SemEval-2014 Task 4: Aspect Based Sentiment Analysis

Maria Pontiki
Institute for Language
and Speech Processing,
"Athena" Research Center
mpontiki@ilsp.gr

Dimitrios Galanis
Institute for Language
and Speech Processing,
"Athena" Research Center
galanisd@ilsp.gr

John Pavlopoulos
Dept. of Informatics,
Athens University of
Economics and Business
annis@aueb.gr

Haris Papageorgiou
Institute for Language
and Speech Processing,
"Athena" Research Center
xaris@ilsp.gr

Ion Androutsopoulos
Dept. of Informatics
Athens University of
Economics and Business
ion@aueb.gr

Suresh Manandhar
Dept. of Computer Science,
University of York
suresh@cs.york.ac.uk

Abstract

Sentiment analysis is increasingly viewed as a vital task both from an academic and a commercial standpoint. The majority of current approaches, however, attempt to detect the overall polarity of a sentence, paragraph, or text span, irrespective of the entities mentioned (e.g., laptops) and their aspects (e.g., battery, screen). SemEval-2014 Task 4 aimed to foster research in the field of aspect-based sentiment analysis, where the goal is to identify the aspects of given target entities and the sentiment expressed for each aspect. The task provided datasets containing manually annotated reviews of restaurants and laptops, as well as a common evaluation procedure. It attracted 163 submissions from 32 teams.

1 Introduction

With the proliferation of user-generated content on the web, interest in mining sentiment and opinions in text has grown rapidly, both in academia and business. Early work in sentiment analysis mainly aimed to detect the overall polarity (e.g., positive or negative) of a given text or text span (Pang et al., 2002; Turney, 2002). However, the need for a more fine-grained approach, such as aspect-based (or 'feature-based') sentiment analysis (ABSA), soon became apparent (Liu, 2012). For example, laptop reviews not only express the overall sentiment about a specific model (e.g., *"This is a great laptop"*), but also sentiments relating to its specific aspects, such as the hardware, software, price, etc. Subsequently, a review may convey opposing sentiments (e.g., *"Its performance is ideal, I wish I could say the same about the price"*) or objective information (e.g., *"This one still has the CD slot"*) for different aspects of an entity.

ABSA is critical in mining and summarizing opinions from on-line reviews (Gamon et al., 2005; Titov and McDonald, 2008; Hu and Liu, 2004a; Popescu and Etzioni, 2005). In this setting, ABSA aims to identify the aspects of the entities being reviewed and to determine the sentiment the reviewers express for each aspect. Within the last decade, several ABSA systems of this kind have been developed for movie reviews (Thet et al., 2010), customer reviews of electronic products like digital cameras (Hu and Liu, 2004a) or netbook computers (Brody and Elhadad, 2010), services (Long et al., 2010), and restaurants (Ganu et al., 2009; Brody and Elhadad, 2010).

Previous publicly available ABSA benchmark datasets adopt different annotation schemes within different tasks. The restaurant reviews dataset of Ganu et al. (2009) uses six coarse-grained aspects (e.g., FOOD, PRICE, SERVICE) and four overall sentence polarity labels (positive, negative, conflict, neutral). Each sentence is assigned one or more aspects together with a polarity label for each aspect; for example, "*The restaurant was expensive, but the menu was great.*" would be assigned the aspect PRICE with negative polarity and FOOD with positive polarity. In the product reviews dataset of Hu and Liu (2004a; 2004b), *aspect terms*, i.e., terms naming aspects (e.g., 'radio', 'voice dialing') together with strength scores (e.g., 'radio': $+2$, 'voice dialing': -3) are pro-

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27–35,
Dublin, Ireland, August 23-24, 2014.

vided. No predefined inventory of aspects is provided, unlike the dataset of Ganu et al.

The SemEval-2014 ABSA Task is based on laptop and restaurant reviews and consists of four subtasks (see Section 2). Participants were free to participate in a subset of subtasks and the domains (laptops or restaurants) of their choice.

2 Task Description

For the first two subtasks (SB1, SB2), datasets on both domains (restaurants, laptops) were provided. For the last two subtasks (SB3, SB4), datasets only for the restaurant reviews were provided.

