root and suffixes. Subsequently, the semantic component establishes the relationships between each word in each sentence. Finally, the posts are placed in bags, which in turn have the annotated posts of blogs that determine the polarity of subjective expressions of the terms included in the posts.

IV. PATTERNS

For this research, we created eight different patterns as shown in Fig. 3, based on the most common structure type. Each pattern describes how the experiment interacts with the proposed framework (see algorithm 1) to achieve a new section of annotated posts. The 8 specialized patterns are represented by the following equations:

$$\forall x, y.Sentence(x, y) \rightarrow Subject(x) \sqcap Predicate(y). \quad (1)$$

$$P1 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \sqcap Noun(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \sqcap Art(y) \\ \sqcap Noun(y) \sqcap Pp(y) \sqcap Art(y) \sqcap Noun(y) \end{cases} \quad (2)$$

$$P2 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \\ \sqcap Noun(x) \sqcap Adj(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \sqcap Art(y) \\ \sqcap Noun(y) \sqcap Adj(y) \sqcap Pp(y) \\ \sqcap Art(y) \sqcap Noun(y) \sqcap Adj(y) \end{cases} \quad (3)$$

$$P3 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \sqcap Noun(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \\ \sqcap Art(y) \sqcap Noun(y) \end{cases} \quad (4)$$

$$P4 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \sqcap Noun(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \\ \sqcap Pp(y) \sqcap Art(y) \sqcap Noun(y) \end{cases} \quad (5)$$

$$P5 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \sqcap Noun(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \\ \sqcap Pp(y) \sqcap Noun(y) \end{cases} \quad (6)$$

$$P6 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \sqcap Noun(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \sqcap Pp(y) \\ \sqcap Noun(y) \end{cases} \quad (7)$$

$$P7 \doteq \forall y.Predicate(y) \rightarrow \exists Verb(y) \sqcap Art(y) \sqcap Noun(y) \quad (8)$$

$$P8 \doteq \begin{cases} \forall y.Predicate(y) \rightarrow \exists Verb(y) \sqcap Art(y) \\ \sqcap Noun(y) \sqcap Pp(y) \sqcap Art(y) \sqcap N(y) \end{cases} \quad (9)$$

Some of the experiments are shown in Table 1 with 8 different subjective tagging schemes mentioned as follows:

