Artificial Intelligence 268 (2019) 85–95

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Language independent sequence labelling for Opinion Target Extraction

Rodrigo Agerri *, German Rigau

IXA NLP Group, University of the Basque Country UPV/EHU, Donostia-San Sebastián, Spain

ARTICLE INFO

Article history:
Received 6 November 2017
Received in revised form 30 November 2018
Accepted 6 December 2018
Available online 10 December 2018

Keywords:
Opinion target extraction
Aspect based sentiment analysis
Information extraction
Clustering
Semi-supervised learning
Natural language processing

ABSTRACT

In this research note 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 best results for six languages in seven different datasets. Furthermore, the results provide further insights into the behaviour of clustering features for sequence labelling tasks. The system and models generated in this work are available for public use and to facilitate reproducibility of results.

1. Introduction

Opinion Mining and Sentiment Analysis (OMSA) are crucial for determining opinion trends and attitudes about commercial products, companies reputation management, brand monitoring, or to track attitudes by mining social media, etc. Furthermore, given the explosion of information produced and shared via the Internet, especially in social media, it is simply not possible to keep up with the constant flow of new information by manual methods.

Early approaches to OMSA were based on document classification, where the task was to determine the polarity (positive, negative, neutral) of a given document or review [17,21]. A well known benchmark for polarity classification at document level is that of [22]. Later on, a finer-grained OMSA was deemed necessary. This was motivated by the fact that in a given review 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”, “restaurant ambience”, etc. Similarly, if the reviews were about consumer electronics such as laptops, then aspects would include “size”, “battery life”, “hard drive capacity”, etc.

In the review shown by Fig. 1 there are three different opinions about two different aspects (categories) of the restaurant, namely, the first two opinions are about the quality of the food and the third one about the general ambience of the place. Furthermore, there are just two opinion targets because the target of the third opinion, the restaurant itself, remains implicit. Finally, each aspect is assigned a polarity; in this case all three opinion aspects are negative.

* Corresponding author.
E-mail addresses: rodrigo.agerri@ehu.eus (R. Agerri), german.rigau@ehu.eus (G. Rigau).

https://doi.org/10.1016/j.artint.2018.12.002
0004-3702/

```
<sentence id="1016296:4">
    <text>Chow fun was dry; pork shu mai was more than usually greasy and had to
        share a table with loud and rude family</text>
    <Opinions>
        <Opinion target="Chow fun" category="FOOD#QUALITY" polarity="negative"
            from="0" to="8"/>
        <Opinion target="pork shu mai" category="FOOD#QUALITY" polarity="negative"
             from="18" to="30"/>
        <Opinion target="NULL" category="AMBIENCE#GENERAL" polarity="negative"
            from="0" to="0"/>
    </Opinions>
</sentence>
```

Fig. 1. Aspect Based Sentiment Analysis example.

In this work we focus on Opinion Target Extraction, which we model as a sequence labelling task. In order to do so, we convert an annotated review such as the one in Fig. 1 into the BIO scheme for learning sequence labelling models [30]. Example (1) shows the review in BIO format. Tokens in the review are tagged depending on whether they are at the beginning (B-target), inside (I-target) or outside (O) of the opinion target expression. Note that the third opinion target in Fig. 1 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 learn language independent models which consist of a set of local, shallow features complemented with semantic distributional features based on clusters obtained from a variety of data sources. We show that 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 [23–25].

The main contribution of this research note is providing an extension or addendum to previous work on sequence labelling [2] by reporting additional experimental results as well as further insights on the performance of our model across languages on a different NLP task such as Opinion Target Extraction (OTE). Thus, we empirically demonstrate the validity and strong performance of our approach for six languages in seven different datasets of the restaurant domain. Every experiment and result presented in this note is novel.

In this sense, we show that our approach is not only competitive across languages and domains for Named Entity Recognition, as shown by [2], but that it can be straightforwardly adapted to different tasks and domains such as OTE. Furthermore, we release the system and every model trained for public use and to facilitate reproducibility of results.

2. Background

Early approaches to Opinion Target Extraction (OTE) were unsupervised, although later on the vast majority of works have been based on supervised and deep learning models. To the best of our knowledge, the first work on OTE was published by Hu and Liu [9]. They created a new task which consisted of generating overviews of the main product features from a collection of customer reviews on consumer electronics. They addressed such task using an unsupervised algorithm based on association mining. Other early unsupervised approaches include Popescu and Etzioni [26] which used a dependency parser to obtain more opinion targets, and Kim and Hovy [13] which aimed at extracting opinion targets in newswire via Semantic Role Labelling. From a supervised perspective, [36] presented an approach which learned the opinion target candidates and a combination of dependency and part-of-speech (POS) paths connecting such pairs. Their results improved the baseline provided by Hu and Liu [9]. Another influential work was Qiu et al. [28], an unsupervised algorithm called Double Propagation which roughly consists of incrementally augmenting a set of seeds via dependency parsing.