Aspect term extraction (SB1): Given a set of review sentences, the task is to identify all aspect terms present in each sentence (e.g., 'wine', 'waiter', 'appetizer', 'price', 'food'). We require all the aspect terms to be identified, including aspect terms for which no sentiment is expressed (neutral polarity). These will be useful for constructing an ontology of aspect terms and to identify frequently discussed aspects.

Aspect term polarity (SB2): In this subtask, we assume that the aspect terms are given (as described in SB1) and the task is to determine the polarity of each aspect term (positive, negative, conflict, or neutral). The conflict label applies when both positive and negative sentiment is expressed about an aspect term (e.g., "*Certainly not the best sushi in New York, however, it is always fresh*"). An alternative would have been to tag the aspect term in these cases with the dominant polarity, but this in turn would be difficult to agree on.

Aspect category detection (SB3): Given a predefined set of aspect categories (e.g., PRICE, FOOD) and a set of review sentences (but without any annotations of aspect terms and their polarities), the task is to identify the aspect categories discussed in each sentence. Aspect categories are typically coarser than the aspect terms as defined in SB1, and they do not necessarily occur as terms in the sentences. For example, in *"Delicious but expensive"*, the aspect categories FOOD and PRICE are not instantiated through specific aspect terms, but are only inferred through the adjectives 'delicious' and 'expensive'. SB1 and SB3 were treated as separate subtasks, thus no information linking aspect terms to aspect categories was provided.

Aspect category polarity (SB4): For this subtask, aspect categories for each review sentence are provided. The goal is to determine the polarity (positive, negative, conflict, or neutral) of each aspect category discussed in each sentence.

Subtasks SB1 and SB2 are useful in cases where no predefined inventory of aspect categories is available. In these cases, frequently discussed aspect terms of the entity can be identified together with their overall sentiment polarities. We hope to include an additional *aspect term aggregation* subtask in future (Pavlopoulos and Androutsopoulos, 2014b) to cluster near-synonymous (e.g., 'money', 'price', 'cost') or related aspect terms (e.g., 'design', 'color', 'feeling') together with their averaged sentiment scores as shown in Fig. 1.

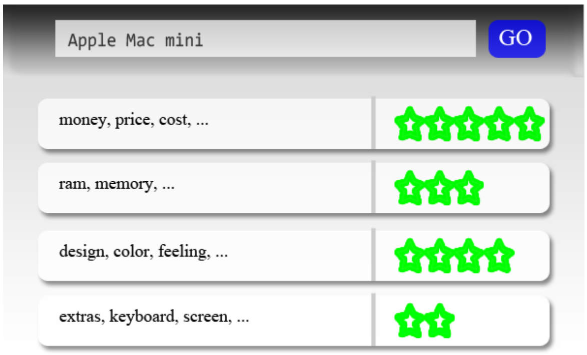


Figure 1: Aggregated aspect terms and average sentiment polarities for a target entity.

Subtasks SB3 and SB4 are useful when a predefined inventory of (coarse) aspect categories is available. A table like the one of Fig. 1 can then also be generated, but this time using the most frequent aspect categories to label the rows, with stars showing the proportion of reviews expressing positive vs. negative opinions for each aspect category.

3 Datasets

3.1 Data Collection

The training and test data sizes are provided in Table 1. The restaurants training data, consisting of 3041 English sentences, is a subset of the dataset from Ganu et al. (2009), which included annotations for coarse aspect categories (as in SB3) and overall sentence polarities. We added annotations for aspect terms occurring in the sentences (SB1), aspect term polarities (SB2), and aspect category polarities (SB4). Additional restaurant reviews were collected and annotated (from scratch) in the same manner and used as test data (800 sentences). The laptops dataset contains 3845 English sentences extracted from laptop custumer reviews. Human annotators tagged the aspect terms (SB1) and their polarities (SB2); 3045 sentences were used for training and 800 for testing (evaluation).

Domain	Train	Test	Total
Restaurants	3041	800	3841
Laptops	3045	800	3845
Total	6086	1600	7686

Table 1: Sizes (sentences) of the datasets.

3.2 Annotation Process

For a given target entity (a restaurant or a laptop) being reviewed, the annotators were asked to provide two types of information: aspect terms (SB1) and aspect term polarities (SB2). For the restaurants dataset, two additional annotation layers were added: aspect category (SB3) and aspect category polarity (SB4).

