On the incidence of part-of-speech on polarity identification of user-generated-content in Spanish

Rafael E. Banchs and Joan Codina

WOMSA: 1st Workshop on Opinion Mining and Sentiment Analysis, Nov. 2009

Barcelona Media
Centre d'Innovació

Motivation

Opinion mining and sentiment analysis technologies…

- Recent outburst of WEB 2.0 has promoted the increment of user-generated-contents

- Comments and opinions constitute a valuable source of collective information

- Very little work has been conducted for the specific case of the Spanish language

Barcelona Media
Centre d'Innovació

Main Objectives

Polarity estimation in Spanish user-generated-contents…

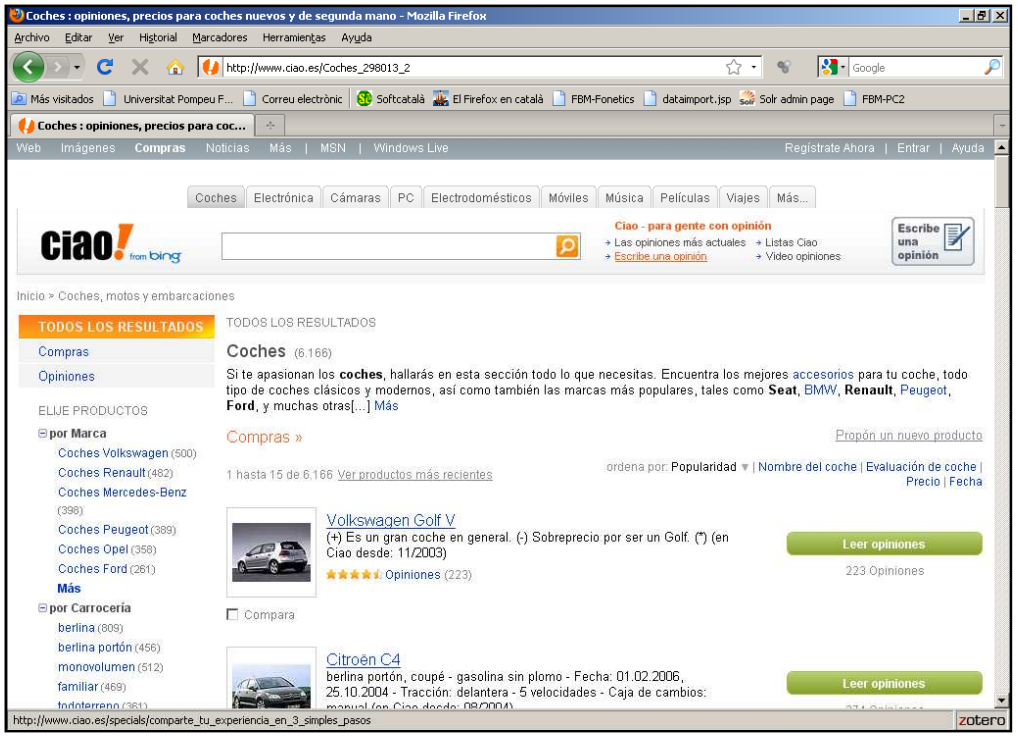
- To collect and prepare a Spanish dataset for opinion mining experimentation
- To explore the effects of Spanish morphology on the task of polarity estimation
- To explore the effects of different part-of-speech on the task of polarity estimation

Barcelona Media
Centre d'Innovació

Collection of the Dataset

Comments from the automotive section of www.ciao.es were crawled:

- Total number of comments: 25,330
- Each with a rating in a 1 to 5 scale
- Average number of words/comment ≈ 250



Barcelona Media
Centre d'Innovació

Preparing the Dataset (I)

Collected data contained a lot of corrupted characters due to encoding problems, which are generally resulting from “cut & paste” edition procedures

Spanish Characters (174 different types)		html Sequences (59 different types)	
corrupted	corrected	corrupted	corrected
Ã¡	á	&#149;	*
Ã©	é	&#324;	ñ
Ã-	í	&#733;	‘
Ã³	ó	&#8212;	“
Ãº	ú	&#9484;	*
Ã±	ñ	&#9668;	–

Preparing the Dataset (II)

