EliXa: A modular and flexible ABSA platform

Iñaki San Vicente, Xabier Saralegi
Elhuyar Foundation
Osinalde industrialdea 3
Usurbil, 20170, Spain
{i.sanvicente,x.saralegi}@elhuyar.com

Rodrigo Agerri
IXA NLP Group
University of the Basque Country (UPV/EHU)
Donostia-San Sebastián
rodrigo.agerri@ehu.eus

Abstract

This paper presents a supervised Aspect Based Sentiment Analysis (ABSA) system. Our aim is to develop a modular platform which allows to easily conduct experiments by replacing the modules or adding new features. We obtain the best result in the Opinion Target Extraction (OTE) task (slot 2) using an off-the-shelf sequence labeler. The target polarity classification (slot 3) is addressed by means of a multiclass SVM algorithm which includes lexical based features such as the polarity values obtained from domain and open polarity lexicons. The system obtains accuracies of 0.70 and 0.73 for the restaurant and laptop domain respectively, and performs second best in the out-of-domain hotel, achieving an accuracy of 0.80.

1 Introduction

Nowadays Sentiment Analysis is proving very useful for tasks such as decision making and market analysis. The ever increasing interest is also shown in the number of related shared tasks organized: TASS (Villena-Román et al., 2012; Villena-Román et al., 2014), SemEval (Nakov et al., 2013; Pontiki et al., 2014; Rosenthal et al., 2014), or the SemSA Challenge at ESWC2014[1]. Research has also been evolving towards specific opinion elements such as entities or properties of a certain opinion target, which is also known as ABSA. The Semeval 2015 ABSA shared task aims at covering the most common problems in an ABSA task: detecting the specific topics an opinion refers to (slot1); extracting the opinion targets (slot2), combining the topic and target identification (slot1&2) and, finally, computing the polarity of the identified word/targets (slot3). Participants were allowed to send one constrained (no external resources allowed) and one unconstrained run for each subtask. We participated in the slot2 and slot3 subtasks.

Our main is to develop an ABSA system to be used in the future for further experimentation. Thus, rather than focusing on tuning the different modules our main goal is to develop a platform to facilitate future experimentation. The EliXa system consists of three independent supervised modules based on the IXA pipes tools (Agerri et al., 2014) and Weka (Hall et al., 2009). Next section describes the external resources used in the unconstrained systems. Sections 3 and 4 describe the systems developed for each subtask and briefly discuss the obtained results.

2 External Resources

Several polarity Lexicons and various corpora were used for the unconstrained versions of our systems. To facilitate reproducibility of results, every resource listed here is publicly available.

2.1 Corpora

For the restaurant domain we used the Yelp Dataset Challenge dataset[2]. Following (Kiritchenko et al., 2014), we manually filtered out categories not corresponding to food related businesses (173 out of 720

[1] http://challenges.2014.eswc-conferences.org/index.php/SemSA

[2] http://www.yelp.com/dataset_challenge

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 748–752,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

were finally selected). A total of 997,721 reviews (117.1M tokens) comprise what we henceforth call the *Yelp food corpus* (C_{Yelp}).

For the laptop domain we leveraged a corpus composed of Amazon reviews of electronic devices (Jo and Oh, 2011). Although only 17,53% of the reviews belong to laptop products, early experiments showed the advantage of using the full corpus for both slot 2 and slot 3 subtasks. The *Amazon electronics corpus* (C_{Amazon}) consists of 24,259 reviews (4.4M tokens). Finally, the English Wikipedia was also used to induce word clusters using word2vec (Mikolov et al., 2013).

2.2 Polarity Lexicons

We generated two types of polarity lexicons to represent polarity in the slot3 subtasks: general purpose and domain specific polarity lexicons.

A general purpose polarity lexicon L_{gen} was built by combining four well known polarity lexicons: SentiWordnet SWN (Baccianella et al., 2010), General Inquirer GI (Stone et al., 1966), Opinion Finder OF (Wilson et al., 2005) and Liu's sentiment lexicon Liu (Hu and Liu, 2004). When a lemma occurs in several lexicons, its polarity is solved according to the following priority order: $Liu > OF > GI > SWN$. The order was set based on the results of (San Vicente et al., 2014). All polarity weights were normalized to a $[-1, 1]$ interval. Polarity categories were mapped to weights for GI (neg_+→-0.8; neg→-0.6; neg_-→-0.2; pos_-→0.2; pos→0.6; pos_+→0.8), Liu and OF (neg→-0.7; pos→0.7 for both). In addition, a restricted lexicon L_{genres} including only the strongest polarity words was derived from L_{gen} by applying a threshold of ±0.6.

Domain	Polarity Lexicon	Total
General	L_{gen}	42,218
General	L_{genres}	12,398
Electronic devices	L_{Amazon}	4,511
Food	L_{Yelp}	4,691

Table 1: Statistics of the polarity lexicons.