The annotators used BRAT (Stenetorp et al., 2012), a web-based annotation tool, which was configured appropriately for the needs of the ABSA task.[1] Figure 2 shows an annotated sentence in BRAT, as viewed by the annotators.

Stage 1: Aspect terms and polarities. During a first annotation stage, the annotators tagged all the single or multiword terms that named particular aspects of the target entity (e.g., *"I liked the service and the staff, but not the food"* → {'service', 'staff', 'food'}, *"The hard disk is very noisy"* → {'hard disk'}). They were asked to tag only aspect terms explicitly naming particular aspects (e.g., *"everything about it"* or *"it's expensive"* do not name particular aspects). The aspect terms were annotated as they appeared, even if misspelled (e.g., *'warrenty'* instead of *'warranty'*). Each identified aspect term also had to be assigned a polarity label (positive, negative, neutral, conflict). For example, *"I hated their fajitas, but their salads were great"* → {'fajitas': negative, 'salads': positive}, *"The hard disk is very noisy"* → {'hard disk': negative}.

Each sentence of the two datasets was annotated by two annotators, a graduate student (annotator A) and an expert linguist (annotator B). Initially, two subsets of sentences (300 from each dataset) were tagged by annotator A and the annotations were inspected and validated by annotator B. The disagreements between the two annotators were confined to borderline cases. Taking into account the types of these disagreements (discussed below), annotator A was provided with additional guidelines and tagged the remainder of the sentences in both datasets.[2] When A was not confident, a decision was made collaboratively with B. When A and B disagreed, a decision was made collaboratively by them and a third expert annotator. Most disagreements fall into one of the following three types:

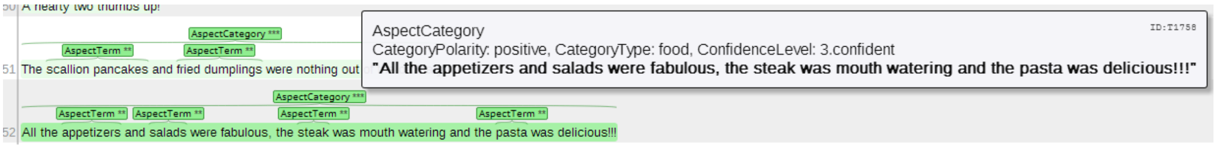
Polarity ambiguity: In several sentences, it was unclear if the reviewer expressed positive or negative opinion, or no opinion at all (just reporting a fact), due to lack of context. For example, in *"12.44 seconds boot time"* it is unclear if the reviewer expresses a positive, negative, or no opinion about the aspect term 'boot time'. In future challenges, it would be better to allow the annotators (and the participating systems) to consider the entire review instead of each sentence in isolation.

Multi-word aspect term boundaries: In several cases, the annotators disagreed on the exact boundaries of multi-word aspect terms when they appeared in conjunctions or disjunctions (e.g., *"selection of meats and seafoods"*, *"noodle and rices dishes"*, *"school or office use"*). In such cases, we asked the annotators to tag as a single aspect term the maximal noun phrase (the entire conjunction or disjunction). Other disagreements concerned the extent of the aspect terms when adjectives that may or may not have a subjective meaning were also present. For example, if 'large' in *"large whole shrimp"* is part of the dish name, then the guidelines require the adjective to be included in the aspect term; otherwise (e.g., in *"large portions"*) 'large' is a subjectivity indicator not to be included in the aspect term. Despite the guidelines, in some cases it was difficult to isolate and tag the exact aspect term, because of intervening words, punctuation, or long-term dependencies.

Aspect term vs. reference to target entity: In some cases, it was unclear if a noun or noun phrase was used as the aspect term or if it referred to the entity being reviewed as whole. In "*This place is awesome*", for example, 'place' most probably refers to the restaurant as a whole (hence, it should not be tagged as an aspect term), but in "*Cozy*

[1] Consult `http://brat.nlplab.org/` for more information about BRAT.

[2] The guidelines are available at: `http://alt.qcri.org/semeval2014/task4/data/uploads/`.



50 A hearty two thumbs up!

51 The scallion pancakes and fried dumplings were nothing out

52 All the appetizers and salads were fabulous, the steak was mouth watering and the pasta was delicious!!!

AspectCategory

CategoryPolarity: positive, CategoryType: food, ConfidenceLevel: 3.confident

"All the appetizers and salads were fabulous, the steak was mouth watering and the pasta was delicious!!!"

