Unsupervised Aspect Term Extraction with B-LSTM & CRF using Automatically Labelled Datasets

Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann
and **Michael Baeriswyl**
Artificial Intelligence and Machine Learning Group — Swisscom AG
`firstName.lastName@swisscom.com`

Abstract

Aspect Term Extraction (ATE) identifies opinionated aspect terms in texts and is one of the tasks in the SemEval Aspect Based Sentiment Analysis (ABSA) contest. The small amount of available datasets for supervised ATE and the costly human annotation for aspect term labelling give rise to the need for unsupervised ATE. In this paper, we introduce an architecture that achieves top-ranking performance for supervised ATE. Moreover, it can be used efficiently as feature extractor and classifier for unsupervised ATE. Our second contribution is a method to automatically construct datasets for ATE. We train a classifier on our automatically labelled datasets and evaluate it on the human annotated SemEval ABSA test sets. Compared to a strong rule-based baseline, we obtain a dramatically higher *F-score* and attain precision values above 80%. Our unsupervised method beats the supervised ABSA baseline from SemEval, while preserving high precision scores.

1 Introduction

For many years now, companies are offering users the possibility of adding reviews in the form of sentences or small paragraphs. Reviews can be beneficial for both customers and companies. On the one hand, people can make better decisions by getting insights about available products and solutions. One the other hand, companies are interested in understanding how and what customers think about their products, which helps in employing marketing solutions and correction strategies. To this end, performing an automated analysis of the user opinions becomes a crucial issue.

Performing sentiment analysis to detect the overall polarity of a sentence or paragraph comes with two major disadvantages. First, sentiment analysis on sentence (or paragraph) level does not fulfill the purpose of getting more accurate and precise information. The polarity refers to a broader context, instead of pinpointing specific targets. Secondly, many sentences or paragraphs contain opposing polarities towards distinct targets, making it impossible to assign an accurate overall polarity.

The need for identifying aspect terms and their respective polarity gave rise to the Aspect Based Sentiment Analysis (ABSA), where the task is first to extract aspects or features of an entity (i.e. Aspect Term Extraction or ATE[1]) from a given text, and second to determine the sentiment polarity (SP), if any, towards each aspect of that entity. The importance of ABSA led to the creation of the ABSA task in the SemEval[2] contest in 2014 ([Pontiki et al., 2014](#)).

Supervised ATE using human annotated datasets leads to high performance for aspect term detection on unseen data, however it has two major drawbacks. First, the size of the labelled datasets is quite small, reducing the performance of the classifiers. Second, human annotation is a very slow and costly procedure.

The drawbacks of supervised ATE can be tackled using unsupervised ATE. The size of the datasets can be significantly increased using targets from publicly available reviews (e.g. *Amazon* or *Yelp*). Reviews are opinion texts and contain plenty of opinionated aspect terms, which makes them perfect candidates for constructing new datasets for ATE. With respect to the second drawback, an au-

---

[1]Also known as Opinion Term Extraction (OTE).

[2]The SemEval (Semantic Evaluation) contest is an ongoing series of evaluations of computational semantic analysis systems.

tomated data labelling process with high precision can replace the slow and error-prone human annotation procedure.

We innovate by performing ATE starting from opinion texts (e.g. reviews). This is a completely unsupervised task since there are no labels to explicitly pinpoint that certain tokens of the text are aspect terms. Reviews may contain labels (e.g. number of stars in a 1-5 star rating system) which are related to their overall polarity. However, such labels are not useful for ATE.

In this work, we present a classifier, which can be used for feature extraction and aspect term detection for both unsupervised and supervised ATE. We validate its suitability for ATE by achieving top-ranking results for supervised ATE using the SemEval-2014 ABSA task datasets[3]. Then, we use it for unsupervised ATE.

Moreover, we contribute by introducing a new, completely automated, unsupervised and domain independent method for annotating raw opinion texts. Then, we use our classifier to perform unsupervised ATE by training it on the automatically labelled datasets obtained with our method. Against all expectations, our unsupervised method beats the supervised ABSA baseline from SemEval-2014, while achieving high precision scores. The latter is very important for unsupervised techniques since we wish to extract non-noisy aspect terms, i.e. minimize the number of false positives.