 ISBN: 978-1-61208-162-5

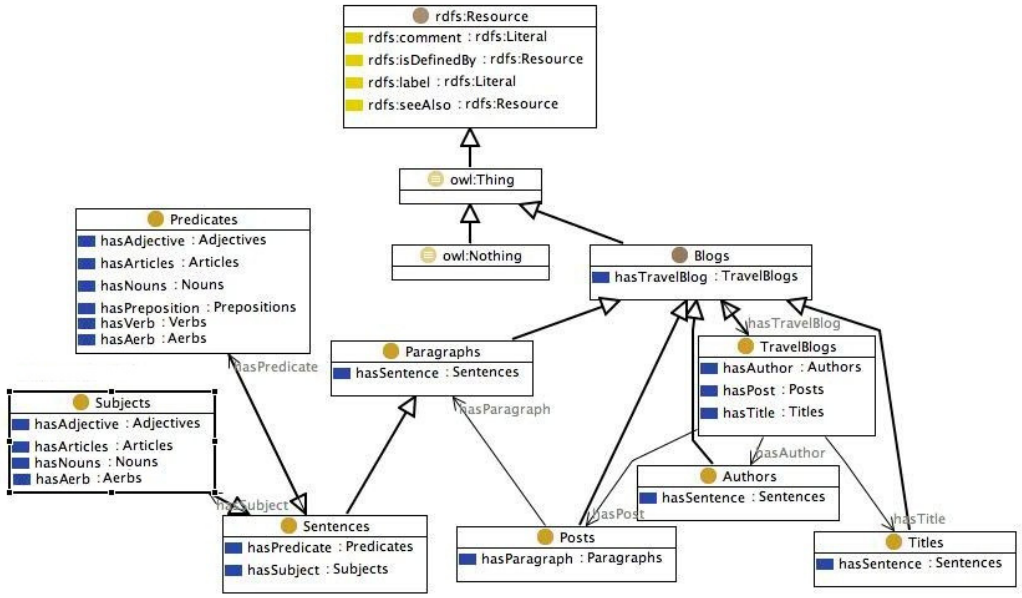

Figure 2. The structure of a blog

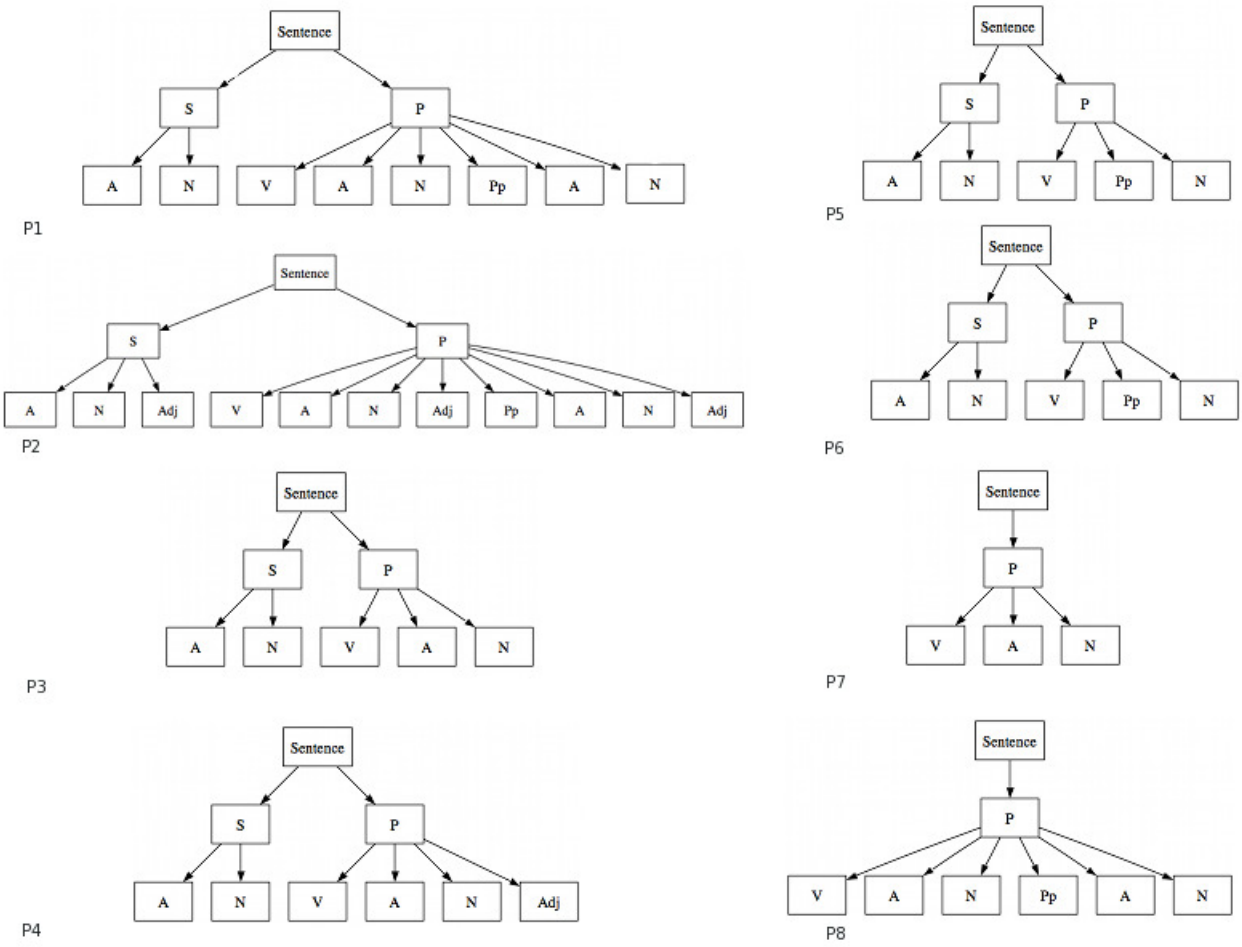

Figure 3. The 8 specialized patterns.