Closer to our work, Jin et al. [12], Li et al. [15] and Jakob and Gurevych [10] approached OTE as a sequence labelling task, modelling the opinion targets using the BIO scheme. The first approach implemented HMM whereas the last two proposed CRFs to solve the problem. In all three cases, their systems included extensive human-designed and linguistically motivated features, such as POS tags, lemmas, dependencies, constituent parsing structure, lexical patterns and semantic features extracted from WordNet [8].

Quite frequently these works used a third party dataset, or a subset of the original one, or created their own annotated data for their experiments. The result was that it was difficult to draw precise conclusions about the advantages or disadvantages of the proposed methods. In this context, the Aspect Based Sentiment Analysis (ABSA) tasks at SemEval [23–25] provided standard training and evaluation data thereby helping to establish a clear benchmark for the OTE task.

Finally, it should be noted that there is a closely related task, namely, the SemEval 2016 task on Stance Detection.[1] Stance detection is related to ABSA, but there is a significant difference. In ABSA the task is to determine whether a piece of text

[1] http://alt.qcri.org/semeval2016/task6/.

```
<sentence id="1016296:4">
    <text>Chow fun was dry; pork shu mai was more than usually greasy and had to
        share a table with loud and rude family</text>
    <Opinions>
        <Opinion target="Chow fun" category="FOOD#QUALITY" polarity="negative"
            from="0" to="8" pfrom=13 pto=16/>
        <Opinion target="pork shu mai" category="FOOD#QUALITY" polarity="negative"
             from="18" to="30" pfrom=53 pto=59/>
        <Opinion target="NULL" category="AMBIENCE#GENERAL" polarity="negative"
            from="0" to="0" pfrom=90 pto=103/>
    </Opinions>
</sentence>
```

Fig. 2. Adding opinion expression annotations to Example (1) in the ABSA 2016 training set.

is positive, negative, or neutral with respect to an aspect and a given target (which in Stance Detection is called "author's favorability" towards a given target). However, in Stance Detection the text may express opinion or sentiment about some other target, not mentioned in the given text, and the targets are predefined, whereas in ABSA the targets are open-ended.

2.1. ABSA tasks at SemEval

Three ABSA editions were held within the SemEval Evaluation Exercises between 2014 and 2016. The ABSA 2014 and 2015 tasks consisted of English reviews only, whereas in the 2016 task 7 more languages were added. Additionally, reviews from four domains were collected for the various sub-tasks across the three editions, namely, Consumer Electronics, Telecommunications, Museums and Restaurant reviews. In any case, 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, for the experiments presented in this paper we decided to focus on the restaurant domain across 6 languages and the three different ABSA editions. Similarly, this section will be focused on reviewing the OTE results for the restaurant domain.

The ABSA task consisted of identifying, for each opinion, the opinion target, the aspect referred to by the opinion and the aspect's polarity. Fig. 1 illustrates the original annotation of a restaurant review in the ABSA 2016 dataset. It should be noted that, out of the three opinion components, only the targets are explicitly represented in the text, which means that OTE can be independently modelled as a sequence labelling problem as shown by Example (1). It is particularly important to notice that the opinion expressions ("dry", "greasy", "loud and rude") are not annotated.

Following previous approaches, the first competitive systems for OTE at ABSA were supervised. Among the participants (for English) in the three editions, one team [31,33] was particularly successful. For ABSA 2014 and 2015 they developed a CRF system with extensive handcrafted linguistic features: POS, head word, dependency relations, WordNet relations, gazetteers and Name Lists based on applying the Double Propagation algorithm [28] on an initial list of 551 seeds. Interestingly, they also introduced word representation features based on Brown and K-mean clusters. For ABSA 2016, they improved their system by using the output of a Recurrent Neural Network (RNN) to provide additional features. The RNN is trained on the following input features: word embeddings, Name Lists and word clusters [32]. They were the best system in 2014 and 2016. In 2015 they obtained the second best result, in which the best system, a preliminary version of the one presented in this note, was submitted by the EliXa team [29].

From 2015 onwards most works have been based on deep learning. Liu et al. [18] applied RNNs on top of a variety of pre-trained word embeddings, while Jebbara and Cimiano [11] presented an architecture in which a RNN based tagger is stacked on top of the features generated by a Convolutional Neural Network (CNN). These systems were evaluated on the 2014 and 2015 datasets, respectively, but they did not go beyond the state-of-the-art.

