GTI at SemEval-2016 Task 5: SVM and CRF for Aspect Detection and Unsupervised Aspect-Based Sentiment Analysis

Tamara Álvarez-López, Jonathan Juncal-Martínez, Milagros Fernández-Gavilanes
Enrique Costa-Montenegro, Francisco Javier González-Castaño
GTI Research Group
AtlantTIC Centre, School of Telecommunication Engineering, University of Vigo
36310 Vigo, Spain
`{talvarez, jonijm, milagros.fernandez, kike}@gti.uvigo.es,`
`javier@det.uvigo.es`

Abstract

This paper describes in detail the approach carried out by the GTI research group for SemEval 2016 Task 5: Aspect-Based Sentiment Analysis, for the different subtasks proposed, as well as languages and dataset contexts. In particular, we developed a system for category detection based on SVM. Then for the opinion target detection task we developed a system based on CRFs. Both are built for restaurants domain in English and Spanish languages. Finally for aspect-based sentiment analysis we carried out an unsupervised approach based on lexicons and syntactic dependencies, in English language for laptops and restaurants domains.

1 Introduction

In the last years, with the growth of Internet, people use it as a means of expressing their opinions and experiences about several subjects. That is the reason why there is a great amount of user generated information available online, through many different platforms, such as blogs, social networks, etc. This information became very valuable for companies, politicians, etc., who are interested in what users say about them or their products. Due to this, Sentiment Analysis (SA) techniques have attracted the interest of researches, trying to process all this amount of information by means of usually supervised methods based on classifiers.

Most of these researches focus on extracting the sentiment of a whole review or text (Liu, 2012). This is enough for many applications and purposes. However, sometimes there is a need for analysing the text in a deeper way, at entity or aspect level. For example, a review in the restaurants domain can include different opinions about different aspects, such as the service or the food quality, so it is interesting to distinguish the different opinions for each of these aspects. This is the reason why some studies emerged about the so-called aspect-based sentiment analysis (Marcheggiani et al., 2014; Lu et al., 2011).

Hence this is the subject of the task 5 of the SemEval 2016 (Pontiki et al., 2016), divided into different subtasks. Groups are asked to detect aspect categories in a review or sentence, which are predefined for each domain and formed by an entity and an attribute. Then, there is a subtask which consists of detecting the opinion target expression, which are related to the categories found. Finally, aspect-based sentiment analysis is required for one of the subtasks, associating a polarity, which can be positive, negative or neutral, to each of the categories found in the sentence or review. Datasets in different languages and domains are available for proving the approaches.

The remainder of this paper is structured as follows. In Section 2 we make a description of the system developed for all the subtasks. Section 3 contains the results of all the different subtasks, as well as detailed scores for each slot. Finally, in section 4 we summarize the main aspects of our system and extract some final conclusions.

2 System Overview

In this section we make a brief description of the system submitted for the different subtasks. We presented our submission for English restaurants

dataset for subtask 1, slots 1, 2 and 3, and subtask 2, slots 1 and 3. For English laptops dataset we sent a submission for subtasks 1 and 2 only in slot 3. Then, the system was also developed for Spanish language and restaurants dataset in subtasks 1, slots 1 and 2 and subtask 2, slot 1. In the next subsections we describe the different stages carried out for obtaining all the different results.

2.1 Preprocessing

As a first step for all the subtasks, each preprocessed social media review must first be broken into tokens, in order to derive the syntactic context. Part-of-speech (POS) tagging and lemmatization are performed to ensure that all the inflected forms of a word are covered. In the case of English, Stanford Tagger is applied due to its better results, however it does not provide lemmatization. That is why using the resulting form and tag, lemma is extracted by means of Freeling Tagger (Atserias et al., 2006; Padró and Stanilovsky, 2012). On the other hand, for Spanish language only Freeling Tagger is used. Freeling is a library that provides multiple languages among which are English and Spanish. Food and drinks recognition is also performed, based on dictionaries [1], in order to identify words referring to those topics for the subsequent processing of the sentences.

POS tagging allows the identification of lexical items that can contribute to the correct recognition of targets in a message. These items are namely adjectives, adverbs, verbs and nouns. The lemmatized and POS-annotated messages are fed to a parser that transforms the output of the tagger into a full parse tree. Finally, the tree is converted to dependencies, and the functions are annotated. The entire process is performed by means of Freeling Parser (Padró and Stanilovsky, 2012).