Table I
SOME OF THE EXPERIMENTS USING FATS

Blog	Post	Pattern	Annotated Post
1	Los dueños no-eran la alegría de la huerta. "The owners were not the joy of vegetable garden".	P1	s[nc[d[los], NP[dueños]], vc[VN[noeran], nc[d[la], NP[alegría], pp[p[de], nc[d[la], huerta]]]]],
2	El viaje perfecto continua sin incidentes malos, con un aprovechamiento máximo. "The perfect trip goes on without any bad incidents, with maximum benefit".	P2	s[nc[d[el], na[NP[viaje], AP[perfecto]]], vc[VP[continua], nc[d[sin], na[NP[incidentes], AN[malos]]], pp[p[con], nc[d[un], na[NP[aprovechamiento], AP[máximo]]]]]],
3	Los italianos son los abiertos. "Italians are open".	P3	s[nc[d[los], NP[italianos]], vc[VP[son], nc[d[los], NP[abiertos]]]],
4	El ambiente esta a la perfección. "The environment is perfect".	P4	s[nc[d[el], NP[ambiente]], vc[VP[esta], pp[p[a], nc[d[la], NP[perfección]]]]],
5	La verdad estabamos muy cansados. "Actually we were very tired."	P5	s[nc[d[la], NP[verdad]], vc[VP[estabamos], pp[p[muy], nc[NN[cansados]]]]],
6	Otra vez estamos super-cansados. "Again we are super tired."	P6	s[nc[d[otra], NP[vez]], vc[VP[estamos], pp[p[super], nc[NN[cansados]]]]],
7	Fuimos a Edimburgo. "We went to Edinburgh."	P7	s[vc[VP[fuimos], pp[p[a], nc[NP[edimburgo]]]]],
8	Tomamos un bus pero nos equivocamos. "We took a bus but we were wrong."	P8	s[vc[VP[tomamos], nc[d[un], NP[bus], pp[p[pero], nc[d[nos], NN[equivocamos]]]]]],

NP (noun positive), NN (noun negative), AP (adjective positive), AN (adjective negative), VP (verb positive), VN (verb negative), AdP (adverb positive), AdN (adverb negative). The part of speech tagging schemes are: s (subject), nc (noun complement), na (noun adjective), d (article), vc (verb complement), pp and p (preposition). Also, Table 1 presents an example of a post annotated using FATS. In this case the annotation scheme [8] was used to tag the posts elements. The bracketing of each element in the sentence involves the delimitation with square brackets.

V. PERFORMANCE RESULTS

In order to evaluate the quality of the structural markup, we calculated the recall and precision rates produced by FATS, as shown in formulas (10) and (11), where RW means Retrieved Words.

$$precision = \frac{|\{\text{relevant}(NP \cup AP \cup AdP \cup VP)\} \cap \{\text{RW}\}|}{|\{\text{RW}\}|} \quad (10)$$

$$recall = \frac{|\{\text{relevant}(NP \cup AP \cup AdP \cup VP)\} \cap \{\text{RW}\}|}{|\{\text{relevant (NP} \cup AP \cup AdP \cup VP)\}|} \quad (11)$$

The first task is to analyze whether each post is grammatically and semantically correct or not as shown in algorithm 1. If the post is incorrect grammatically/semantically, FATS ends the task of analysis and cannot continue with the process of annotation. However, anytime a post is both grammatically and semantically correct FATS produces a YES answer, at the same time comparing each element with the proposed patterns (see Section IV). The results are shown in Table 2 where a collection of 180 sentences of the 100 posts are tagged.

Algorithm 1 Annotating Posts of Blogs

```
1:  procedure APB(Posts,NumberofPosts)
2:      for i ← 1, NumberofPosts do
3:          for j ← 1, NumberofSentences(i) do
4:              if sentence(j) = pattern then
5:                  for k ← 1, NumberofElements do
6:                      element(k) ← syntactic(element(k));
7:                      element(k) ← mofpho(element(k));
8:                      element(k) ← semantic(element(k));
9:                      element(k) ← annotate(element(k));
10:                 end for
11:             end if
12:         end for
13:     end for
14: end procedure
```

Table II
RECALL AND PRECISION FOR THE POST

	NP	AP	VP	AdP
Recall	.93	.93	.93	.93
Precision	.93	.94	.94	.85

VI. CONCLUSIONS AND FUTURE WORKS

This paper presented FATS, a framework for tagging posts of blogs by using a table of symbols of subjectivity, a lexicon of reference [9], and eight patterns of analysis. Our framework automatically builds a subjectively annotated domain specific corpus by analyzing morphosyntactic, syntactic and semantic levels of posts with at least 90% recall and precision. The corpus resulting from this research can be used as it is for other applications because it is easy to integrate in other processes. We can see as shown in Fig. 4 that the accuracy of FATS in the tagging of NP, AP and VP is higher than in the tagging of AdP. However, the difference