Poria et al. [27] presented a 7 layer deep CNN combining word embeddings trained on a 5 billion word corpus extracted from Amazon [19], POS tag features and manually developed linguistic patterns based on syntactic analysis and SenticNet [5] a concept-level knowledge based build for Sentiment Analysis applications. They only evaluate their system on the English 2014 ABSA data, obtaining best results up to date on that benchmark.

More recently, Wang et al. [34] proposed a coupled multi-layer attention (CMLA) network where each layer consists of a couple of attentions with tensor operators. Unlike previous approaches, their system does not use complex linguistic-based features designed for one specific language. However, whereas 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 ABSA training and testing data with the required opinion expressions. Although Wang et al. [34] did not release the datasets with the annotated opinion expressions, Fig. 2 illustrates what these annotations would look like. Thus, two new attributes (`pfrom` and `pto`) annotate the opinion expressions for each of the three opinions ("dry", "greasy" and "loud and rude", respectively). 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 [16] develop a multi-task learning framework consisting of two LSTMs equipped with extended memories and neural memory operations. As Wang et al. [34], they use opinion expressions annotations for a joint modelling of opinion targets and expressions. However, unlike Wang et al. [34] they do not manually annotate the opinion expressions.

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.

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

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. They do not provide evaluation results on the 2014 and 2015 restaurant datasets.

With respect to other languages, the IIT-T team presented systems for 4 out of the 7 languages in ABSA 2016, obtaining the best score for French and Dutch, second in Spanish but with very poor results for English, well below the baseline. The GTI team [3] implemented a CRF system using POS, lemmas and bigrams as features. They obtained the best result for Spanish and rather modest results for English.

Summarizing, the most successful systems for OTE have been based on supervised approaches with rather elaborate, complex and linguistically inspired features. Poria et al. [27] obtain best results on the ABSA 2014 data by means of a CNN with word embeddings trained on 5 billion words from Amazon, POS features, manual patterns based on syntactic analysis and SenticNet. More recently, the CMLA deep learning model has established new state-of-the-art results for the 2015 dataset, whereas Li and Lam [16] provide the state of the art for the 2016 benchmark. Thus, there is not currently a multilingual system that obtains competitive results across (at least) several of the languages included in ABSA.

As usual, most of the work has been done for English, with the large majority of the previous systems providing results only for one of the three English ABSA editions and without exploring the multilingual aspect. This could be due to the complex and language-specific systems that performed best for English [27], or perhaps because the CMLA approach of Wang et al. [34] would require, in addition to the opinion targets, the gold standard annotations of the opinion expressions for each of the 6 languages other than English 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 labelling system. In this section we will describe each of the resources used.

3.1. ABSA datasets

Table 1 shows the ABSA datasets from the restaurants domain for English, Spanish, French, Dutch, Russian and Turkish. From left to right each row displays the number of tokens, number of targets and the number of multiword targets for each training and test set. For English, it should be noted that 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.

Additionally, we think it is also interesting to note the low number of targets that are multiwords. To provide a couple of examples, for Spanish only the %35.59 of the targets are multiwords whereas for Dutch the percentage goes down to %25.68. If we compare these numbers with the CoNLL 2002 data for Named Entity Recognition (NER), a classic sequence labelling task, we find that in the ABSA data there is less than half the number of multiword targets than the number of multiword entities that can be found in the CoNLL Spanish and Dutch data (%35.59 vs %74.33 for Spanish and %25.68 vs %44.96 for Dutch).

3.2. Unlabelled corpora

Apart from the manually annotated data, we also leveraged large, publicly available, unlabelled data to train the clusters: (i) Brown 1000 clusters and (ii) Clark and Word2vec clusters in the 100–800 range.

Table 2
Unlabelled corpora to induce clusters. For each corpus and cluster type the number of words (in millions) is specified. Average training times: depending on the number of words, Brown clusters training time required between 5 h and 48 h. Word2vec required 1–4 hours whereas Clark clusters training lasted between 5 hours and 10 days.

	Million words in corpus		Million words for training		
			Brown	Clark	Word2vec
en	Yelp Academic Dataset	225	156	225	225
	Yelp food	117	82	117	117
	Yelp food-hotels	102	73	102	102
	Wikipedia (20141208)	1700	790	790	1700
es	Wikipedia (20140810)	428	246	246	428
fr	Wikipedia (20140804)	547	280	280	547
nl	Wikipedia (20140804)	235	128	128	235
ru	Wikipedia (20140727)	338	158	158	338
tr	Wikipedia (20140806)	48	33	48	48

In order to induce clusters from the restaurant domain we used the *Yelp Academic Dataset*,[2] from which three versions were created. First, the full dataset, containing 225M tokens. Second, a subset consisting of filtering out those categories that do not correspond directly to food related reviews [14]. Thus, out of the 720 categories contained in the Yelp Academic Dataset, we kept the reviews from 173 of them. This *Yelp food* dataset contained 117M tokens in 997,721 reviews. Finally, we removed two more categories (Hotels and Hotels & Travel) from the *Yelp food* dataset to create the *Yelp food-hotels* subset containing around 102M tokens. For the rest of the languages we used their corresponding Wikipedia dumps. The pre-processing and tokenization is performed with the IXA pipes tools [1].