2.2 Subtask 1: Sentence-level Aspect-Based Sentiment Analysis (ABSA)

This subtask contains different slots, having participated in three of them, which are slot 1, slot 2 and slot 3. The system for Spanish and English language is exactly the same for both slots 1 and 2.

---

[1]Taken from the lists available at https://es.speaklanguages.com/inglés/vocabulario/comidas

2.2.1 Slot 1 Aspect category detection

The aim of this task is to assign to each sentence a category, which is a tuple (entity, attribute), from a given set of 12 different predefined categories. To do this, we used a linear SVM classifier combined with word lists. These word lists are created from the training file provided by the organization, which was composed of 2000 sentences, grouped in 350 reviews. Different datasets were provided for several languages and topics. Our system was developed for restaurants dataset, both in English and Spanish.

The library libsvm (Chang and Lin, 2011) was used to implement the SVM classifier, using the following features for each sentence:

- Words: those words appearing in the sentence, which are nouns, verbs or adjectives are extracted.

- Lemmas: lemmas from nouns, verbs and adjectives are selected.

- POS tags: part of speech from nouns, verbs and adjectives in the sentence.

- Bigrams: all the bigrams found in the sentence.

We developed 12 different binary classifiers, one for each possible category. If the output of one classifier for a particular sentence is "1", then we add the related category to the sentence. If more than one category is found for the same sentence, we add all of them to the list of categories. After this, the outputs are improved by means of our word lists, as we can see in Algorithm 1, executed for each sentence. The word lists were created automatically from the training file, extracting all the nouns and adjectives appearing in sentences from the same category, and manually filtered later in order to remove noisy items. Six different lists are composed, containing terms related to: ambience, service, prices, quality, style options and location.

The inputs defined for the following algorithm are the list of categories obtained from SVM for each sentence (CList(s)) and the six word lists created previously. The output is the new list per sentence, containing the old categories from SVM and the new ones added.

Algorithm 1: Combining SVM outputs with word lists for a sentence s.

Input: CList(s), ambienceL, serviceL, locationL, pricesL, qualityL, styleL
Output: newList(s)

1   newList(s) = CList(s);
2   **foreach** *unigram(s)* **do**
3   **if** *unigram(s)* ∈ *ambienceL* **then**
4   newList(s) = newList(s) ∪ {AMBIENCE#GENERAL}
5   **end**
6   **if** *unigram(s)* ∈ *serviceL* **then**
7   newList(s) = newList(s) ∪ {SERVICE#GENERAL}
8   **end**
9   **if** *unigram(s)* ∈ *locationL* **then**
10   newList(s) = newList(s) ∪ {LOCATION#GENERAL}
11   **end**
12   **if** *unigram(s)* ∈ *pricesL* **then**
13   **if** *FOOD#A* ∈ *CList(s)* **then**
14   newList(s) = newList(s) ∪ {FOOD#PRICES}
15   **else**
16   **if** *DRINKS#A* ∈ *CList(s)* **then**
17   newList(s) = newList(s) ∪ {DRINKS#PRICES}
18   **else**
19   newList(s) = newList(s) ∪ {RESTAURANT#PRICES}
20   **end**
21   **end**
22   **end**
23   **if** *unigram(s)* ∈ *qualityL* **then**
24   **if** *DRINKS#A* ∈ *CList(s)* **then**
25   newList(s) = newList(s) ∪ {DRINKS#QUALITY}
26   **else**
27   newList(s) = newList(s) ∪ {FOOD#QUALITY}
28   **end**
29   **end**
30   **if** *unigram(s)* ∈ *styleL* **then**
31   **if** *DRINKS#A* ∈ *CList(s)* **then**
32   newList(s) = newList(s) ∪ {DRINKS#STYLEOPTIONS}
33   **else**
34   newList(s) = newList(s) ∪ {FOOD#STYLEOPTIONS}
35   **end**
36   **end**
37   **end**

2.2.2 Slot 2 Opinion target expression

For this slot, teams were asked to extract the exact expressions or words in the sentence, in which an opinion is expressed. The implementation for this slot is made by means of CRFs, using CRF++ tool (Kudo, 2005) and the training file provided for building the model. A training file is needed to build as input for the CRF, whose structure is as follows. In the first column, all the words for every sentence are written, then in the second column, the corresponding lemma. The third column represents the tag and the last one represents if the word is an aspect or not or if it is included in a multiword aspect. Then for creating the model we take into account all these features, as well as all the possible bigrams in each sentence. In the output, if no target is found, no opinion is returned for that sentence.