Domain specific polarity lexicons L_{Yelp} and L_{Amazon} were automatically extracted from C_{Yelp} and C_{Amazon} reviews corpora. Reviews are rated in a $[1..5]$ interval, being 1 the most negative and 5 the most positive. Using the Log-likelihood ratio (LLR) (Dunning, 1993) we obtained the ranking of the words which occur more with negative and positive reviews respectively. We considered reviews with 1 and 2 rating as negative and those with 4 and 5 ratings as positive. LLR scores were normalized to a $[-1, 1]$ interval and included in L_{Yelp} and L_{Amazon} lexicons as polarity weights.

3 Slot2 Subtask: Opinion Target Extraction

The Opinion Target Extraction task (OTE) is addressed as a sequence labeling problem. We use the *ixa-pipe-nerc* Named Entity Recognition system[3] (Agerri et al., 2014) off-the-shelf to train our OTE models; the system learns supervised models via the Perceptron algorithm as described by (Collins, 2002). *ixa-pipe-nerc* uses the Apache OpenNLP project implementation of the Perceptron algorithm[4] customized with its own features. Specifically, *ixa-pipe-nerc* implements basic non-linguistic local features and on top of those a combination of word class representation features partially inspired by (Turian et al., 2010). The word representation features use large amounts of unlabeled data. The result is a quite simple but competitive system which obtains the best constrained and unconstrained results and the first and third best overall results.

The local features implemented are: current token and token shape (digits, lowercase, punctuation, etc.) in a 2 range window, previous prediction, beginning of sentence, 4 characters in prefix and suffix, bigrams and trigrams (token and shape). On top of them we induce three types of word representations:

- Brown (Brown et al., 1992) clusters, taking the 4th, 8th, 12th and 20th node in the path. We induced 1000 clusters on the Yelp reviews dataset described in section 2.1 using the tool implemented by Liang[5].
- Clark (Clark, 2003) clusters, using the standard configuration to induce 200 clusters on the Yelp reviews dataset and 100 clusters on the food portion of the Yelp reviews dataset.

[3] https://github.com/ixa-ehu/ixa-pipe-nerc

[4] http://opennlp.apache.org/

[5] https://github.com/percyliang/brown-cluster

- Word2vec (Mikolov et al., 2013) clusters, based on K-means applied over the extracted word vectors using the skip-gram algorithm[6]; 400 clusters were induced using the Wikipedia.

The implementation of the clustering features looks for the cluster class of the incoming token in one or more of the clustering lexicons induced following the three methods listed above. If found, then we add the class as a feature. The Brown clusters only apply to the token related features, which are duplicated. We chose the best combination of features using 5-fold cross validation, obtaining 73.03 F1 score with local features (e.g. constrained mode) and 77.12 adding the word clustering features, namely, in unconstrained mode. These two configurations were used to process the test set in this task. Table 2 lists the official results for the first 4 systems in the task.

System (type)	**Precision**	**Recall**	**F1 score**
Baseline	55.42	43.4	48.68
EliXa (u)	68.93	71.22	**70.05**
NLANGP (u)	70.53	64.02	67.12
EliXa (c)	67.23	66.61	66.91
IHS-RD-Belarus (c)	67.58	59.23	63.13

Table 2: Results obtained on the slot2 evaluation on restaurant data.

The results show that leveraging unlabeled text is helpful in the OTE task, obtaining an increase of 7 points in recall. It is also worth mentioning that our constrained system (using non-linguistic local features) performs very closely to the second best overall system by the NLANGP team (unconstrained). Finally, we would like to point out to the overall low results in this task (for example, compared to the 2014 edition), due to the very small and difficult training set (e.g., containing many short samples such as "Tasty Dog!") which made it extremely hard to learn good models for this task. The OTE models will be made freely available in the *ixa-pipe-nerc* website in time for SemEval 2015.

4 Slot3 Subtask: Sentiment Polarity

The EliXa system implements a single multiclass SVM classifier. We use the SMO implementation provided by the Weka library (Hall et al., 2009). All the classifiers built over the training data were evaluated via 10-fold cross validation. The complexity parameter was optimized as ($C = 1.0$). Many configurations were tested in this experiments, but in the following we only will describe the final setting.

4.1 Baseline

The very first features we introduced in our classifier were token ngrams. Initial experiments showed that lemma ngrams (lgrams) performed better than raw form ngrams. One feature per lgram is added to the vector representation, and lemma frequency is stored. With respect to the ngram size used, we tested up to 4-gram features and improvement was achieved in laptop domain but only when not combined with other features.

4.2 PoS

PoS tag and lemma information, obtained using the IXA pipes tools (Agerri et al., 2014), were also included as features. One feature per PoS tag was added again storing the number of occurrences of a tag in the sentence. These features slightly improve over the baseline only in the restaurant domain.

4.3 Window

Given that a sentence may contain multiple opinions, we define a window span around a given opinion target (5 words before and 5 words after). When the target of an opinion is null the whole sentence is taken as span. Only the restaurant and hotel domains contained gold target annotations so we did not use this feature in the laptop domain.