The number of words used for each dataset, language and cluster type are described in Table 2. For example, the first row reads "Yelp Academic Dataset containing 225M words was used; after pre-processing, 156M words were taken to induce Brown clusters, whereas Clark and Word2vec clusters were trained on the whole corpus". As explained in [2], we pre-process the corpus before training Brown clusters, resulting in a smaller dataset than the original. Additionally, due to efficiency reasons, when the corpus is too large we use the pre-processed version to induce the Clark clusters.

3.3. System

We use the sequence labeller implemented within IXA pipes [2]. It learns supervised models based on the Perceptron algorithm [7]. To avoid duplication of efforts, it uses the Apache OpenNLP project implementation[3] customized with its own features. 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 [4] clusters, taking the 4th, 8th, 12th and 20th node in the path; (ii) Clark [6] clusters and, (iii) Word2vec [20] 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 ("not found" otherwise). 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.

Fig. 3 depicts how our system relates, via clusters, unseen words with those words that have been seen as targets during the training process. Thus, the tokens 'french-onions' and 'salmon' would be annotated as opinion targets because they occur in the same clusters as seen words which in the training data are labelled as targets.

The word representation features are *combined* and *stacked* using the clustering lexicons induced over the different data sources listed in Table 2. In other words, *stacking* means adding various clustering features of the same type obtained from different data sources (for example, using clusters trained on Yelp and on Wikipedia); *combining* refers to combining different types of clustering features obtained from the same data source (e.g., using features from Brown and Clark clustering lexicons).

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 that has proven to be highly successful in the most popular Named Entity Recognition and Classification (NER) benchmarks, both in out-of-domain and in-domain evaluations.

[2] http://www.yelp.com/dataset_challenge.
[3] http://opennlp.apache.org/.

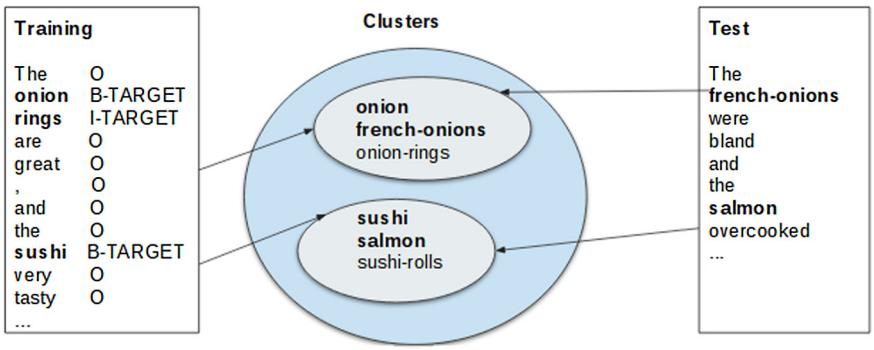

Fig. 3. Unigram matching in clustering features.

Table 3
ABSA SemEval 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.

Features	2014			2015			2016		
	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**

Furthermore, it was demonstrated that the system also performed robustly across languages without any language-specific tuning. Details of the system's implementation, including detailed description of the local and clustering features, can be found in [2],[4] including a section on how to combine the clustering features.

A preliminary version of this system [29] was the winner of the OTE sub-task in the ABSA 2015 edition (English only). In the next section we show that this system obtains state-of-the-art results not only across domains and languages for NER, but also for other tasks such as Opinion Target Extraction. The results reported are obtained using the official ABSA evaluation scripts [23–25].

4. Experimental results

In this section we report on the experiments performed using the system and data described above. First we will present the English results for the three ABSA editions as well as a comparison with previous work. After that we will do the same for 5 additional languages included in the ABSA 2016 edition: Dutch, French, Russian, Spanish and Turkish. The local and clustering features, as described in Section 3.3, are the same for every language and evaluation setting. The only change is the clustering lexicons used for the different languages. As stated in section 3.3, the best cluster combination is chosen via 5-fold cross validation (CV) on the training data. We first try every permutation with the Clark and Word2vec clusters. Once the best combination is obtained, we then try with the Brown clusters obtaining thus the final model for each language and dataset.

4.1. English

Table 3 provides detailed results on the Opinion Target Extraction (OTE) task 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, it is also interesting to note the huge gains obtained by the clustering features, 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. Fourth, the clustering features improve the recall by 12–15 points in the 2015 and 2016 data, and around 7 points for 2014. Finally, while in 2014 the precision also increases, in the 2015 setting it degrades almost by 4 points in F1 score.