 ISBN: 978-1-61208-162-5

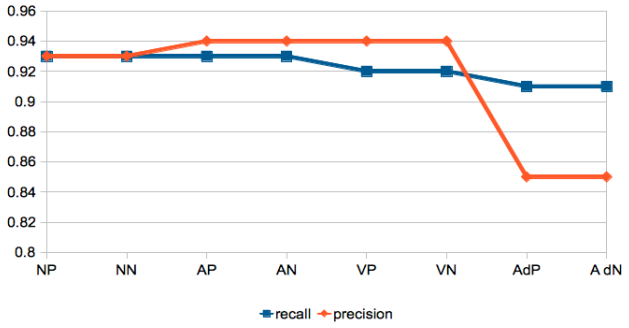

Figure 4. Precision for the posts.

is negligible and thus, it does not represent a limit for the present research. Future work will be done with ontologies to enrich and improve FATS.

ACKNOWLEDGMENT.

We are grateful to the Sciences Research Council (CONACYT) and COECYTJAL for funding this research project.

REFERENCES

[1] T. B.-L. Jim Hendler, "From the semantic web to social machines: A research challenge for ai on the world wide web," *Artif. Intell*, vol. 2, no. 174, pp. 156–161, 2010.

[2] http://en.wordpress.com/types-of-blogs/ (Accesed: June 2011).

[3] http://code.google.com/apis/blogger/ (Accesed: June 2011).

[4] http://mybloglog.com (Accesed: June 2011).

[5] T. S. A. in the UK, "Accessibility guidelines for written resources," *www.scoutbase.org.uk*, 2011.

[6] M. Muller-Prove, "Taxonomien und folksonomien tagging als neues hci-element (in german)," *I-com*, vol. 6, no. 1, pp. 14–18, 2007.

[7] NISO, "Guidelines for the construction, format, and management of monolingual controlled vocabularies." *ANSI/NISO Z39.19-2005 Bethesda, MD: National Information Standards Organization.*, 2005.

[8] http://www.ilc.cnr.it/EAGLES/annotate/annotate.html (Accesed: June 2011).

[9] M. R. Villarreal, "Corpus de blogs de viajes: análisis lingüístico para el reconocimiento de la valoración de la información. (in spanish)," *Proyecto Fin de Carrera. E. U. I. T. de Telecomunicación. Universidad Politécnica de Madrid.*, 2009.

[10] J. Wiebe, "Tracking point of view in narrative," *Proceedings of the NLPKE*, no. 174, 2008.

[11] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis." *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 353–363, 2006.

[12] A. Andreevskaia and S. Bergler, "Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses," *In Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics*, 2006.

[13] A. Esuli and F. Sebastiani, "Determining term subjectivity and term orientation for opinion mining," *In Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics*, pp. 193–200, 2006.

[14] J. Wiebe and R. Mihalcea, "Word sense and subjectivity," *In Proceedings of the 21st International Conference on Computational Linguistics*, 2006.

[15] S.-M. Kim and E. Hovy, "Automatic detection of opinion bearing words and sentences," *roceedings of the 2nd International Joint Conference on Natural Language Processing*, pp. 61–66, 2005.

[16] K. I. Y. M. K. T. Kobayashi, Nozomi and T. Fukushima, "Collecting evaluative expressions for opinion extraction," *Proceedings of the First International Joint Conference on Natural Language Processing*, 2004.

[17] SWANN, "Dictionary of socio linguistics," *The University of Alabama Press*, 2004.

[18] W. d. G. Jrgen Handke, *The structure of the lexicon: human versus machine*, 1995.

[19] R. Higashinaka, M. Walker, and R. Prasad, "Learning to generate naturalistic utterances using reviews in spoken dialogue systems," *ACM Transactions on Speech and Language Processing (TSLP)*, 2007.

[20] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06*, 2006, pp. 417–422.

[21] http://nlg18.csie.ntu.edu.tw:8080/opinion/userform.jsp (Accesed: June 2011).

[22] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Opinionfinder: a system for subjectivity analysis," in *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Association for Computational Linguistics, 2005, pp. 34–35.

[23] J. Kuhn, "Parsing word-aligned parallel corpora in a grammar induction context," in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ser. ParaText '05. Association for Computational Linguistics, 2005, pp. 17–24.

[24] T. Schmidt, "Creating and working with spoken language corpora in exmaralda," in *LULCL II: Lesser Used Languages and Computer Linguistics II*, 2009, pp. 151–164.

[25] C. Müller, "Representing and accessing multi-level annotations in mmax2," in *Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 73–76.

[26] http://crawler.archive.org/index.html (Accesed: June 2011).

 ISBN: 978-1-61208-162-5