Language Independent Sequence Labelling for Opinion Target Extraction (Extended Abstract)*

Rodrigo Agerri† and **German Rigau**
HiTZ Centre, IXA Group, University of the Basque Country UPV/EHU
{rodrigo.agerri, german.rigau}@ehu.eus

Abstract

In this paper we present a language independent system to model Opinion Target Extraction (OTE) as a sequence labelling task. The system consists of a combination of clustering features implemented on top of a simple set of shallow local features. Experiments on the well known Aspect Based Sentiment Analysis (ABSA) benchmarks show that our approach is very competitive across languages, obtaining, at the time of writing, best results for six languages in seven different datasets. Furthermore, the results provide further insights into the behaviour of clustering features for sequence labeling tasks. Finally, we also show that these results can be outperformed by recent advances in contextual word embeddings and the transformer architecture. The system and models generated in this work are available for public use and to facilitate reproducibility of results.

1 Introduction

Early approaches to Opinion Mining and Sentiment Analysis (OMSA) were based on document classification, where the task was to determine the polarity (positive, negative, neutral) of a given document or review [Pang and Lee, 2008; Liu, 2012]. Later on, a finer-grained OMSA has been proposed motivated by the fact that a given review may contain more than one opinion about a variety of aspects or attributes of a given product is usually conveyed. Thus, Aspect Based Sentiment Analysis (ABSA) was defined as a task which consisted of identifying several components of a given opinion: the opinion holder, the target, the opinion expression (the textual expression conveying polarity) and the aspects or features. Aspects are mostly domain-dependent. In restaurant reviews, relevant aspects would include "food quality", "price", "service", etc. Similarly, reviews about consumer electronics would include aspects such as "size", "battery life", "hard drive capacity", and so on.

In this work we focus on Opinion Target Extraction (OTE), which we model as a sequence labelling task. Example (1) shows a review in which the target terms are tagged as being at the beginning (B-target), inside (I-target) or outside (O) of the opinion target expression (note that the target of the third opinion in this review is implicit).

(1) **Chow/B-target fun/I-target** was/O dry/O; **pork/B-target shu/I-target mai/I-target** was/O more/O than/O usually/O greasy/O and/O had/O to/O share/O a/O table/O with/O loud/O and/O rude/O family/O.

We present a language independent system which consists of a set of local, shallow features complemented with semantic distributional features based on clusters obtained from a variety of data sources. Our approach, despite the lack of hand-engineered, language-specific features, obtains state-of-the-art results in 7 datasets for 6 languages on the ABSA benchmarks (at the time of publication). This is due to the use of dense, cluster-based word representations obtained from large amounts of unlabeled data. Furthermore, we update the original journal paper by including results using XLM-RoBERTa, a transformer architecture [Devlin *et al.*, 2018] which uses contextual embeddings pre-trained in a large language model for 100 languages [Conneau *et al.*, 2019].

The main contributions of this paper are the following: we provide a simple and fast approach to OTE based on a framework developed for Named Entity Recognition (NER) [Agerri and Rigau, 2016]. We empirically demonstrate the validity and strong performance of our approach for six languages in seven different datasets of the restaurant domain. We show that our approach is not only competitive across languages and domains for NER, but that it can be straightforwardly adapted to different tasks and domains such as OTE. Furthermore, the system and models are available for public use and to facilitate reproducibility of results[1]. Finally, we compare with state-of-the-art results obtained by fine-tuning XLM-RoBERTa both for each language and in a zero-shot setting [Jebbara and Cimiano, 2019].

2 Background

The Aspect Based Sentiment Analysis (ABSA) tasks at SemEval [Pontiki *et al.*, 2014; Pontiki *et al.*, 2015; Pontiki *et al.*, 2016] provided standard training and evaluation data thereby

[*] This paper is an extended abstract of the Artificial Intelligence Journal publication [Agerri and Rigau, 2019], https://doi.org/10.1016/j.artint.2018.12.002

[†] Contact Author

[1] https://github.com/ixa-ehu/ixa-pipe-opinion

Language	ABSA	No. of Tokens and Opinion Targets					
		Train			Test		
		Token	B-target	I-target	Token	B-target	I-target
en	2014	47028	3687	1457	12606	1134	524
en	2015	18488	1199	538	10412	542	264
en	2016	28900	1743	797	9952	612	274
es	2016	35847	1858	742	13179	713	173
fr	2016	26777	1641	443	11646	650	239
nl	2016	24788	1231	331	7606	373	81
ru	2016	51509	3078	953	16999	952	372
tr	2016	12406	1374	516	1316	145	61