Table 4 compares our results with previous work. MIN refers to the multi-task learning framework consisting of two LSTMs equipped with extended memories and neural memory operations with manually developed rules for detecting

[4] Table 3 and pages 68–71.

Table 4
ABSA SemEval 2014–2016: Comparison of English results in terms of F1 scores; ∗ refers to models enriched with human-engineered linguistic features.

System	ABSA 2014	ABSA 2015	ABSA 2016
MIN∗ [16]	–	–	73.44
CNN-SenticNet [27]	86.20	–	–
CNN-SenticNet∗ [27]	**87.17**	–	–
LSTM [18]	81.15	64.30	–
WDEmb [35]	84.31	69.12	–
WDEmb∗ [35]	84.97	69.73	–
RNCRF [34]	84.05	67.06	–
RNCRF∗ [34]	85.29	70.73	–
DLIREC-NLANGP [31–33]	84.01	67.11	72.34
BY+CYF100-CYR200+W2VW400	84.11	**70.90**	**73.51**
Baseline	47.16	48.06	44.07

opinion expressions [16]. CNN-SenticNet is the 7 layer CNN with Amazon word embeddings, POS, linguistic rules based on syntax patterns and SenticNet [27].

LSTM is a Long Short Term Memory neural network built on top of word embeddings as proposed by Liu et al. [18]. WDEmb [35] uses word and dependency path, linear context and dependency context embedding features the input to a CRF. RNCRF is a joint model with CRF and a recursive neural network whereas CMLA is the Coupled Multilayer Attentions model described in section 2.1, both systems proposed by Wang et al. [34]. DLIREC-NLANGP is the winning system at ABSA 2014 and 2016 [31–33] while the penultimate row refers to our own system for all the three benchmarks (details in Table 3).

The results of Table 4 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 the deep learning approaches (enriched with human-engineered linguistic features) obtain comparable or 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 2.1). 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 expressions 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, as detailed in Table 1. 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.

4.2. Multilingual

We trained our system 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 on the training data. The features are exactly the same as those used for English, the only change is the data on which the clusters are trained. Table 5 reports on the detailed results obtained for each of the languages. In bold we show the best model chosen via 5-fold CV. Moreover, we also show the best models using only one of each of the clustering features.

The first difference with respect to the English results is that the Brown clustering features are, in three out of five settings, detrimental to performance. Second, that combining clustering features is only beneficial for Spanish. Third, the overall results are in general lower than those obtained in the 2016 English data. Finally, the difference between the best results and the results using the Local features is lower than for English, even though the Local results are similar to those obtained with the English datasets (except for Turkish, but this is due to the significantly smaller size of the data, as shown in Table 1).

We believe that all these four issues are caused, at least partially, by the lack of domain-specific clustering features used for the multilingual experiments. In other words, while for the English experiments we leveraged the Yelp dataset to train the clustering algorithms, in the multilingual setting we first tried with already available clusters induced from the Wikipedia. Thus, it is to be expected that the gains obtained by clustering features obtained from domain-specific data such as Yelp would be superior to those achieved by the clusters trained on out-of-domain data.

In spite of this, Table 6 shows that our system outperforms the best previous approaches across the five languages. In some cases, such as Turkish and Russian, the best previous scores were the baselines provided by the ABSA organizers, but for Dutch, French and Spanish our system is significantly better than current state-of-the-art. In particular, and despite using

Table 5
ABSA SemEval 2016 multilingual results.

Language	Features	Precision	Recall	F1
es	Local (L)	79.17	59.19	67.74
	L + BW	67.96	63.67	65.75
	L + CW600	73.22	64.80	68.75
	L + W2VW300	75.50	63.53	69.00
	L + **CW600** + **W2VW300**	75.36	65.22	**69.92**
fr	Local (L)	66.92	66.41	66.67
	L + BW	63.39	72.46	67.62
	L + **CW100**	69.94	69.08	**69.50**
	L + W2VW100	66.52	68.77	67.62
nl	Local (L)	73.14	55.50	63.11
	L + BW	68.59	57.37	62.48
	L + CW100	66.94	65.15	66.03
	L + **W2VW400**	68.27	64.61	**66.39**
ru	Local (L)	64.87	61.87	63.33
	L + BW	61.32	64.60	62.92
	L + **CW500**	64.21	66.91	**65.53**
	L + W2VW700	64.41	64.81	64.61
tr	Local (L)	56.82	51.72	54.15
	L + **BW**	62.69	57.93	**60.22**
	L + CW200	58.28	60.69	59.46
	L + W2VW300	59.09	53.79	56.32

Table 6
ABSA SemEval 2016: Comparison of multilingual results in terms of F1 scores.