2.2.3 Slot 3 Sentiment polarity

This slot is implemented only for English language, both restaurants and laptops datasets. Our system is fully unsupervised, this can explain the low results obtained for this slot. An adjustment was made to the system already implemented for sentiment analysis in the whole sentence, which was presented in Semeval 2015, task 10: sentiment analysis in Twitter (Fernández-Gavilanes et al., 2015), which was also unsupervised. For this dataset, a new polarity lexicon was generated automatically from the training dataset, applying a polarity rank algorithm, as explained in the mentioned article. Then, it was merged with SOCAL (Taboada et al., 2011) and AFINN (Nielsen, 2011) lexicons, which are general context ones, by applying an average for those words which appeared in more than one of them.

Our system for the restaurant dataset implements the following syntactic rules:

- If there is no opinion or only one target expression in the sentence, the system automatically takes the polarity of the whole sentence and assign it to all the categories which appear in this sentence.

- If there is only one different target expression but appearing more than once, we check if there is an adversative clause in the sentence built with "but" particle. If not, we also take the

polarity of the whole sentence for all the opinions. If the previous condition is fulfilled, we will take the polarity of the first clause of the sentence, which is the piece of sentence placed before the "but" and then apply a polarity linear system, which consists of summing up all the polarities found in the dictionary created. For the next opinions which have the same target, we will follow the same procedure but with the piece of sentence after the "but". For this linear approach, we take negations in account only for adjectives, flipping the polarity of the adjectives which come inmediately after a negation particle, as "no" or "not".

- When there are several different opinion targets, we split the sentence to detect the scope of each target and apply the same linear polarity algorithm explained in the previous point. To detect the scope of the target, we take the words which appear before and after the target, splitting by punctuation marks (";", ",", ".", "?", "!", "-").

For the laptops dataset, since there are no opinion target expressions, we take the polarity of the whole sentence to assign the polarity of each category.

2.3 Subtask 2: Text-level ABSA

Subtask 2 is similar to subtask 1, but instead of implementing aspect detection at sentence-level, it is performed at text-level. Participants are asked to implement slots 1 and 3 for this subtask. We participate in slot 1 for Spanish and English language, following the same procedure for both. Slot 3 is just implemented for English language for restaurants and laptops datasets.

2.3.1 Slot 1 Aspect category detection

Once we performed aspect category detection at sentence-level, we use this output as input for text-level detection. All the categories found are grouped at sentence-level and added all of them at review-level. Besides this, if RESTAURANT#GENERAL is not explicitly assigned to any sentence of the review, we add it anyway.

2.3.2 Slot 3 Sentiment polarity

Similarly to slot 1, we use the output from subtask 1 slot 3 as input for this slot. All the polarities found are again grouped for all the sentences contained in the review and added them to text-level. If there are different polarities for the same category, some rules are applied: if polarities are negative and neutral, negative is finally assigned; if there are positive and neutral opinions, positive polarity is assigned; if there are positive and negative opinions for the same category, the tag "conflict" is assigned to that category at review-level.

Moreover, as RESTAURANT#GENERAL is compulsory for every review, if no sentence has this category assigned, we take into account all the polarities of the other categories found and then assign the polarity for this category. Again, if there are different polarities containing positive and negative, "conflict" tag is assigned. The same process is followed for laptops dataset, with the LAPTOPS#GENERAL category.

3 Experimental Results

In this section, we describe the experiments carried out for the different subtasks and slots and the datasets provided by the organization. These datasets are composed of several reviews, splitted in sentences, for restaurants and laptops topics. The performance of slots 1 and 2, for both subtasks, are measured by means of the F-score, while slot 3 is evaluated by means of the accuracy.

Table 1 represents the precision, recall and F-score obtained for restaurants datasets and all the slots submitted. For English language, an unconstrained system was presented, while for Spanish language both constrained and unconstrained systems were submitted. The constrained approaches do not need any external resources, but only the training files provided, while in the unconstrained ones, food and drinks lexicon was used in the pre-processing step for identifying different foods and drinks.