Table 1: ABSA SemEval 2014-2016 datasets for the restaurant domain. B-target indicates the number of opinion targets in each set; I-target refers to the number of multiword targets.

helping to establish a clear benchmark for the OTE task. The ABSA 2014 and 2015 tasks consisted of English reviews only, whereas in the 2016 task 7 more languages were added. The only constant in each of the ABSA editions was the inclusion, for the Opinion Target Extraction (OTE) sub-task, of restaurant reviews for every language. Thus, we decided to focus on the restaurant domain across 6 languages and the three different ABSA editions. The ABSA task consisted of identifying, for each opinion, the opinion target, the aspect referred to by the opinion and the aspect's polarity. It should be noted that, out of the three opinion components, only the targets are explicitly annotated in the text. Opinion expressions such as "dry", "greasy" or "loud and rude" are not annotated.

Among the participants (for English) one team [Toh and Wang, 2014; Toh and Su, 2015] was quite successful. For the first two editions they developed a CRF system with extensive handcrafted linguistic features. For ABSA 2016, added the output of a Recurrent Neural Network (RNN) to provide additional features. They were the best system in 2014 and 2016. In 2015 the best system was a preliminary version of the one presented in this work [San Vicente *et al.*, 2015].

From 2015 onwards most works are based on deep learning. [Poria *et al.*, 2016] presented a 7 layer deep CNN combining word embeddings trained on a 5 billion word corpus extracted from Amazon [McAuley and Leskovec, 2013], POS tag features and manually developed linguistic patterns based on syntactic analysis and SenticNet [Cambria *et al.*, 2014]. They only evaluate their system on the English 2014 data, obtaining best results up to date on that benchmark. More recently, [Wang *et al.*, 2017] proposed a coupled multi-layer attention (CMLA) network where each layer consists of a couple of attentions with tensor operators. While previous successful approaches modelled OTE as an independent task, in the CMLA model the attentions interactively learn both the opinion targets and the opinion expressions. As opinion expressions are not available in the original ABSA datasets, they had to manually annotate the training and testing data with the required opinion expressions. Using this new manual information to train their CMLA network they reported the best results so far for ABSA 2014 and 2015 (English only).

Finally, [Li and Lam, 2017] develop a multi-task learning framework consisting of two LSTMs equipped with extended memories and neural memory operations. As [Wang *et al.*, 2017], they use opinion expressions annotations for a joint modeling of opinion targets and expressions. However, unlike [Wang *et al.*, 2017] they do not manually annotate the opinion expressions. Instead they manually add sentiment lexicons and rules based on dependency parsing in order to find the opinion words required to train their system. Using this hand-engineered system, they report state of the art results only for English on the ABSA 2016 dataset. Summarizing, up to the publication of our journal paper there was not a multilingual system that obtained competitive results across the languages included in the ABSA benchmark. This could be due to the complex and language-specific systems that performed best for English [Poria *et al.*, 2016], or perhaps because the CMLA approach of [Wang *et al.*, 2017] would require, in addition to the opinion targets, the gold standard annotations of the opinion expressions for every language in the ABSA datasets.

3 Methodology

The work presented in this research note requires the following resources: (i) Aspect Based Sentiment Analysis (ABSA) data for training and testing; (ii) large unlabelled corpora to obtain semantic distributional features from clustering lexicons; and (iii) a sequence labeling system: ixa-pipe-opinion[2].

Table 1 shows the ABSA datasets from the restaurant domain for English, Spanish, French, Dutch, Russian and Turkish. For English, the size of the 2015 set is less than half with respect to the 2014 dataset in terms of tokens, and only one third in number of targets. The French, Spanish and Dutch datasets are quite similar in terms of tokens although the number of targets in the Dutch dataset is comparatively smaller, possibly due to the tendency to construct compound terms in that language. The Russian dataset is the largest whereas the Turkish set is by far the smallest one.