Figure 2: A sentence in the BRAT tool, annotated with four aspect terms (‘appetizers’, ‘salads’, ‘steak’, ‘pasta’) and one aspect category (FOOD). For aspect categories, the whole sentence is tagged.

place and good pizza” it probably refers to the ambience of the restaurant. A broader context would again help in some of these cases.

We note that laptop reviews often evaluate each laptop as a whole, rather than expressing opinions about particular aspects. Furthermore, when they express opinions about particular aspects, they often do so by using adjectives that refer implicitly to aspects (e.g., ‘expensive’, ‘heavy’), rather than using explicit aspect terms (e.g., ‘cost’, ‘weight’); the annotators were instructed to tag only explicit aspect terms, not adjectives implicitly referring to aspects. By contrast, restaurant reviews contain many more aspect terms (Table 2, last column).[3]

Dataset	Pos.	Neg.	Con.	Neu.	Tot.
LPT-TR	987	866	45	460	2358
LPT-TE	341	128	16	169	654
RST-TR	2164	805	91	633	3693
RST-TE	728	196	14	196	1134

Table 2: Aspect terms and their polarities per domain. LPT and RST indicate laptop and restaurant reviews, respectively. TR and TE indicate the training and test set.

Another difference between the two datasets is that the neutral class is much more frequent in (the aspect terms of) laptops, since laptop reviews often mention features without expressing any (clear) sentiment (e.g., “*the latest version does not have a disc drive*”). Nevertheless, the positive class is the majority in both datasets, but it is much more frequent in restaurants (Table 2). The majority of the aspect terms are single-words in both datasets (2148 in laptops, 4827 in restaurants, out of 3012 and 4827 total aspect terms, respectively).

Stage 2: Aspect categories and polarities. In this task, each sentence needs to be tagged with the aspect categories discussed in the sentence. The aspect categories are FOOD, SERVICE, PRICE, AMBIENCE (the atmosphere and environment of a restaurant), and ANECDOTES/MISCELLANEOUS (sentences not belonging in any of the previous aspect categories). [4] For example, *“The restaurant was expensive, but the menu was great”* is assigned the aspect categories PRICE and FOOD. Additionally, a polarity (positive, negative, conflict, neutral) for each aspect category should be provided (e.g., *“The restaurant was expensive, but the menu was great”* → {PRICE: negative, FOOD: positive}.

One annotator validated the existing aspect category annotations of the corpus of Ganu et al. (2009). The agreement with the existing annotations was 92% measured as average F_1. Most disagreements concerned additions of missing aspect category annotations. Furthermore, the same annotator validated and corrected (if needed) the existing polarity labels per aspect category annotation. The agreement for the polarity labels was 87% in terms of accuracy and it was measured only on the common aspect category annotations. The additional 800 sentences (not present in Ganu et al.’s dataset) were used for testing and were annotated from scratch in the same manner. The distribution of the polarity classes per category is presented in Table 3. Again, ‘positive’ is the majority polarity class while the dominant aspect category is FOOD in both the training and test restaurant sentences.

Determining the aspect categories of the sentences and their polarities (Stage 2) was an easier task compared to detecting aspect terms and their polarities (Stage 1). The annotators needed less time in Stage 2 and it was easier to reach agreement. Exceptions were some sentences where it was difficult to decide if the categories AMBIENCE or ANECDOTES/MISCELLANEOUS applied (e.g., *“One of my Fav spots in the city”*). We instructed the annotators to classify those sentences only in ANECDOTES/MISCELLANEOUS, if they conveyed

[3] We count aspect term *occurrences*, not distinct terms.

[4] In the original dataset of Ganu et al. (2009), ANECDOTES and MISCELLANEOUS were separate categories, but in practice they were difficult to distinguish and we merged them.

	Positive		Negative		Conflict		Neutral		Total	
Category	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
FOOD	867	302	209	69	66	16	90	31	1232	418
PRICE	179	51	115	28	17	3	10	1	321	83
SERVICE	324	101	218	63	35	5	20	3	597	172
AMBIENCE	263	76	98	21	47	13	23	8	431	118
ANECD./MISC.	546	127	199	41	30	15	357	51	1132	234
Total	2179	657	839	159	163	52	500	94	3713	1025

Table 3: Aspect categories distribution per sentiment class.

general views about a restaurant, without explicitly referring to its atmosphere or environment.