The rest of this paper is organized as follows. Section 2 presents the related work for ATE. Our approach for supervised and unsupervised ATE is described in Sections 3 and 4 respectively. Section 5 presents our experiments and results for both supervised and unsupervised ATE. Finally, Section 6 focuses on our conclusions and future work.

2 Related Work

Research in the area of both supervised and unsupervised ATE has flourished after the creation of the SemEval ABSA task in 2014. The winners of the SemEval-2014 ABSA contest (Toh and Wang, 2014) use supervised methods for ATE. They extract features, similar to those used in traditional Name Entity Recognition (NER) systems (Tkachenko and Simanovsky, 2012) using

the provided training sets. Moreover, they exploit external sources, such as the WordNet lexicographer files (Miller, 1995) and word clusters (e.g. Brown clusters (Turian et al., 2010) or K-means[4]). Toh and Su (2015) suggest using gazetteers (Kazama and Torisawa, 2008) and word embeddings (Mikolov et al., 2013) for ATE. Toh and Su (2016) use the probability output of an Recurrent Neural Network (RNN) to further enrich the feature space.

Independently of the feature extraction techniques, supervised ATE is treated as a sequential labelling task. Top-ranking participants in the SemEval ABSA contest use Conditional Random Fields (CRF) or Support Vector Machine (SVM) as sequential labelling classifiers (Toh and Wang, 2014; Toh and Su, 2015; Chernyshevich, 2014; Brun et al., 2014).

There is also related work with respect to unsupervised ATE. Liu et al. (2015) exploit syntactic rules to automatically detect aspect terms. (Garcia-Pablos et al., 2015; Garcia-Pablos and Rigau, 2014) use a graph representation to describe the interactions between aspect terms and opinion words in raw text. Graph nodes are ranked using PageRank and high-ranked nodes are used to create a set of aspect terms. Then, they use this set to annotate unseen data by simply performing exact or lemma matching.

Systems similar to (Hercig et al., 2016; Yin et al., 2016; Soujanya et al., 2016) perform semi-supervised ATE since they use human annotated datasets for training but enrich their feature space using features extracted by exploiting large unlabelled corpora. Pavlopoulos and Androutsopoulos (2015) present a method for constructing new datasets for ATE, however they use non-standard evaluation metrics. Finally, systems like (Garcia-Pablos et al., 2017) focus on classifying the aspect terms into categories. We do not compare against such systems, since they do not perform the same task and are not equivalent to ours.

In all but one[5] aforementioned cases, the evaluation of the model is performed using the *F-score*, as defined in (Pontiki et al., 2014). In case of unsupervised ATE, achieving higher precision is more important than higher recall as highlighted in (Liu et al., 2015).

---

[3]The SemEval ABSA datasets contain human annotation for ATE for both the laptop and the restaurant domains only in 2014.

[4]https://en.wikipedia.org/wiki/K-means_clustering

[5]Pavlopoulos and Androutsopoulos (2015) use a non-standard definition of precision and recall.

We perform both supervised and unsupervised ATE using a model that utilizes continuous word representations and performs feature extraction and sequential labelling simultaneously while training. In case of supervised ATE, the training datasets are those of the SemEval ABSA task (human annotated). In case of unsupervised ATE, we annotate raw opinion texts (e.g. reviews) with a completely automated and unsupervised process, which we introduce. To the best of our knowledge, we are the first to train a classifier using an automatically labelled dataset and perform evaluation on human annotated datasets.

3 Supervised Aspect Term Extraction

The ATE task can be modelled as a token-based classification task, where labels are assigned to the tokens of a sequence, depending on whether they are aspect terms or not. For supervised ATE, we apply a classification pipeline that consists of 3 steps: (i) data preprocessing, (ii) model training and (iii) model evaluation.

The feature extraction is performed from a two-layer bidirectional long short-term memory (B-LSTM) network while the model is training, similar to the way a Convolutional Neural Network (CNN) extracts features while performing image classification. Therefore, we do not explicitly include this step in the aforementioned pipeline.