Apart from the manually annotated data, we also leveraged large, publicly available, unlabeled data to train the clusters: (i) Brown 1000 clusters and (ii) Clark and Word2vec clusters in the 100-800 range. In order to induce clusters from the restaurant domain we used 450M words from the *Yelp Academic Dataset*[3]. For the rest of the languages we used their corresponding Wikipedia dumps[4]. The pre-processing and

[2] https://github.com/ixa-ehu/ixa-pipe-opinion

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

[4] More details in the original paper, Table 2.

	2014			2015			2016		
Features	P	R	F1	P	R	F1	P	R	F1
Local (L)	81.84	74.69	78.10	**76.82**	54.43	63.71	74.41	61.76	67.50
L + BY	77.84	84.57	81.07	71.73	63.65	67.45	**74.49**	71.08	72.74
L + CYF100-CYR200	**82.91**	84.30	83.60	73.25	61.62	66.93	74.12	72.06	73.07
L + W2VW400	76.82	82.10	79.37	74.42	59.04	65.84	73.04	65.52	69.08
L + **ALL**	81.15	**87.30**	**84.11**	72.90	**69.00**	**70.90**	73.33	**73.69**	**73.51**

Table 2: ABSA 2014-2016 English results. BY: Brown Yelp 1000 classes; CYF100-CYR200: Clark Yelp Food 100 classes and Clark Yelp Reviews 200 classes; W2VW400: Word2vec Wikipedia 400 classes; ALL: BY+CYF100-CYR200+W2VW400.

tokenization was performed with the IXA pipes tools [Agerri *et al.*, 2014].

3.1 ixa-pipe-opinion

We adapt for this task the sequence labeler implemented in [Agerri and Rigau, 2016]. By design, the sequence labeller aims to establish a simple and shallow feature set, avoiding any linguistic motivated features, with the objective of removing any reliance on costly extra gold annotations and/or cascading errors across annotations. The system consists of: (i) Local, shallow features based mostly on orthographic, word shape and n-gram features plus their context; and (ii) three types of simple clustering features, based on unigram matching: (i) Brown [Brown *et al.*, 1992] clusters, taking the 4th, 8th, 12th and 20th node in the path; (ii) Clark [Clark, 2003] clusters and, (iii) Word2vec [Mikolov *et al.*, 2013] clusters, based on K-means applied over the extracted word vectors using the skip-gram algorithm.

The clustering features look 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 the class is added as feature. As we work on a 5 token window, for each token and clustering lexicon at least 5 features are generated. For Brown, the number of features generated depend on the number of nodes found in the path for each token and clustering lexicon used. To choose the best combination of clustering features we tried, via 5-fold cross validation on the training set, every possible permutation of the available Clark and Word2vec clustering lexicons obtained from the data sources. Once the best combination of Clark and Word2vec clustering lexicons per data source was found, we tried to combine them with the Brown clusters. The result is a rather simple but very competitive system which is basically based on generating denser, cluster-based word representations for improving its performance[5].

4 Experimental Results

Table 2 presents ixa-pipe-opinion's results for English. We show in bold our best model (ALL) chosen via 5-fold CV on the training data. Moreover, we also show the results of the best models using only one type of clustering feature, namely, the best Brown, Clark and Word2vec models, respectively. The first noteworthy issue is that the same model obtains the best results on the three English datasets. Second, the cluster-based word representations have a huge impact, between 6-7 points in F1 score across the three ABSA datasets. Third, the results show that the combination of clustering features induced from different data sources is crucial.

Table 3 compares our results with previous work. MIN refers to the multi-task learning framework in [Li and Lam, 2017]. CNN-SenticNet is the 7 layer CNN with Amazon word embeddings, POS, linguistic rules based on syntax patterns and SenticNet [Poria *et al.*, 2016]. LSTM is the system proposed by [Liu *et al.*, 2015]. WDEmb refers to [Yin *et al.*, 2016]. RNCRF is a joint model with CRF and a recursive neural network whereas CMLA is the Coupled Multilayer Attentions model described in Section 2, both systems proposed by [Wang *et al.*, 2017]. DLIREC-NLANGP represents the winning systems in 2014 and 2016 [Toh and Wang, 2014; Toh and Su, 2015; Toh and Su, 2016] while the penultimate row refers to our own system as presented in Table 2.