The collected dataset was severely biased towards the positive-polarity…

- A balanced subset was selected

Ratings	0	1	2	3	4	5
Collected Comments	2	561	967	2098	7951	13751
Selected Comments	2	561	967	0	0	1530

Negative Samples

Positive Samples

Preparing the Dataset (III)

Morpho-syntactic annotation of the selected experimental dataset…

- Lexical unit and sentence segmentation
- Part-of-speech and lemma annotation

Resources used:

- OpenNLP toolkit (for segmentation)
- PAROLE tag set and TreeTagger algorithm with Spanish models trained with LEXEP corpus (for annotation)

Barcelona Media
Centre d'Innovació

Methodological Framework

A standard machine learning procedure:

- Supervised approach using binary classification engines based on Support Vector Machines
- Vector space model representation (bag-of-words) with standard TF-IDF weighting, and non stopword removal
- Twenty-fold cross-validation approach, with random selection of train and test sets (no overlapping)
- Average accuracies (and standard deviations) used as performance metric (and confidence interval estimate)

Barcelona Media
Centre d'Innovació

Experimental Setting

Two groups of experiments were conducted:

- Morphology effect on polarity detection evaluation
- Part-of-speech effect on polarity detection evaluation

Two experimental datasets were used:

Dataset	Polarity	Comments	Tokens	Full Forms	Lemmas	Length
Large	Positive	1,530	448.8 K	23.4 K	15.8 K	293.38
	Negative	1,530	423.4 K	22.4 K	15.1 K	276.75
	Both	3,060	872.3 K	34.6 K	23.3 K	285.07
Small	Positive	241	81.5 K	8.9 K	6.2 K	338.45
	Negative	232	72.9 K	7.1 K	5.0 K	314.38
	Both	473	154.5 K	12.5 K	8.5 K	326.65

Barcelona Media
Centre d'Innovació

Morphology Effects (I)

A comparative experiment:

- Reference classifier: trained by considering lemmas
- Contrastive classifier: trained by considering full forms

Dataset	Reference (lemmas)	Contrastive (full forms)
Small	75.96 (1.52)	77.47 (1.90)
Large	72.25 (2.61)	73.85 (2.51)

Barcelona Media
Centre d'Innovació

Morphology Effects (II)

A comparative experiment:

- Reference classifier: trained by considering lemmas
- Contrastive classifier: trained by considering full forms

Dataset	Class	Reference (lemmas)		Contrastive (full forms)	
		Precision	Recall	Precision	Recall
Small	**Positive**	74.72 (3.52)	77.90 (4.32)	76.81 (3.59)	78.70 (3.67)
	Negative	77.00 (2.83)	73.30 (5.46)	78.19 (2.69)	76.00 (4.91)
Large	**Positive**	71.78 (3.60)	73.20 (5.74)	72.29 (3.99)	75.55 (4.08)
	Negative	72.82 (4.49)	71.15 (4.61)	74.38 (4.02)	70.90 (5.18)

Barcelona Media
Centre d'Innovació

Part-Of-Speech Effects (I)