3.1 Data Preprocessing

We break down each sentence into tokens using the spaCy parser[6] and follow the traditional IOB format (short for Inside, Outside, Beginning) for sequential labelling. Tokens that represent the aspect terms of the sentence are labelled with B. In case an aspect term consists of multiple tokens, the first token receives the B label and the rest receive the I label. Tokens that are not aspect terms are labelled with O. Given the sentence "The internal speakers are amazing." with target "internal speakers", the labelling would be as follows: (The|O) (internal|B) (speakers|I) (are|O) (amazing|O) (.|O).

3.2 Classifier Architecture

We employ a two-layer B-LSTM to extract features for each token, which are used by a CRF for token-based classification. Features are created by exploiting the word morphology and the structure of the sentence. The architecture is depicted in Fig. 1 and is inspired by the NER system presented in (Yang et al., 2016). However, we employ LSTM cells and use word embeddings from fastText[7].

First B-LSTM layer: Randomly initialized character embeddings of each token are given as input into the first B-LSTM layer, which aims at learning new word embeddings. The first and second directions (left $\rightarrow$ right and left $\leftarrow$ right) of the first B-LSTM layer are responsible for learning word embeddings by exploiting the prefix and the suffix of each token respectively.

Second B-LSTM layer: For each token of a sentence, we create a feature vector by combining (i) the extracted word embeddings from the first B-LSTM layer and (ii) pre-trained word embeddings using fastText. These feature vectors are given as input to the second B-LSTM layer, which extracts a feature vector for each token by exploiting the structure of the sentence. Similar to the first B-LSTM layer, the first and second directions are responsible for extracting features using the previous and the next tokens of each word.

CRF layer: The final layer uses the extracted feature vectors in order to perform sequential labelling.

4 Unsupervised Aspect Term Extraction

The human annotation process — required to identify aspect terms in small sentences and construct datasets for supervised ATE — comes at a high cost, mainly for two reasons:

1. Human annotated datasets typically consist of a few thousand sentences[8] extracted from large corpora of domain-specific reviews. The **small amount of data** reduces the performance of classifiers.

2. **Human annotation** is very slow, costly and risky. Annotators may introduce noise in the datasets by labelling words incorrectly, either because they are sloppy workers or because they do not know exactly what aspect terms are. For example, given the sentence "Works well, and I am extremely happy to be back to an apple OS.", human annotators may consider the word "works" as a target[9]. However, aspect terms are nouns and noun

---

[6]https://spacy.io/docs/

[7]https://github.com/facebookresearch/fastText

[8]The datasets of the SemEval ABSA task consist of approximately 3000 sentences for English.

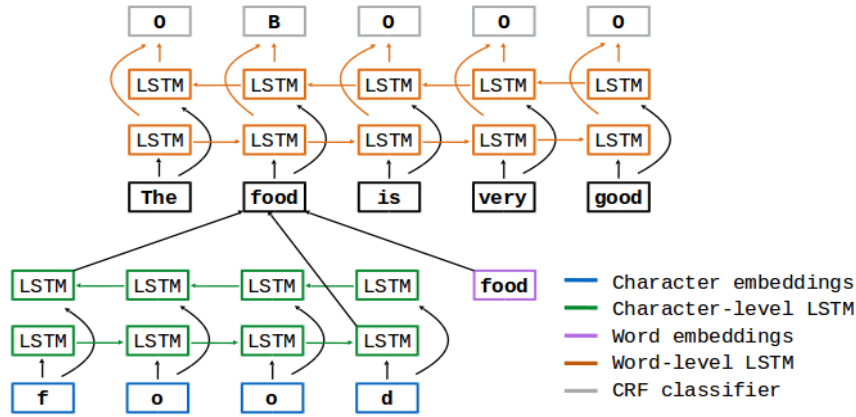[9]Example taken from the golden annotated dataset for laptop reviews of the SemEval-2014 ABSA task.

Figure 1: Sequential labelling using B-LSTM & CRF classifier.

phrases (Liu et al., 2015), therefore the verb "works" should not be considered as a target. We employ unsupervised ATE in order to overcome both problems. We tackle the first problem by using large datasets of opinion texts (e.g. reviews). Such datasets are ideal for ATE since they contain a plethora of opinionated aspect terms. In order to tackle the second problem, we introduce and use an automated and unsupervised method for labelling the tokens of the aforementioned datasets using the IOB format. We consider only noun and noun phrases as candidate aspect terms and focus on token labelling with high precision in order to reduce falsely annotated aspect terms. In that way, we minimize the cost, the time and the mistakes introduced by the human annotation process.