4.4 Polarity Lexicons

The positive and negative scores we extracted as features from both general purpose and domain specific lexicons. Both scores are calculated as the sum of every positive/negative score in the corresponding lexicon divided by the number of words in the sentence. Features obtained from the general lexicons provide a slight improvement. L_{genres} is better for restaurant domain, while L_{gen} is better for laptops. Domain specific lexicons L_{Amazon} and L_{Yelp} also help as shown by tables 3 and 4.

[6] https://code.google.com/p/word2vec/

4.5 Word Clusters

Word2vec clustering features combine best with the rest as shown by table 3. These features only were useful for the restaurant domain, perhaps due to the small size of the laptops domain data.

4.6 Feature combinations

Every feature, when used in isolation, only marginally improves the baseline. Some of them, such as the E&A features (using the gold information from the slot1 subtask) for the laptop domain, only help when combined with others. Best performance is achieved when several features are combined. As shown by tables 4 and 5, improvement over the baseline ranges between 2,8% and 1,9% in the laptop and restaurant domains respectively.

Classifier	Acc Rest
Baseline (organizers)	78.8
Baseline	
1lgram	80.11
2lgram	79.3
$1lgram + E\&A$	79.8
$1lgram(w5)$	80.41
$1lgram + PoS$	80.59 (c)
Lexicons	
$1lgram + L_{gen}$	80.6
$1lgram + L_{genres}$	81
$1lgram + L_{Yelp}$	80.9
Combinations	
$1lgram(w5) + w2v(C_{Yelp}) + L_{genres} + L_{Yelp} + PoS$	82.34 (u)

Table 3: Slot3 ablation experiments for restaurants. (c) and (u) refer to constrained and unconstrained tracks.

4.7 Results

Table 5 shows the result achieved by our sentiment polarity classifier. Although for both restaurant and laptops domains we obtain results over the baseline both performance are modest.

In contrast, for the out of domain track, which was evaluated on hotel reviews our system obtains the third highest score. Because of the similarity of the domains, we straightforwardly applied our restaurant domain models. The good results of the constrained system could mean that the feature combination used may be robust across domains. With respect to the unconstrained system, we suspect that such a good performance is achieved due to the fact that word cluster information was very adequate for the hotel domain, because C_{yelp} contains a 10.55% of hotel reviews.

Classifier	Acc Lapt
Baseline (organizers)	78.3
Baseline	
1lgram	79.33
2lgram	79.7
$1lgram + clusters(w2v)$	79.23
$1lgram + E\&A$	79.23
$1lgram + PoS$	78.88
Lexicons	
$1lgram + L_{gen}$	79.2
$1lgram + L_{genres}$	79
$1lgram + L_{Amazon}$	79.7
Combinations	
$1lgram + PoS + E\&A$	79.99 (c)
$2lgram + PoS + E\&A$	78.27
$1lgram + L_{genres} + L_{Amazon} + PoS + E\&A$	80.85 (u)

Table 4: Slot3 ablation experiments for laptops; (c) and (u) refer to constrained and unconstrained tracks.

System	**Rest.**	**Lapt.**	**Hotel**
Baseline	63.55	69.97	71.68 (majority)
Sentiue	**78.70 (1)**	**79.35 (1)**	71.68 (4)
lsislif	75.50 (3)	77.87 (3)	**85.84 (1)**
EliXa (u)	70.06(10)	72.92 (7)	79.65 (3)
EliXa (c)	67.34 (14)	71.55 (9)	74.93 (5)

Table 5: Results obtained on the slot3 evaluation on restaurant data; ranking in brackets.

5 Conclusions

We have presented a modular and supervised ABSA platform developed to facilitate future experimentation in the field. We submitted runs corresponding to the slot2 and slot3 subtasks, obtaining competitive results. In particular, we obtained the best results in slot2 (OTE) and for slot3 we obtain 3rd best result in the out-of-domain track, which is nice for a supervised system. Finally, a system for topic detection (slot1) is currently under development.

6 Acknowledgments

This work has been supported by the following projects: ADi project (Etortek grant No. IE-14-382), NewsReader (FP7-ICT 2011-8-316404), SKaTer (TIN2012-38584-C06-02) and Tacardi (TIN2012-38523-C02-01).

References

Rodrigo Agerri, Josu Bermudez, and German Rigau. 2014. Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, pages 26–31, Reykjavik, Iceland, May.

S. Baccianella, A. Esuli, and F. Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation (LREC-2010), Malta.*, volume 25.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 59–66.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computacional Linguistics*, 19(1):61–74, March.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, november.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 815–824, New York, NY, USA. ACM.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland, August.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval*, volume 14.

Iñaki San Vicente, Rodrigo Agerri, and German Rigau. 2014. Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL2014*, pages 88–97, Gothenburg, Sweden.

P. Stone, D. Dunphy, M. Smith, and D. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge (MA): MIT Press.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July.

Julio Villena-Román, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González-Cristóbal. 2012. Tass-workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural*, 50:37–44.

Julio Villena-Román, Janine García-Morera, Sara Lana-Serrano, and José Carlos González-Cristóbal. 2014. Tass 2013 - a second step in reputation analysis in spanish. *Procesamiento del Lenguaje Natural*, 52(0).

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, page 347–354.