System	2014	2015	2016
MIN$*$	-	-	73.44
CNN-SenticNet	86.20	-	-
CNN-SenticNet$*$	**87.17**	-	-
LSTM	81.15	64.30	-
WDEmb	84.31	69.12	-
WDEmb$*$	84.97	69.73	-
RNCRF	84.05	67.06	-
RNCRF$*$	85.29	70.73	-
DLIREC-NLANGP	84.01	67.11	72.34
ixa-pipe-opinion	84.11	**70.90**	**73.51**
Baseline	47.16	48.06	44.07

Table 3: ABSA SemEval 2014-2016: Comparison of English results to previous work in terms of F1 scores; $*$ refers to models enriched with human-engineered linguistic features.

The results of Table 3 show that our system, despite its simplicity, is highly competitive, obtaining the best results on the 2015 and 2016 datasets and a competitive performance on the 2014 benchmark. In particular, we outperform much more complex and language-specific approaches tuned via language-specific features, such as that of DLIREC-NLANGP. Furthermore, while some of the deep learning approaches (enriched with human-engineered linguistic features) obtain better results on the 2014 data, that is not the case for the 2015 and 2016 benchmarks, where our system outperforms also the MIN and CMLA models (systems which require manually added rules and gold-standard opinion expressions to obtain their best results, as explained in Section 2). In this sense, this means that our system obtains better results than MIN and CMLA by learning the targets independently instead of jointly learning the target and those ex-

[5] More details in Section 3 of original paper

	2014	2015	2016					
System	en	en	en	es	fr	nl	ru	tr
ixa-pipe-opinion	84.11	70.90	73.51	69.92	69.50	66.39	65.53	60.22
XLM-RoBERTa zero-shot	-	-	-	74.16	69.05	69.07	65.16	55.55
XLM-RoBERTa	87.62	75.91	79.18	77.20	78.07	74.90	74.52	68.55

Table 4: Overview of F1 scores and comparison with current state-of-the art cross-lingual systems.

pressions that convey the polarity of the opinion, namely, the opinion expression.

There seems to be also a correlation between the size of the datasets and performance, given that the results on the 2014 data are much higher than those obtained using the 2015 and 2016 datasets. This might be due to the fact that the 2014 training set is substantially larger. In fact, the smaller datasets seem to affect more the deep learning approaches (LSTM, WDEmb, RNCRF) where only the MIN and CMLA models obtain similar results to ours, albeit using manually added language-specific annotations. Finally, it would have been interesting to compare MIN, CNN-SenticNet and CMLA with our system on the three ABSA benchmarks, but their systems are not publicly available.

4.1 Multilingual

We trained ixa-pipe-opinion for 5 other languages on the ABSA 2016 datasets, using the same strategy as for English. We choose the best Clark-Word2vec combination (with and without Brown clusters) via 5-cross validation. The features are the same as those for English, the only change the data used to train the clusters (Wikipedia in this case). Table 5 shows that our system outperforms every previous approach for every language. In some cases, such as Turkish and Russian, previous scores were simply baselines provided by the ABSA organizers, but for the rest our system is still significantly better than previous state-of-the-art. In particular, and despite using the same system for every language, we improve over GTI's submission, which implemented a CRF system with linguistic features specific to Spanish [Álvarez-López *et al.*, 2016].

Language	System	F1
es	GTI	68.51
	L + **CW600 + W2VW300**	**69.92**
	Baseline	51.91
fr	IIT-T	66.67
	L + **CW100**	**69.50**
	Baseline	45.45
nl	IIT-T	56.99
	L + **W2VW400**	**66.39**
	Baseline	50.64
ru	Danii.	33.47
	L + **CW500**	**65.53**
	Baseline	49.31
tr	L + **BW**	**60.22**
	Baseline	41.86

Table 5: ABSA SemEval 2016: Comparison of multilingual results in terms of F1 scores.

The first difference with respect to the English results is that combining clustering features is only beneficial for Spanish. Second, the overall results are lower than those obtained in the 2016 English data. This is probably due to training the clusters on Wikipedia data (as opposed to English for which we used Yelp) which is far from optimal for this task. Thus, it would be expected to obtain better results if domain-specific unlabeled data was used to obtain the cluster-based word representations.