We use the publicly available datasets of *Amazon* and *Yelp* for laptop and restaurant reviews respectively and perform some data cleaning such as removing URLs from the reviews.

4.1 Automated Data Labelling

Using raw opinion texts (e.g. reviews) for ATE is a completely unsupervised task since there are no labels to explicitly pinpoint that certain tokens of the text are aspect terms. Reviews frequently contain labels (e.g. number of stars in a 1-5 star rating system) related to their overall polarity but these are not useful for ATE.

Our goal is to label each token of the unlabelled opinion texts in an automated way using the IOB format with unsupervised methods. While labelling aspect terms, we focus on high precision, a property that guarantees that the resulting datasets will contain as little noisy aspect terms as possible. The importance of high precision is also high-

lighted in (Liu et al., 2015), where authors construct syntactic rules primarily by focusing on this criterion.

Algorithm 1 describes the automated data labelling method. First, we create a list of quality phrases and prune it using a desired threshold value. Then, we iterate through all sentences and annotate tokens that obey certain syntactic rules as aspect terms. We repeat this procedure for multi-word aspect terms and finally label the tokens using the IOB format.

Algorithm 1 Automated Data Labelling

1: *qual_phrases = run_autophrase(corpus)*
2: *candidates = prune(qual_phrases, q_{th})*
3: **for** sentence **in** corpus **do**
4:     labels = []
5:     **for** token **in** sentence **do**
6:         **if** token **in** candidates **then**
7:             *l = get_label(token, rules, lexicon)*
8:             *labels.append(l)*
9:     *assign_iob_tags(sentence, labels)*

4.1.1 Quality Phrase List

We start by building a sorted list of the form (*quality phrase, q*), where $q \in [0, 1]$ represents the quality value of each phrase. The quality phrases — which we use as candidate aspect terms — are *n-grams* that appear in the raw review corpora and exceed a minimum support threshold[10]. The list of quality phrases is built by applying the AutoPhrase algorithm (Shang et al., 2017) on the review datasets for laptops and restaurants. The quality of each phrase is determined via a

---

[10]Support is an indication of how frequently the *n-gram* appears in the dataset.

classification task with decision trees that takes into account a list of high quality phrases using *Wikipedia*. The values of the features (e.g. *tf-idf*) used in the decision trees to predict the quality of each phrase are more informative when the provided corpora are domain dependent. Therefore, we apply AutoPhrase on each dataset separately, rather than combining the two datasets.

The extracted quality phrases, together with a set of syntactic rules, contribute in the automated data labelling process, which is based on 3 pillars:

1. a sentiment lexicon
2. a pruned list of quality phrases
3. syntactic rules able to capture aspect terms

Existing ATE systems (Garcia-Pablos et al., 2015), although unsupervised, exploit also syntactic rules derived from supervised tools (e.g. parsers). Moreover, they require domain-dependent human input (e.g. seed words) to perform double-propagation. We avoid that by using a sentiment lexicon.

4.1.2 Sentiment Lexicon

In many cases, aspect terms have modifiers (e.g. "This is a great screen") or are objects of verbs (e.g. "I love the screen of this laptop") that express a sentiment. Therefore, we make use of a sentiment lexicon[11], which is necessary in order to perform a look-up on whether modifiers and verbs express a sentiment or not.

4.1.3 Pruned Quality Phrases

We prune our quality phrases since they contain both true and noisy aspect term candidates. More concretely, we filter the list of quality phrases in order to keep *n-grams* with a quality above a certain threshold.

We present an example to show the value of the pruning step. The list of quality phrases extracted using the *Amazon* review dataset on laptops contains the *1-gram* "couch" and the *2-gram* "touch pad" with quality 0.67 and 0.95 respectively. However, the presence of the word "couch" as an aspect term in laptop reviews is completely arbitrary. Therefore, we prune the list of quality phrases using an empirical quality threshold of $q_{th} = 0.7$ and $q_{th} = 0.6$ for the laptop and restaurant domain respectively. We set these thresholds manually after inspecting the lists of qual-

---

[11]We use the sentiment lexicon of Bing Liu: https://www.cs.uic.edu/~liub/FBS/ sentiment-analysis.html

ity phrases and detecting the quality value under which a lot of domain-irrelevant candidate aspect terms appear.