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

the same system for every language, we improve over GTI's submission, which implemented a CRF system with linguistic features specific to Spanish [3].

5. Discussion and error analysis

Considering the simplicity of our approach, we obtain best results for 6 languages and 7 different settings in the Opinion Target Extraction (OTE) benchmark for the restaurant domain using the ABSA 2014–2016 datasets.

These results are obtained without linguistic or manually-engineered features, relying on injecting external knowledge from the combination of clustering features to obtain a robust system across languages, outperforming other more complex and language-specific systems. Furthermore, the feature set used is the same for every setting, reducing human intervention to a minimum and establishing a clear methodology for a fast and easy creation of competitive OTE multilingual taggers.

The results also confirm the behaviour of these clustering algorithms to provide features for sequence labelling tasks such as OTE and Named Entity Recognition (NER), as previously discussed in [2]. Thus, in every evaluation setting the best results using Brown clusters as features were obtained when data close to the application domain and text genre, even if relatively small, was used to train the Brown algorithm. This can be clearly seen if we compare the English with the multilingual results. For English, the models including Brown clusters improve the Local features over 3–5 points in F1 score, whereas for Spanish, Dutch and Russian, they worsen performance. The reason is that for English the Yelp dataset is used whereas

Table 7
False Positives and Negatives for every ABSA 2014–2016 setting.

Error type	2014	2015	2016					
	en	en	en	es	fr	nl	ru	tr
FP	**230**	151	**189**	165	194	117	**390**	62
FN	143	**169**	163	**248**	**202**	**132**	312	**65**

Table 8
Top five false positive (FP) and negative (FN) errors for English, Spanish and French.

	2014		2015		2016					
	en		en		en		es		fr	
FP	place	21	place	16	place	16	comida	11	restaurant	13
	money	6	food	6	food	16	restaurante	10	cuisine	9
	spot	4	waitress	4	restaurant	11	atención	7	terrasse	8
	pizza	3	chicken	4	service	7	platos	6	repas	7
	sushi	3	salmon	3	wait	3	servicio	4	plats	6
FN	place	4	restaurant	8	place	7	restaurante	12	restaurant	5
	food	3	place	7	sushi	3	platos	7	cuisine	5
	waiting	2	food	5	restaurant	3	trato	6	carte	5
	taste	2	Casa La Femme	4	Ray's	3	comida	6	plats	4
	selection	2	The Four Seasons	3	menu	3	carta	6	table	3

for the rest of languages the clusters are induced using the Wikipedia, effectively an out-of-domain corpus. The exception is Turkish, for which a 6 point gain in F1 score is obtained, but we believe that is probably due to the small size of the training data used for training the Local model.

In contrast, Word2vec clusters clearly benefit from larger amounts of data, as illustrated by the best English Word2vec model being the one trained using the Wikipedia, and not the Yelp dataset, which is closer to the application domain. Finally, the Clark algorithm seems to be the most versatile as it consistently outperforms the other two clustering methods in 4 out of the 8 evaluation settings presented.

Summarizing: (i) Brown clusters perform better when leveraged from source data close to the application domain, even if small in size; (ii) Clark clusters are the most robust of the three with respect to the size and domain of the data used; and (iii) for Word2vec size is the crucial factor. The larger the source data the better the performance. Thus, instead of choosing over one clustering type or the other, our system provides a method to effectively combining them, depending on the data sources available, to obtain robust and language independent sequence labelling systems.

Finally, results show that our models are particularly competitive when the amount of training data available is small, allowing us to compete with more complex systems including also manually-engineered features, as shown especially by the English results on the 2015 and 2016 data.

5.1. Error analysis

We will now discuss the shortcomings and most common errors performed by our system for the OTE task. By looking at the overall results in terms of *precision* and *recall*, it is possible to see the following patterns: With respect to the Local models, precision is consistently better than recall or, in other words, the coverage of the Local models is quite low. Tables 3 and 5 show that adding clustering features to the Local models allows to improve the recall for every evaluation setting, although with different outcomes. Overall, precision suffers, except for French.[5] Furthermore, in three cases (English 2014, 2016 and Russian) precision is lower than recall, whereas the remaining 5 evaluations show that, despite large improvements in F1 score, most errors in our system are caused by *false negatives*, as it can be seen in Table 7.

Table 8 displays the top 5 most common false positives and false negative errors for English, Spanish and French.[6] By inspecting our system's output, and both the test and training sets, we found out that there were three main sources of errors: (a) errors caused by ambiguity in the use of certain source forms that may or may not refer to an opinion target; (b) span errors, where the target has only been partially annotated; and (c) unknown targets, which the system was unable to annotate by generalizing on the training data or clusters.