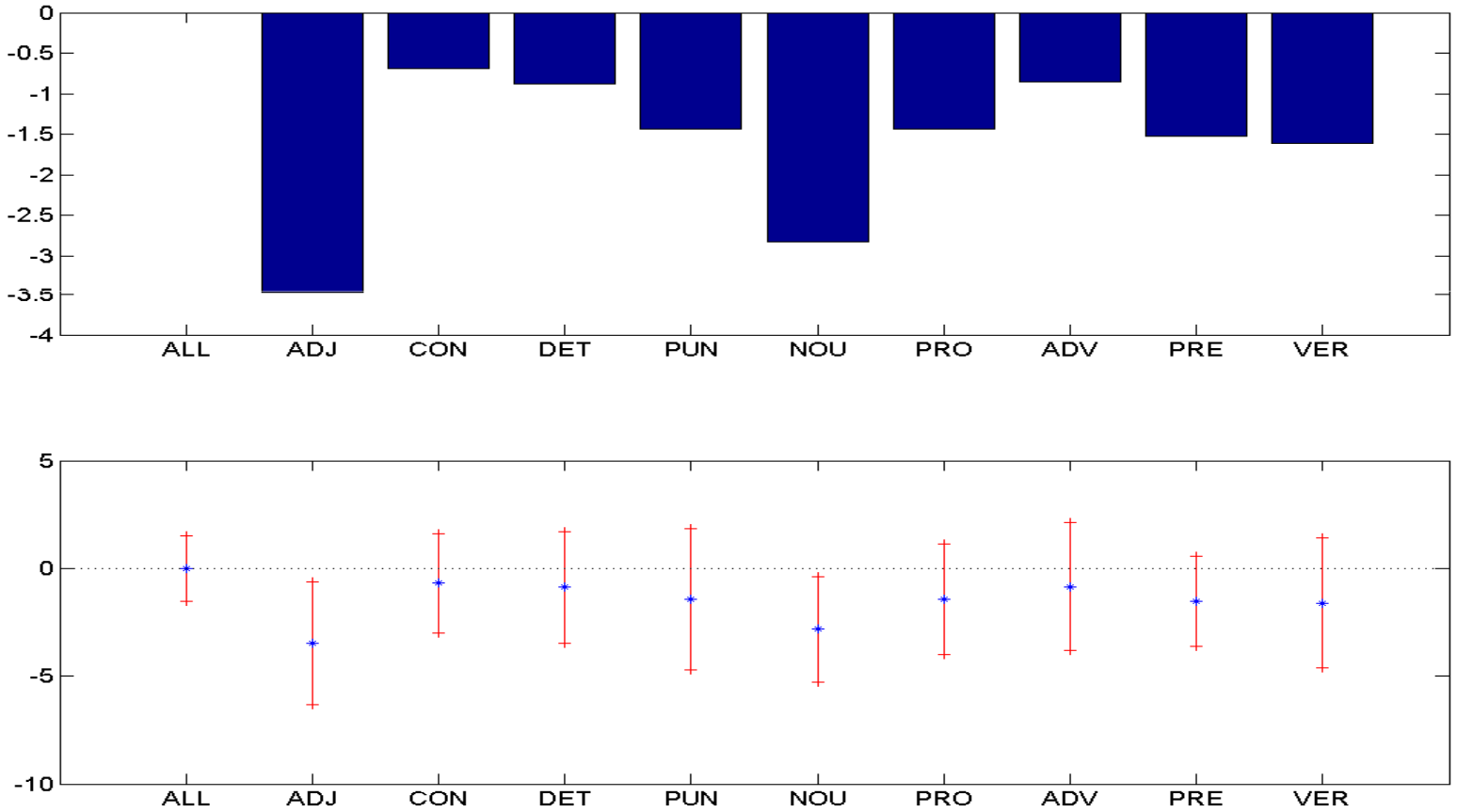
A comparative experiment:

- Reference classifier: considering all part-of-speech
- 9 contrastive classifiers: one part-of-speech removed

POS Category	Identifier	Corresponding POS Tags
Adjectives	ADJ	AQ, AC
Conjunctions	CON	CC, CS
Determiners	DET	DA, DD, DE, DI, DN, DP, DT
Punctuation	PUN	Fa, Fc, Fd, Fe, Fg, Fh, Fi, Fp, Fs, Fx, Fz
Nouns	NOU	NC, NP
Pronouns	PRO	PO, PD, PI, PN, PP, PR, PT, PX
Adverbs	ADV	RG, RN
Prepositions	PRE	SP
Verbs	VER	VA, VM, VS