While the pruning step removes irrelevant phrases, as shown above, it also means that *n-grams* such as "set up", which are true aspect term candidates are removed from the list due to low quality ($q_{set\ up} = 0.32$). However, reducing noisy aspect term candidates (e.g. "couch" with $q = 0.67$) is more important than keeping all aspect term candidates since we wish to annotate aspect terms with high precision.

We can make the data labelling method completely automated by setting a fixed quality threshold q_{th} for pruning the list of quality phrases. We highlight that a fixed threshold of $q_{th} = 0.7$ leads to a good — but not optimal — trade-off between high precision values and good *F-score* for ATE.

4.1.4 Syntactic Rules for ATE

The pruned quality phrases and the sentiment lexicon are combined with syntactic rules in order to extract aspect terms from sentences. Before applying any syntactic rule, we validate if a token is a potential aspect term by checking if it (i) is not a stopword, (ii) is present in the pruned quality phrases and (iii) has a POS tag that is present in [NOUN, PRON, PROPN, ADJ, ADP, CONJ]. Table 1 tabulates all rules used for ATE and gives examples of reviews with the respective extracted aspect terms. For simplicity, we adopt a syntactic rule notation similar to the one used in (Liu et al., 2015). The functions used in Table 1 have the following interpretation:

- $depends(d, t_i, t_j)$ is true if the syntactic dependency between the tokens t_i and t_j is d.
- $opinion\_word(t_i)$ is true if the token t_i is in the sentiment lexicon.
- $mark\_target(t_i)$ means that we mark the token t_i as aspect term.
- $is\_aspect(t_i)$ is true if the token t_i is already marked as aspect term.

4.1.5 Language and Domain Adaptation

The automated data labelling method requires adaptation in order to be used in different languages. More concretely, we need to adapt (i) the syntactic rules of Table 1, (ii) the sentiment lexicon and (iii) the tools required from Autophrase (e.g. part-of-speech tagger) to the target language. We can use the automated data labelling method for ATE dataset construction in a completely

Rules	Example	Extracted Targets
$depends(dobj, t_i, t_j)$ **and** $opinion\_word(t_j)$ **then** $mark\_target(t_i)$	I like the screen	screen
$depends(nsubj, t_i, t_j)$ **and** $depends(acomp, t_k, t_j)$ **and** $opinion\_word(t_k)$ **then** $mark\_target(t_i)$	The internal speakers are amazing	internal speakers
$depends(nsubj, t_i, t_j)$ **and** $depends(advmod, t_j, t_j)$ **and** $opinion\_word(t_k)$ **then** $mark\_target(t_i)$	The touchpad works perfectly	touchpad
$depends(pobj$ **or** $dobj, t_i, t_j)$ **and** $depends(amod, t_k, t_i)$ **and** $opinion\_word(t_k)$ **then** $mark\_target(t_i)$	This laptop has great price	price
$depends(cc$ **or** $conj, t_i, t_j)$ **and** $is\_aspect(t_j)$ **then** $mark\_target(t_i)$	Screen and speakers are awful	screen speakers
$depends(compound, t_i, t_j)$ **and** $is\_aspect(t_j)$ **then** $mark\_target(t_i)$	The wifi card is not good	wifi card

Table 1: Syntactic rules for aspect term extraction.

domain-independent fashion. To do so, we only need to set the pruning threshold q_{th} of the quality phrase list to a fixed value (Section 4.1.3). Our experiments reveal that setting $q_{th} = 0.7$ results in a good trade-off between high precision and *F-score*, independently of the laptop or the restaurant domain.

4.2 Model Training and Evaluation

We train a B-LSTM & CRF classifier to perform unsupervised ATE for both domains using the automatically labelled datasets constructed in Section 4.1. The classifier is evaluated on the human annotated test datasets of the SemEval-2014 ABSA contest.