It can be seen that there is not much difference between constrained and unconstrained systems for Spanish language, so we can assume that the recognition of different names of foods or drinks does not increase the knowledge of the classifiers, perform-

		Prec.	Rec.	F
	Subt1-Slot1	72.14	63.79	67.71
EN - U	Subt1-Slot2	69.45	63.89	66.55
	Subt2-Slot1	87.00	81.19	83.99
	Subt1-Slot1	74.82	66.80	70.59
SP - U	Subt1-Slot2	69.94	66.90	68.39
	Subt2-Slot1	86.31	81.89	84.04
	Subt1-Slot1	74.59	65.98	70.02
SP - C	Subt1-Slot2	69.45	67.60	68.51
	Subt2-Slot1	86.63	81.89	84.19

Table 1: Measures for restaurants dataset, slots 1 and 2.

ing almost equally. Moreover, we can state that our system perfoms as well for English as for Spanish language.

In Table 2, the detailed scores for slot 3 are shown in English language, for restaurants dataset, likewise in Table 3 for laptops dataset.

		Prec.	Rec.	F	Acc.
	P	84.66	76.76	80.52	
Subt1	N	60.5	59.31	59.9	69.96
	NEU	10.48	25.00	14.77	
	P	87.2	76.22	81.34	
Subt2	N	62.75	38.1	47.41	64.11
	NEU	18.18	8.7	11.76	
	CONFL.	7.61	63.64	13.59	

Table 2: Detailed scores for slot 3, restaurants dataset in English language.

		Prec.	Rec.	F	Acc.
	P	68.78	87.94	77.19	
Subt1	N	63.39	42.34	50.77	67.29
	NEU	0	0	0	
	P	74.64	76.63	75.62	
Subt2	N	60.81	27.78	38.14	58.35
	NEU	12.12	12.9	12.5	
	CONFL.	10.99	71.43	19.05	

Table 3: Detailed scores for slot 3, laptops dataset in English language.

As it can be seen in Table 2 and Table 3, the results obtained for the sentiment slot are not quite competitive with the other teams. This can be due to the fact that our system is fully unsupervised, while the others are usually supervised systems, based on training. Moreover, we performed a simple adaptation from our original system, made for sentiment analysis in Twitter, presented to SemEval 2015, so there is still a lot of improvement on this field.

4 Conclusions

This paper describes the participation of the GTI group, AtlantTIC Research Center, University of Vigo, in the SemEval 2016, Task 5: Aspect-Based Sentiment Analysis. We developed a supervised system based on SVM classifiers for category detection, and CRFs for opinion target detection. Then, for the aspect-based sentiment analysis we submitted a fully unsupervised system, based on syntactic dependencies and context-based polarity lexicons.

Test sets				Position
			Slot1	10/20
		Subtask1	Slot2	4/15
	REST		Slot3	19/20
EN		Subtask2	Slot1	1/3
			Slot3	4/4
	LAPT	Subtask1	Slot3	14/15
		Subtask2	Slot3	4/4
		Subtask1	Slot1	1/6
SP	REST		Slot2	1/3
		Subtask2	Slot1	1/2

Table 4: Position of our approach in the different datasets and subtasks submitted, according to the results published by the organisation.

As we can see in Table 4, competitive results were obtained for aspect and category detection, being in first position for Spanish language, both in subtask 1 and subtask 2. Moreover, in subtask 2, which is aspect detection at review level, we also achieved the first position for English language in restaurants datasets. However, our system did not perform as well as expected in slot 3, maybe due to the fact of the lack of supervision for our model. It results not competitive against other supervised approaches, although its main advantage is that there is no need of training sets, which is time and resource consuming in order to manually tag them.

Acknowledgments

References

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró.

2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of LREC*, volume 6, pages 48–55.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Milagros Fernández-Gavilanes, Tamara Álvarez López, Jonathan Juncal-Martínez, Enrique Costa-Montenegro, and Francisco Javier González-Castaño. 2015. GTI: An Unsupervised Approach for Sentiment Analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 533–538, Denver, Colorado, June. Association for Computational Linguistics.

Taku Kudo. 2005. Crf++: Yet another crf toolkit. *Software available at http://crfpp. sourceforge. net*.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 81–88. IEEE.

Diego Marcheggiani, Oscar Täckström, Andrea Esuli, and Fabrizio Sebastiani. 2014. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In *Advances in Information Retrieval*, pages 273–285. Springer.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *LREC2012*.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeny Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June 2016. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.