With respect to type (a), it is useful to look at the most common errors for all three languages, namely, 'place', 'food' and 'restaurant', which are also among the top 5 most frequent targets in the gold standard sets. By looking at Examples (1–3) we would say that in all three cases 'place' should be annotated as opinion target. However, (2) is a false positive (FP), (3) is a false negative (FN) and (1) is an example from the training set in which 'place' is annotated as target. This is the case

[5] It also goes up for Turkish, but as already commented, we believe that due to the small size of the Turkish training set, clustering features allow to improve both precision and recall.

[6] According to the authors' knowledge of languages to comment on specific examples from the data.

with many instances of 'place' for which there seems to be some inconsistency in the actual annotation of the training and test set examples.[7]

Example (1): Avoid this place!
Example (2): this place is a keeper!
Example (3): it is great place to watch sporting events.

For other frequent type (a) errors, ambiguity is the main problem. Thus, in Spanish the use of 'comida'[8] and 'restaurante'[9] is highly ambiguous and causes many FPs and FNs because sometimes it is actually an opinion target whereas in many other cases it is just referring to the meal or the restaurant themselves without expressing any opinion about them. The same phenomenon occurs for "food" and "restaurant" in English and for 'cuisine' and 'restaurant' in French.

Span type (b) errors are typically caused by long opinion targets such as "filet mignon on top of spinach and mashed potatoes" for which our system annotates "filet" and "spinach" as separate targets, or "chicken curry and chicken tikka masala" which is wrongly tagged as one target. These cases are difficult because on the surface they look similar but the first one refers to one dish only, hence one target, whereas the second one refers to two separate dishes for which two different opinion targets should be annotated. Of course, these cases are particularly hurtful because they count as both FP and FN.

Finally, type (c) errors are usually caused by lack of generalization of our system to deal with unknown targets. Example (4–7) contain various mentions to the "Ray's" restaurant, which is in the top 5 errors for the English 2016 test set.

Example (4): After 12 years in Seattle Ray's rates as the place we always go back to.
Example (5): We were only in Seattle for one night and I'm so glad we picked Rays for dinner!
Example (6): I love Dungeness crabs and at Ray's you can get them served in about 6 different ways!
Example (7): Imagine my happy surprise upon finding that the views are only the third-best thing about Ray's!
Example (8): Ray's is something of a Seattle institution

Examples (4), (5) and (7) are FNs, (6) is a FP caused by wrongly identifying the target as "Ray's you", whereas (8) is not event annotated in the gold standard or by our system, although it should had been.

6. Concluding remarks

In this research note we provide additional empirical experimentation to [2], reporting best results for Opinion Target Extraction for 6 languages and 7 datasets using the same set of simple, shallow and language independent features. Furthermore, the results provide some interesting insights with respect to the use of clusters to inject external knowledge via semi-supervised features.

First, Brown clusters are particularly beneficial when trained on domain-related data. This seems to be the case in the multilingual setting, where the Brown clusters (trained on out-of-domain Wikipedia data) worsen the system's performance for every language except for Turkish.

Second, the results also show that Clark and Word2vec improve results in general, even if induced on out-of-domain data. Thirdly, for best performance it is convenient to combine clusters obtained from diverse data sources, both from in- and out-of-domain corpora.

Finally, the results indicate that, even when the amount of training data is small, such as in the 2015 and 2016 English benchmarks, our system's performance remains competitive thanks to the combination of clustering features. This, together with the lack of linguistic features, facilitates the easy and fast development of systems for new domains or languages. These considerations thus confirm the hypotheses stated in [2] with respect to the use of clustering features to obtain robust sequence taggers across languages and tasks.

The system and models for every language and dataset are available as part of the *ixa-pipe-opinion* module for public use and reproducibility of results.[10]

Acknowledgements

First, we would like to thank the anonymous reviewers for their comments to improve the paper. We would also like to thank Iñaki San Vicente for his help obtaining the Yelp data. This work has been supported by the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE), under the projects TUNER (TIN2015-65308-C5-1-R) and CROSSTEXT (TIN2015-72646-EXP).

[7] Interannotator agreement (91% F1) was only reported for a small subset of the Spanish data.
[8] In English: "food" or "meal", depending on the context.
[9] In English: "restaurant".
[10] https://github.com/ixa-ehu/ixa-pipe-opinion.