5 Experiments and Results

We perform experiments for supervised and unsupervised ATE in the laptop and the restaurant domain and evaluate our classifier using the CoNLL[12] *F-score*. Compared to other supervised learning methods, we reach the performance of SemEval-2014 ABSA winners in the restaurant domain. For laptops, our supervised system exceeds the best *F-score* of the SemEval-2014 ABSA contest by approximately 3%. With respect to unsupervised ATE, our technique achieves (i) very high precision and (ii) an *F-score* that exceeds the supervised baseline of the SemEval ABSA.

5.1 Experiments for Supervised ATE

For supervised learning, we perform experiments using the human annotated training and test sets provided by the SemEval-2014 ABSA contest for
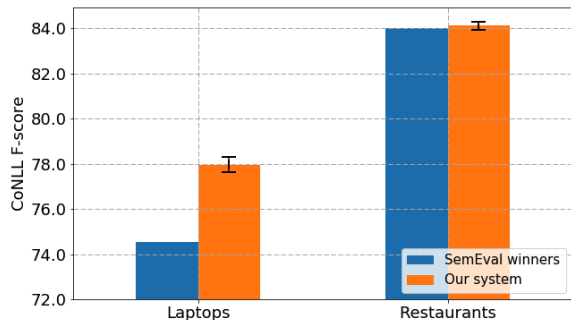

Figure 2: Results for supervised ATE using the B-LSTM & CRF architecture. We compare against the winners of the SemEval-2014 ABSA contest.

the laptop and restaurant domain. Our classifier uses the B-LSTM & CRF architecture presented in Fig. 1 and its implementation is based on (Dernoncourt et al., 2017).

We use a random 80-20% split on the original training set of SemEval-2014 ABSA contest in order to create a new train and validation set. We keep the test set for our final evaluation. For most of the parameters, we use the default values of (Dernoncourt et al., 2017). However, we use the *adam* optimizer with learning rate $\alpha = 0.01$ and a batch size of 64. Moreover, we use the pre-trained word embeddings of fastText.

We train the classifier using the reduced training set for a maximum number of 100 epochs. After each epoch, we evaluate our model using the CoNLL *F-score* on the validation set. Moreover, we use early stopping with a patience of 20 epochs. This means that the experiment terminates earlier if the CoNLL *F-score* of the validation set does not improve for 20 consecutive epochs. At the end of each experiment we choose the model of the epoch that gives the best performance on the

---

[12]http://www.cnts.ua.ac.be/conll2003/

validation set and make predictions on the test set. We repeat the aforementioned procedure for 50 experiments and present the experimental results for both domains in Fig 2.

The *F-score* of the SemEval-2014 ABSA winners is 74.55 and 84.01 for the laptop and the restaurant domain respectively. The B-LSTM & CRF classifier achieves an *F-score* of 77.96 ± 0.38 for laptops and an *F-score* of 84.12 ± 0.2 for restaurants with a confidence interval of 95%. With our performance, we would have surely won in the laptop domain and probably also in the restaurant domain.

5.2 Experiments for Unsupervised ATE

We also perform experiments for ATE with unsupervised learning. For training, we use the automatically labelled datasets (hereafter denoted as ALD) obtained using the methodology described in Section 4.1 with $q_{th} = 0.7$ and $q_{th} = 0.6$ for the laptop and the restaurant domain respectively. For testing, we use the human labelled datasets (hereafter denoted as HLD) of the SemEval-2014 ABSA task.

Our main goal is to evaluate our unsupervised technique on human annotated datasets. To the best of our knowledge, the largest available human annotated datasets for ATE are provided by the SemEval ABSA task and contain laptop and restaurant reviews. Therefore, our analysis focuses only on these two domains.

We start by creating a rule-based baseline model to make predictions for the HLD simply by applying techniques presented in Section 4.1. This baseline (presented in the following section) does not rely on any machine learning techniques for the annotation procedure.

We aim at beating the rule-based baseline by using machine learning. To this end, we use the ALD to train our classifier. For unsupervised ATE, we run two types of experiments. The first one uses the traditional IOB labelling format and is stricter. The second one is more relaxed and uses only B and O labels (i.e. I labels are converted to B). The intuition is that aspect terms can be identified by separating B and I labels from O. Therefore, I and B labels are treated equally against O.