Finally, and as an update to the original journal paper, we have evaluated XLM-RoBERTa[6] in two settings: (i) by fine-tuning in each of the languages and, (ii) in a zero-shot setting, namely, training in English and evaluating in each of the respective languages. XLM-RoBERTa [Conneau *et al.*, 2019] is a system based on the transformer architecture [Devlin *et al.*, 2018] which provides a pre-trained language model trained on 2.5 TB of Common Crawl text. These type of language models allows to build rich representations of text and have enabled improvements across most NLP tasks. Table 4 reports the results for both settings. It can be seen that XLM-RoBERTa obtains huge gains over the previous state-of-the-art results which were reported by our own ixa-pipe-opinion system (except for English 2014). The results also show that for languages other than English, the differences are larger, which probably reflects the non-optimal Wikipedia-based clusters used for those languages.

5 Concluding Remarks

We present a simple and general approach to sequence labeling that, at the time of publication, obtained state-of-the-art results in 7 datasets for 6 languages on the ABSA benchmarks [Pontiki *et al.*, 2016]. We also show that our approach can be straightforwardly adapted to different tasks and domains such as OTE or NER. This is mostly due the cluster-based word representations obtained from large amounts of unlabeled data. Finally, we also include here state-of-the-art results using XLM-RoBERTa, showing the improvements that can be obtained by using multilingual and richer context-based word embeddings and the transformer architecture.

Acknowledgments

This work has been partially funded by the Spanish Ministry of Science, Innovation and Universities (DeepReading RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE) and by *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018* (BigKnowledge). Rodrigo Agerri is funded by the RYC-2017-23647 fellowship and acknowledges the donation of a Titan V GPU by the NVIDIA Corporation.

[6] Fine-tuned over 10 epochs, learning rate 5e5, batch size 32 and max length 128 for every language.

References

[Agerri and Rigau, 2016] Rodrigo Agerri and German Rigau. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82, 2016.

[Agerri *et al.*, 2014] Rodrigo Agerri, Josu Bermudez, and German Rigau. IXA pipeline: Efficient and ready to use multilingual NLP tools. In *LREC*, 2014.

[Álvarez-López *et al.*, 2016] Tamara Álvarez-López, Jonathan Juncal-Martínez, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, and Francisco Javier González-Castaño. GTI at semeval-2016 task 5: Svm and crf for aspect detection and unsupervised aspect-based sentiment analysis. In *SemEval*, 2016.

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

[Cambria *et al.*, 2014] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *AAAI*, 2014.

[Clark, 2003] Alexander Clark. Combining distributional and morphological information for part of speech induction. In *EACL*, 2003.

[Conneau *et al.*, 2019] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv:1911.02116*, 2019.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Jebbara and Cimiano, 2019] Soufian Jebbara and Philipp Cimiano. Zero-shot cross-lingual opinion target extraction. In *NAACL*, pages 2486–2495, 2019.

[Li and Lam, 2017] Xin Li and Wai Lam. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2892, 2017.

[Liu *et al.*, 2015] Pengfei Liu, Shafiq Joty, and Helen Meng. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443. Association for Computational Linguistics, 2015.

[Liu, 2012] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[McAuley and Leskovec, 2013] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

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

[Pang and Lee, 2008] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[Pontiki *et al.*, 2014] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *SemEval*, 2014.

[Pontiki *et al.*, 2015] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *SemEval*, 2015.

[Pontiki *et al.*, 2016] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. Semeval-2016 task 5: Aspect based sentiment analysis. In *SemEval*, 2016.

[Poria *et al.*, 2016] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49, 2016.

[San Vicente *et al.*, 2015] Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. Elixa: A modular and flexible absa platform. In *SemEval*, 2015.

[Toh and Su, 2015] Zhiqiang Toh and Jian Su. Nlangp: Supervised machine learning system for aspect category classification and opinion target extraction. In *SemEval*, 2015.

[Toh and Su, 2016] Zhiqiang Toh and Jian Su. Nlangp at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In *SemEval*, 2016.

[Toh and Wang, 2014] Zhiqiang Toh and Wenting Wang. Dlirec: Aspect term extraction and term polarity classification system. In *SemEval*, 2014.

[Wang *et al.*, 2017] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *AAAI*, pages 3316–3322, 2017.

[Yin *et al.*, 2016] Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. Unsupervised word and dependency path embeddings for aspect term extraction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2979–2985. AAAI Press, 2016.