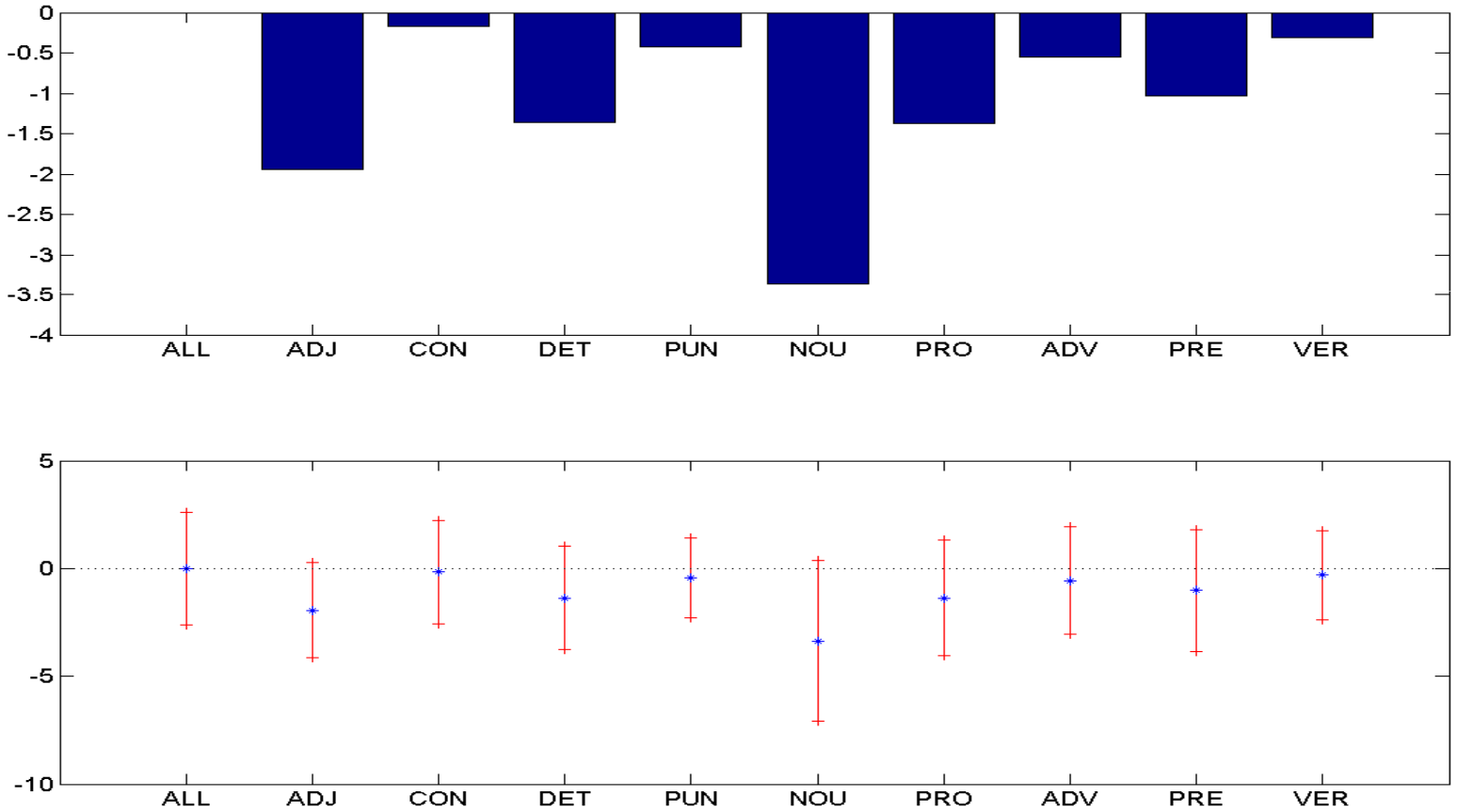
Part-Of-Speech Effects (II)

Results considering the small dataset:



Part-Of-Speech Effects (III)

Results considering the large dataset:



Barcelona Media
Centre d'Innovació

Part-Of-Speech Effects (IV)

Lemma vocabularies for the two most influential syntactic categories, and their absolute size differences between the small and large datasets

Dataset	Adjectives	Nouns
Small	1,749	4,067
Large	4,743	11,157
Absolute increment	2,994	7,090

Barcelona Media
Centre d'Innovació

Conclusions

Some conclusions derived from described experiments:

- Although Spanish morphology seems to be contributing to polarity detection, the observed effect is not statistically significant
- There is a clear tendency for nouns and adjectives to play an important role on polarity estimation, however this evidence was not statistically significant for experiments considering the large dataset
- A greater increment of noun vocabulary than adjective vocabulary seems to be explaining the greater incidence of nouns in experiments considering the large dataset

Barcelona Media
Centre d'Innovació

Future Work

Our future work in this area will focus on:

- Improving the quality of the constructed dataset (complete comments are actually a mix-up of positive and negative opinions), and scaling it up to sentence level.
- Improving the quality of morpho-syntactic annotation of user-generated-contents in Spanish.
- Exploring alternatives for dealing with special, but common, cases such as: negations, comparisons, etc.
- Addressing related problems such as opinion summarization and objective/subjective nature detection.

Barcelona Media

QUESTIONS...