Rule-based Baseline Model

The methodology described in Section 4.1 is used in order to make predictions on the HLD for laptops and restaurants, i.e. the rule-based baseline

	Labels: IOB		Labels: OB		
	P	**F_1**	**P**	**F_1**	
Rule-based	65.13	24.35	76.65	23.76	Laptops
SVM	61.64	37.94	72.02	43.29	
B-LSTM & CRF	**66.67**	**42.09**	**74.51**	**44.37**	
SemEval		35.64			
Rule-based	84.26	28.74	96.67	27.37	Restaurants
SVM	67.28	48.08	80.83	57.36	
B-LSTM & CRF	**74.03**	**53.93**	**83.19**	**63.09**	
SemEval		47.15			

Table 2: Experiments for unsupervised ATE. We compare B-LSTM & CRF classifier against the rule-based baseline, an SVM classifier and the baseline of the SemEval-2014 ABSA contest.

model does not use any machine learning algorithm. During the annotation process, a token of the HLD is labelled as a target if (i) it belongs in the pruned quality phrases list and (ii) satisfies at least one of the rules in Table 1. A comparison between the predicted and the golden labels of the HLD gives a quality estimation of the syntactic rules we create and acts as a baseline.

SVM

We train a linear SVM classifier[13] in order to create a second baseline model that uses machine learning. For SVM, we use the baseline features presented in (Stratos and Collins, 2015) and build 1-0 feature vectors by exploiting the word morphology and the sentence structure (i.e. adjacent words of each token). The training and the evaluation are done using the ALD and the HLD respectively.

In addition, we wish to show the trade-off between precision and recall for different values of q_{th}. We perform experiments for different values of q_{th} and validate that the higher q_{th} the higher the precision and the lower the recall. For example, an SVM classifier trained on an ALD with $q_{th} = 0.7$ achieves an $F_1 = 39.63$ and $P = 71.54$ (Table 2 shows results for $q_{th} = 0.6$ for restaurants). We choose to keep $q_{th} = 0.6$ for the restaurant domain because we are interested in a good combination of high precision and *F-score*.

---

[13]We use the implementation of LIBLINEAR (Fan et al., 2008).

B-LSTM & CRF

We employ the B-LSTM & CRF classifier using the ALD as training set and the HLD as test set, i.e. the evaluation is performed on the human annotated datasets of SemEval-2014 ABSA task. In addition, we use the ABSA training sets of SemEval-2014 as validation sets.

The maximum number of epochs and the patience are set to 20 and 5 respectively. As stopping criterion, we simply choose the epoch that achieves the best *F-score* on the validation set. In all our experiments, we compare the performance of the B-LSTM & CRF classifier with the respective performance of the rule-based baseline and the SVM model. We do not report any confidence intervals for the B-LSTM & CRF classifier because the training time increases dramatically in the case of unsupervised ATE due to the increased size of the dataset. Conducting one experiment usually takes more than $15h$, which means that a round of at least 20 experiments, that would allow for defining confidence intervals, would be computationally intensive. For this reason, we leave the report of confidence intervals for unsupervised ATE for future work. However, we repeat up to 3 experiments for each case and verify that the CoNLL *F-score* and the precision are always higher compared to SVM. Results for the laptop domain can be visualized in Fig. 3. We do not present any figures for the restaurant domain since the learning curves are very similar to the ones of the laptop domain.

We draw several conclusions by observing the results tabulated in Table 2. First, the B-LSTM & CRF classifier achieves higher *F-score* for both domains compared to the rule-based baseline model and the SVM classifier. The *F-score* relative improvement between the rule-based baseline and the B-LSTM & CRF classifier is 73% and 88% for the laptop and the restaurant domain respectively. At the same time, we preserve high precision and attain values above 80%. Finally, our unsupervised method beats the supervised baseline *F-score* from SemEval ABSA.