References

[1] R. Agerri, J. Bermudez, G. Rigau, IXA pipeline: efficient and ready to use multilingual NLP tools, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 2014, pp. 3823–3828.
[2] R. Agerri, G. Rigau, Robust multilingual named entity recognition with shallow semi-supervised features, Artif. Intell. 238 (2016) 63–82.
[3] T. Àlvarez-López, J. Juncal-Martínez, M. Fernández-Gavilanes, E. Costa-Montenegro, F.J. González-Castaño, Gti at semeval-2016 task 5: svm and crf for aspect detection and unsupervised aspect-based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, 2016, pp. 306–311.
[4] P.F. Brown, P.V. Desouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai, Class-based n-gram models of natural language, Comput. Linguist. 18 (1992) 467–479.
[5] E. Cambria, D. Olsher, D. Rajagopal, Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis, in: Twenty-eighth AAAI Conference on Artificial Intelligence, 2014.
[6] A. Clark, 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 – vol. 1, Association for Computational Linguistics, 2003, pp. 59–66.
[7] M. Collins, 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 – vol. 10, 2002, pp. 1–8.
[8] C. Fellbaum, G. Miller (Eds.), Wordnet: An Electronic Lexical Database, MIT Press, Cambridge (MA), 1998.
[9] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.
[10] N. Jakob, I. Gurevych, Extracting opinion targets in a single-and cross-domain setting with conditional random fields, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010, pp. 1035–1045.
[11] S. Jebbara, P. Cimiano, Aspect-based relational sentiment analysis using a stacked neural network architecture, in: ECAI 2016 – 22nd European Conference on Artificial Intelligence, 29 August–2 September 2016, The Hague, The Netherlands – Including Prestigious Applications of Artificial Intelligence (PAIS 2016), 2016, pp. 1123–1131.
[12] W. Jin, H.H. Ho, R.K. Srihari, A novel lexicalized hmm-based learning framework for web opinion mining, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 465–472.
[13] S.M. Kim, E. Hovy, Extracting opinions, opinion holders, and topics expressed in online news media text, in: Proceedings of the Workshop on Sentiment and Subjectivity in Text, Association for Computational Linguistics, 2006, pp. 1–8.
[14] S. Kiritchenko, X. Zhu, C. Cherry, S. Mohammad, NRC-Canada-2014: detecting aspects and sentiment in customer reviews, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 2014, pp. 437–442.
[15] F. Li, C. Han, M. Huang, X. Zhu, Y.J. Xia, S. Zhang, H. Yu, Structure-aware review mining and summarization, in: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 2010, pp. 653–661.
[16] X. Li, W. 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, 2017, pp. 2886–2892.
[17] B. Liu, Sentiment analysis and opinion mining, Synth. Lect. Hum. Lang. Technol., vol. 5, 2012, pp. 1–167.
[18] P. Liu, S. Joty, H. 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, Association for Computational Linguistics, 2015, pp. 1433–1443.
[19] J. McAuley, J. Leskovec, Hidden factors and hidden topics: understanding rating dimensions with review text, in: Proceedings of the 7th ACM Conference on Recommender Systems, ACM, 2013, pp. 165–172.
[20] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[21] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retr. 2 (2008) 1–135.
[22] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? sentiment classification using machine learning techniques, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2002, pp. 79–86.
[23] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S.M. Jiménez-Zafra, G. Eryiğit, Semeval-2016 task 5: aspect based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 19–30.
[24] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, Semeval-2015 task 12: aspect based sentiment analysis, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 486–495.
[25] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, Semeval-2014 task 4: aspect based sentiment analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 2014, pp. 27–35.
[26] A.M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 339–346.
[27] S. Poria, E. Cambria, A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network, Knowl.-Based Syst. 108 (2016) 42–49.
[28] G. Qiu, B. Liu, J. Bu, C. Chen, Opinion word expansion and target extraction through double propagation, Comput. Linguist. 37 (2011) 9–27.
[29] I.N. San Vicente, X. Saralegi, R. Agerri, Elixa: a modular and flexible absa platform, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 748–752.
[30] E.F. Tjong Kim Sang, Introduction to the CoNLL-2002 shared task: language-independent named entity recognition, in: Proceedings of CoNLL-2002, Taipei, Taiwan, 2002, pp. 155–158.
[31] Z. Toh, J. Su, Nlangp: supervised machine learning system for aspect category classification and opinion target extraction, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, 2015, pp. 496–501.
[32] Z. Toh, J. Su, Nlangp at semeval-2016 task 5: improving aspect based sentiment analysis using neural network features, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016, pp. 282–288.
[33] Z. Toh, W. Wang, Dlirec: aspect term extraction and term polarity classification system, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 235–240.
[34] W. Wang, S.J. Pan, D. Dahlmeier, X. Xiao, Coupled multi-layer attentions for co-extraction of aspect and opinion terms, in: AAAI, 2017, pp. 3316–3322.
[35] Y. Yin, F. Wei, L. Dong, K. Xu, M. Zhang, M. Zhou, Unsupervised word and dependency path embeddings for aspect term extraction, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press, 2016, pp. 2979–2985.
[36] L. Zhuang, F. Jing, X.Y. Zhu, Movie review mining and summarization, in: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, ACM, 2006, pp. 43–50.