3.3 Format and Availability of the Datasets

The datasets of the ABSA task were provided in an XML format (see Fig. 3). They are available with a non commercial, no redistribution license through META-SHARE, a repository devoted to the sharing and dissemination of language resources (Piperidis, 2012).[5]

4 Evaluation Measures and Baselines

The evaluation of the ABSA task ran in two phases. In Phase A, the participants were asked to return the aspect terms (SB1) and aspect categories (SB3) for the provided test datasets. Subsequently, in Phase B, the participants were given the gold aspect terms and aspect categories (as in Fig. 3) for the sentences of Phase A and they were asked to return the polarities of the aspect terms (SB2) and the polarities of the aspect categories of each sentence (SB4).[6] Each participating team was allowed to submit up to two runs per subtask and domain (restaurants, laptops) in each phase; one constrained (C), where only the provided training data and other resources (e.g., publicly available lexica) excluding additional annotated sentences could be used, and one unconstrained (U), where additional data of any kind could be used for training. In the latter case, the teams had to report the resources they used.

To evaluate aspect term extraction (SB1) and aspect category detection (SB3) in Phase A, we used the F_1 measure, defined as usually:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{1}$$

where precision (P) and recall (R) are defined as:

$$P = \frac{|S \cap G|}{|S|}, R = \frac{|S \cap G|}{|G|} \tag{2}$$

Here S is the set of aspect term or aspect category annotations (in SB1 and SB3, respectively) that a system returned for all the test sentences (of a domain), and G is the set of the gold (correct) aspect term or aspect category annotations.

To evaluate aspect term polarity (SB2) and aspect category polarity (SB4) detection in Phase B, we calculated the accuracy of each system, defined as the number of correctly predicted aspect term or aspect category polarity labels, respectively, divided by the total number of aspect term or aspect category annotations. Recall that we used the gold aspect term and category annotations in Phase B.

We provided four baselines, one per subtask:[7]

Aspect term extraction (SB1) baseline: A sequence of tokens is tagged as an aspect term in a test sentence (of a domain), if it is listed in a dictionary that contains all the aspect terms of the training sentences (of the same domain).

Aspect term polarity (SB2) baseline: For each aspect term t in a test sentence s (of a particular domain), this baseline checks if t had been encountered in the training sentences (of the domain). If so, it retrieves the k most similar to s training sentences (of the domain), and assigns to the aspect term t the most frequent polarity it had in the k sentences. Otherwise, if t had not been encountered in the training sentences, it is assigned the most frequent aspect term polarity label of the

[5] The datasets can be downloaded from `http://metashare.ilsp.gr:8080/`. META-SHARE (`http://www.meta-share.org/`) was implemented in the framework of the META-NET Network of Excellence (`http://www.meta-net.eu/`).

[6] Phase A ran from 9:00 GMT, March 24 to 21:00 GMT, March 25, 2014. Phase B ran from 9:00 GMT, March 27 to 17:00 GMT, March 29, 2014.

[7] Implementations of the baselines and further information about the baselines are available at: `http://alt.qcri.org/semeval2014/task4/data/uploads/`.

```
<sentence id="11351725#582163#9">
   <text>Our waiter was friendly and it is a shame that he didnt have a supportive
staff to work with.</text>
        <aspectTerms>
            <aspectTerm term="waiter" polarity="positive" from="4" to="10"/>
            <aspectTerm term="staff" polarity="negative" from="74" to="79"/>
        </aspectTerms>
        <aspectCategories>
            <aspectCategory category="service" polarity="conflict"/>
        </aspectCategories>
</sentence>
```

Figure 3: An XML snippet that corresponds to the annotated sentence of Fig. 2.

training set. The similarity between two sentences is measured as the Dice coefficient of the sets of (distinct) words of the two sentences. For example, the similarity between "*this is a demo*" and "*that is yet another demo*" is $\frac{2\cdot 2}{4+5} = 0.44$.

Aspect category extraction (SB3) baseline: For every test sentence s, the k most similar to s training sentences are retrieved (as in the SB2 baseline). Then, s is assigned the m most frequent aspect category labels of the k retrieved sentences; m is the most frequent number of aspect category labels per sentence among the k sentences.

Aspect category polarity (SB4): This baseline assigns to each aspect category c of a test sentence s the most frequent polarity label that c had in the k most similar to s training sentences (of the same domain), considering only training sentences that have the aspect category label c. Sentence similarity is computed as in the SB2 baseline.

For subtasks SB2 and SB4, we also use a majority baseline that assigns the most frequent polarity (in the training data) to all the aspect terms and aspect categories. The scores of all the baselines and systems are presented in Tables 4–6.

5 Evaluation Results

The ABSA task attracted 32 teams in total and 165 submissions (systems), 76 for phase A and 89 for phase B. Based on the human-annotation experience, the expectations were that systems would perform better in Phase B (SB3, SB4, involving aspect categories) than in Phase A (SB1, SB2, involving aspect terms). The evaluation results confirmed our expectations (Tables 4–6).

5.1 Results of Phase A

The aspect term extraction subtask (SB1) attracted 24 teams for the laptops dataset and 24 teams for the restaurants dataset; consult Table 4.

Laptops		Restaurants	
Team	F_1	Team	F_1
IHS_RD.	74.55†	DLIREC	84.01*
DLIREC	73.78*	XRCE	83.98
DLIREC	70.4	NRC-Can.	80.18
NRC-Can.	68.56	UNITOR	80.09
UNITOR	67.95*	UNITOR	79.96*
XRCE	67.24	IHS_RD.	79.62†
SAP_RI	66.6	UWB	79.35*
IITP	66.55	SeemGo	78.61
UNITOR	66.08	DLIREC	78.34
SeemGo	65.99	ECNU	78.24
ECNU	65.88	SAP_RI	77.88
SNAP	62.4	UWB	76.23
DMIS	60.59	IITP	74.94
UWB	60.39	DMIS	72.73
JU_CSE.	59.37	JU_CSE.	72.34
lsis_lif	56.97	Blinov	71.21*
USF	52.58	lsis_lif	71.09
Blinov	52.07*	USF	70.69
UFAL	48.98	EBDG	69.28*
UBham	47.49	UBham	68.63*
UBham	47.26*	UBham	68.51
SINAI	45.28	SINAI	65.41
EBDG	41.52*	V3	60.43*
V3	36.62*	UFAL	58.88
COMMIT.	25.19	COMMIT.	54.38
NILCUSP	25.19	NILCUSP	49.04
iTac	23.92	SNAP	46.46
		iTac	38.29
Baseline	35.64	Baseline	47.15

Table 4: Results for aspect term extraction (SB1). Stars indicate unconstrained systems. The † indicates a constrained system that was not trained on the in-domain training dataset (unlike the rest of the constrained systems), but on the union of the two training datasets (laptops, restaurants).

Restaurants		Restaurants	
Team	F_1	Team	Acc.
NRC-Can.	88.57	NRC-Can.	82.92
UNITOR	85.26*	XRCE	78.14
XRCE	82.28	UNITOR	76.29*
UWB	81.55*	SAP_RI	75.6
UWB	81.04	SeemGo	74.63
UNITOR	80.76	SA-UZH	73.07
SAP_RI	79.04	UNITOR	73.07
SNAP	78.22	UWB	72.78
Blinov	75.27*	UWB	72.78*
UBham	74.79*	lsis_lif	72.09
UBham	74.24	UBham	71.9
EBDG	73.98*	EBDG	69.75
SeemGo	73.75	SNAP	69.56
SINAI	73.67	COMMIT.	67.7
JU_CSE.	70.46	Blinov	65.65*
lsis_lif	68.27	Ualberta.	65.46
ECNU	67.29	JU_CSE.	64.09
UFAL	64.51	ECNU	63.41
V3	60.20*	UFAL	63.21
COMMIT.	59.3	iTac	62.73*
iTac	56.95	ECNU	60.39*
		SINAI	60.29
		V3	47.21
		Baseline	65.65
Baseline	63.89	Majority	64.09

Table 5: Results for aspect category detection (SB3) and aspect category polarity (SB4). Stars indicate unconstrained systems.

Overall, the systems achieved significantly higher scores (+10%) in the restaurants domain, compared to laptops. The best F_1 score (74.55%) for laptops was achieved by the IHS_RD. team, which relied on Conditional Random Fields (CRF) with features extracted using named entity recognition, POS tagging, parsing, and semantic analysis. The IHS_RD. team used additional reviews from Amazon and Epinions (without annotated terms) to learn the sentiment orientation of words and they trained their CRF on the union of the restaurant and laptop training data that we provided; the same trained CRF classifier was then used in both domains.

The second system, the unconstrained system of DLIREC, also uses a CRF, along with POS and dependency tree based features. It also uses features derived from the aspect terms of the training data and clusters created from additional reviews from YELP and Amazon. In the restaurants domain, the unconstrained system of DLIREC ranked first with an F_1 of 84.01%, but the best unconstrained system, that of XRCE, was very close (83.98%). The XRCE system relies on a parser to extract syntactic/semantic dependencies (e.g., 'dissapointed'–'food'). For aspect term extraction, the parser's vocabulary was enriched with the aspect terms of the training data and a term list extracted from Wikipedia and Wordnet. A set of grammar rules was also added to detect multiword terms and associate them with the corresponding aspect category (e.g., FOOD, PRICE).

The aspect category extraction subtask (SB3) attracted 18 teams. As shown in Table 5, the best score was achieved by the system of NRC-Canada (88.57%), which relied on five binary (one-vs-all) SVMs, one for each aspect category. The SVMs used features based on various types of n-grams (e.g., stemmed) and information from a lexicon learnt from YELP data, which associates aspect terms with aspect categories. The latter lexicon significantly improved F_1. The constrained UNITOR system uses five SVMs with bag-of-words (BoW) features, which in the unconstrained submission are generalized using distributional vectors learnt from Opinosis and TripAdvisor data. Similarly, UWB uses a binary MaxEnt classifier for each aspect category with BoW and TF-IDF features. The unconstrained submission of UWB also uses word clusters learnt using various methods (e.g., LDA); additional features indicate which clusters the words of the sentence being classified come from. XRCE uses information identified by its syntactic parser as well as BoW features to train a logistic regression model that assigns to the sentence probabilities of belonging to each aspect category. A probability threshold, tuned on the training data, is then used to determine which categories will be assigned to the sentence.

5.2 Results of Phase B

The aspect term polarity detection subtask (SB2) attracted 26 teams for the laptops dataset and 26 teams for the restaurants dataset. DCU and NRC-Canada had the best systems in both domains (Table 6). Their scores on the laptops dataset were identical (70.48%). On the laptops dataset, the DCU system performed slightly better (80.95% vs. 80.15%). For SB2, both NRC-Canada and DCU relied on an SVM classifier with features mainly based on n-grams, parse trees, and several out-of-domain, publicly available sentiment lexica (e.g., MPQA, SentiWordnet and Bing Liu's Opinion Lexicon). NRC-Canada also used two automatically compiled polarity lexica for restaurants and laptops, obtained from YELP and Amazon data, respectively. Furthermore, NRC-Canada showed by ablation experiments that the most useful features are those derived from the sentiment lexica. On the other hand, DCU used only publicly available lexica, which were manually adapted by filtering words that do not express sentiment in laptop and restaurant reviews (e.g., 'really') and by adding others that were missing and do express sentiment (e.g., 'mouthwatering').

The aspect category polarity detection subtask (SB4) attracted 20 teams. NRC-Canada again had the best score (82.92%) using an SVM classifier. The same feature set as in SB2 was used, but it was further enriched to capture information related to each specific aspect category. The second team, XRCE, used information from its syntactic parser, BoW features, and an out-of-domain sentiment lexicon to train an SVM model that predicts the polarity of each given aspect category.

6 Conclusions and Future Work

We provided an overview of Task 4 of SemEval-2014. The task aimed to foster research in aspect-based sentiment analysis (ABSA). We constructed and released ABSA benchmark datasets containing manually annotated reviews from two domains (restaurants, laptops). The task attracted 163 submissions from 32 teams that were evaluated in four subtasks centered around aspect terms (detecting aspect terms and their polarities) and coarser aspect categories (assigning aspect categories and aspect category polarities to sentences). The task will be repeated in SemEval-2015 with additional datasets and a domain-adaptation subtask.[8] In the future, we hope to add an aspect term aggregation subtask (Pavlopoulos and Androutsopoulos, 2014a).

Laptops		Restaurants	
Team	Acc.	Team	Acc.
DCU	70.48	DCU	80.95
NRC-Can.	70.48	NRC-Can.	80.15†
SZTE-NLP	66.97	UWB	77.68*
UBham	66.66	XRCE	77.68
UWB	66.66*	SZTE-NLP	75.22
lsis_lif	64.52	UNITOR	74.95*
USF	64.52	UBham	74.6
SNAP	64.06	USF	73.19
UNITOR	62.99	UNITOR	72.48
UWB	62.53	SeemGo	72.31
IHS_RD.	61.62	lsis_lif	72.13
SeemGo	61.31	UWB	71.95
ECNU	61.16	SA-UZH	70.98
ECNU	61.16*	IHS_RD.	70.81
SINAI	58.71	SNAP	70.81
SAP_RI	58.56	ECNU	70.72
UNITOR	58.56*	ECNU	70.72*
SA-UZH	58.25	INSIGHT.	70.72
COMMIT	57.03	SAP_RI	69.92
INSIGHT.	57.03	EBDG	68.6
UMCC.	57.03*	UMCC.	66.84*
UFAL	56.88	UFAL	66.57
UMCC.	56.11	UMCC.	66.57
EBDG	55.96	COMMIT	65.96
JU_CSE.	55.65	JU_CSE.	65.52
UO_UA	55.19*	Blinov	63.58*
V3	53.82	iTac	62.25*
Blinov	52.29*	V3	59.78
iTac	51.83*	SINAI	58.73
DLIREC	36.54	DLIREC	42.32*
DLIREC	36.54*	DLIREC	41.71
IITP	66.97	IITP	67.37
Baseline	51.37	Baseline	64.28
Majority	52.14	Majority	64.19

Table 6: Results for the aspect term polarity subtask (SB2). Stars indicate unconstrained systems. The † indicates a constrained system that was not trained on the in-domain training dataset (unlike the rest of the constrained systems), but on the union of the two training datasets. IITP's original submission files were corrupted; they were resent and scored after the end of the evaluation period.

Acknowledgements

We thank Ioanna Lazari, who provided an initial version of the laptops dataset, Konstantina Papanikolaou, who carried out a critical part of the annotation process, and Juli Bakagianni, who supported our use of the META-SHARE platform. We are also very grateful to the participants for their feedback. Maria Pontiki and Haris Papageorgiou were supported by the IS-HELLEANA (09-

[8] Consult http://alt.qcri.org/semeval2015/task12/.

72-922) and the POLYTROPON (KRIPIS-GSRT, MIS: 448306) projects.

References

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of NAACL*, pages 804–812, Los Angeles, California.

Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric K. Ringger. 2005. Pulse: Mining customer opinions from free text. In *IDA*, pages 121–132, Madrid, Spain.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *Proceedings of WebDB*, Providence, Rhode Island, USA.

Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews. In *Proceedings of KDD*, pages 168–177, Seattle, WA, USA.

Minqing Hu and Bing Liu. 2004b. Mining opinion features in customer reviews. In *Proceedings of AAAI*, pages 755–760, San Jose, California.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Chong Long, Jie Zhang, and Xiaoyan Zhu. 2010. A review selection approach for accurate feature rating estimation. In *Proceedings of COLING (Posters)*, pages 766–774, Beijing, China.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, Philadelphia, Pennsylvania, USA.

John Pavlopoulos and Ion Androutsopoulos. 2014a. Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method. In *Proceedings of LASM-EACL*, pages 44–52, Gothenburg, Sweden.

John Pavlopoulos and Ion Androutsopoulos. 2014b. Multi-granular aspect aggregation in aspect-based sentiment analysis. In *Proceedings of EACL*, pages 78–87, Gothenburg, Sweden.

Stelios Piperidis. 2012. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of LREC-2012*, pages 36–42, Istanbul, Turkey.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP*, pages 339–346, Vancouver, British Columbia, Canada.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of EACL*, pages 102–107, Avignon, France.

Tun Thura Thet, Jin-Cheon Na, and Christopher S. G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Information Science*, 36(6):823–848.

Ivan Titov and Ryan T. McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL*, pages 308–316, Columbus, Ohio, USA.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*, pages 417–424, Philadelphia, Pennsylvania, USA.