6 Conclusion and Future Work

We present a B-LSTM & CRF classifier which we use for feature extraction and aspect term detection for both supervised and unsupervised ATE. We validate this classifier by performing supervised ATE and achieving top-ranking performance
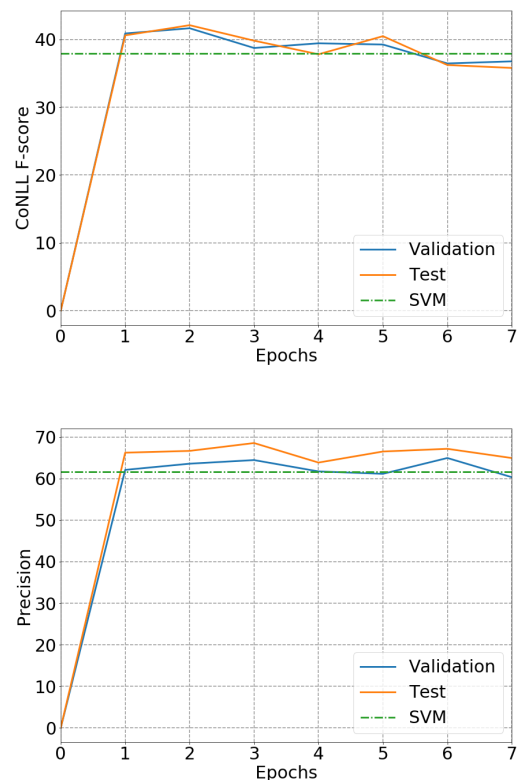

Figure 3: *F-score* (top) and precision (bottom) comparison between B-LSTM & CRF and SVM for unsupervised ATE in the laptop domain. B, I and O labels are used.

on the human annotated datasets of the SemEval-2014 ABSA contest for the laptop and restaurant domain. Moreover, we introduce a new, automated, unsupervised and domain independent method to label tokens of raw opinion texts as aspect terms with high precision. We use the automatically labelled datasets to train the B-LSTM & CRF classifier, which we evaluate on human annotated datasets. Against all odds, our unsupervised method beats the supervised ABSA baseline *F-score* from SemEval, while preserving high precision scores.

As future work, we plan to perform ATE for different domains (e.g. hotels) using our methods. Moreover, we plan to work towards adapting our techniques to multilingual datasets (e.g. French, Spanish, etc.). We would also investigate the idea of exploiting the available ratings (e.g. 1-5 stars) of the review datasets in order to construct new datasets for ATE. This would allow us to perform ATE with distant supervision.

References

Caroline Brun, Diana Nicoleta Popa, and Claude Roux. 2014. Xrce: Hybrid classification for aspect-based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*.

Maryna Chernyshevich. 2014. IHS R&D Belarus: cross-domain extraction of product features using crf. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks .

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* .

Aitor Garcia-Pablos, Montse Cuadros, and German Rigau. 2015. V3: Unsupervised aspect based sentiment analysis for semeval-2015 task 12.

Aitor Garcia-Pablos, Montse Cuadros, and German Rigau. 2017. W2VLDA: Almost unsupervised system for aspect based sentiment analysis .

Aitor Garcia-Pablos and German Rigau. 2014. Unsupervised acquisition of domain aspect terms for aspect based opinion mining.

Tomáš Hercig, Tomáš Brychcín, Lukáš Svoboda, Michal Konkol, and Josef Steinberger. 2016. Unsupervised methods to improve aspect-based sentiment analysis in Czech. *Computación y Sistemas* .

Junichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. *Proceedings of ACL-08: HLT* .

Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for opinion target extraction .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS Proceedings*.

George A. Miller. 1995. Wordnet: A lexical database for english. In *Communications of the ACM*.

John Pavlopoulos and Ion Androutsopoulos. 2015. Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method.

Maria Pontiki, Harris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2017. Automated phrase mining from massive text corpora. *CoRR* abs/1702.04457.

Poria Soujanya, Cambria Erik, and Gelbukh Alexander. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* .

Karl Stratos and Michael Collins. 2015. Simple semi-supervised pos tagging .

Maksim Tkachenko and Andrey Simanovsky. 2012. Named entity recognition: Exploring features .

Zhiqiang Toh and Jian Su. 2015. NLANG: supervised machine learning system for aspect category classification and opinion target extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Zhiqiang Toh and Jian Su. 2016. NLANG at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In *Proceedings of SemEval-2016*.

Zhiqiang Toh and Wenting Wang. 2014. DLIREC: aspect term extraction and term polarity classification system. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch .

Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction .