**SemEval 2016, Task 5**

**Aspect Based Sentiment Analysis (ABSA)**

**Evaluation**

**ABSA 2016 task**

**Subtask 1 (SB1): Sentence-Level ABSA**

Given a review text about a laptop, restaurant, hotel (or other entity) the goal in ABSA 2016 is to identify tuples (at the sentence level) that contain the following types of information[[1]](#footnote-1):

* **Slot 1**: Aspect Category (Entity and Attribute). Identify every entity E and attribute A pair E#A towards which an opinion is expressed in the given text. Each E#A pair defines an aspect category of the given text.
* **Slot 2**: Opinion Target Expression (OTE). An opinion target expression (OTE) is an expression used in the given text to refer to the reviewed entity E of a pair E#A[[2]](#footnote-2).
* **Slot 3:** Sentiment Polarity. Each identified <category, target> pair has to be assigned a polarity, from a set P = {positive, negative, neutral}.

Examples of opinion tuples for the restaurant domain are shown in Fig 1.

|  |
| --- |
| Review id:"1004293"  Judging from previous posts this used to be a good place, but not any longer.  {category:"RESTAURANT#GENERAL", target:"place", from:"51", to="56", polarity:"negative" }  We, there were four of us, arrived at noon - the place was empty - and the staff acted  like we were imposing on them and they were very rude.  {category:"SERVICE#GENERAL", target:"staff", from:"75", to:"80", polarity:"negative" }  They never brought us complimentary noodles, ignored repeated requests for sugar,  and threw our dishes on the table.  {category:"SERVICE#GENERAL", target:"NULL", from:"-", to:"-", polarity:"negative"}  The food was lousy - too sweet or too salty and the portions tiny.  {category="FOOD#QUALITY", target:"food", from:"4", to:"8", polarity="negative" }  {category:"FOOD#STYLE\_OPTIONS", target:"portions", from:"52", to:"60", polarity:"negative" }  After all that, they complained to me about the small tip.  {category:"SERVICE#GENERAL", target:"NULL", from:"-", to:"-", polarity:"negative" }    Avoid this place!  {category:"RESTAURANT#GENERAL", target:"place", from:"11", to:"16", polarity:"negative"} |

Figure 1: ABSA 2016 opinion tuples for a restaurant review.

**Subtask 2 (SB2): Text-Level ABSA**

The target is to identify <Slot1,Slot3> tuples at the text level. For example for the previous review the following tuples should be identified.

{category:"RESTAURANT#GENERAL", polarity:"negative" }

{category:"SERVICE#GENERAL", polarity:"negative" }

{category:"FOOD#QUALITY", polarity:"negative" }

{category:"FOOD#STYLE\_OPTIONS", polarity:"negative" }

**Evaluation**

**Subtask 1 (SB1)**

**Slot 1:** The evaluation assesses whether a system identifies and returns the set of aspect categories towards which an opinion is expressed. In particular, precision, recall and F-1 scores are calculated by comparing the list of the categories that a system returned (for a sentence) to the corresponding gold list. These lists are constructed by extracting the values of Slot 1 (category). For example for the 4th sentence of Fig 1 the list is {(FOOD#QUALITY), (FOOD#STYLE\_OPTIONS)}. The calculation ignores duplicate occurrences of categories. For example, for the following sentence the categories list is {(FOOD#QUALITY)}.

|  |
| --- |
| Furthermore, the rice had no seasoning, so the sushi was bland and disgusting.  {category="FOOD#QUALITY", target="rice, from="17", to="21", polarity="negative"}  {category="FOOD#QUALITY", target="sushi", from="47", to="52", polarity="negative"} |

You can evaluate your system in category extraction by running the following command[[3]](#footnote-3).

java -cp ./A.jar absa16.Do Eval -prd pred.xml -gld teGld.xml -evs 1 -phs A -sbt SB1

pred.xml contains the predicted annotations and teGld.xml the gold annotations. Both xml files should be in the same format as the provided training data[[4]](#footnote-4).

**Slot 2:** The evaluation assesses whether a system identifies and returns the set of targets, i.e. the expressions that are used in a sentence to refer to the reviewed entities. In particular, precision, recall and F-1 scores are calculated by comparing the list of the targets that a system returned (for a sentence) to the corresponding gold list. These lists are constructed using the target offsets. For example for the 4th sentence of Fig 1 the extracted list is {(4, 8), (52, 60)}. The calculation discards NULL targets since they do not correspond to explicit target mentions. For example, for the following sentence the constructed list is {(19,29)}.

|  |
| --- |
| Terrible, terrible management - deserves to be shut-down.  {category="SERVICE#GENERAL", target="management", from="19" to="29", polarity="negative"}  {category="RESTAURANT#GENERAL", target=”NULL”, from="-", to="-", polarity="negative"} |

Duplicate targets are also ignored[[5]](#footnote-5), so for the next sentence the target list is {(51, 55)}

|  |
| --- |
| I expected quite a bit more from such an expensive menu.  {category="FOOD#PRICES ", target=" menu ", from="51", to="55", polarity=" negative"}  {category="FOOD#QUALITY", target=" menu ", from="51", to="55", polarity=" negative"} |

You can evaluate your system in target extraction by running the following command.

java -cp ./A.jar absa16.Do Eval -prd pred.xml -gld teGld.xml -evs 2 -phs A -sbt SB1

**Slot 1&2:** <category, target> evaluation assesses whether a system identifies the targets, the aspects categories and constructs the corresponding tuples. Again precision, recall and F1 scores are calculated by comparing the <category, target> tuples of a system to the gold ones. You can evaluate your system in <category, target> extraction by running the following command.

java -cp ./A.jar absa16.Do Eval -prd pred.xml -gld teGld.xml -evs 3 -phs A -sbt SB1

The compared lists in this case contain the target offsets and the category values. For example for the 4th sentence this list is the following.

{(FOOD#QUALITY,4,8), (FOOD#STYLE\_OPTIONS,52,60)}.

**Slot 3:** For polarity classification evaluation we use the total accuracy score. To evaluate your system run the command shown below. The pred.xml should contain the gold annotations for the categories and targets (in the same order as in teGld.xml) and the corresponding predicted polarities.

java -cp ./A.jar absa16.Do Eval -prd pred.xml -gld teGld.xml -evs 5 -phs B -sbt SB1

The program will print the total accuracy score as well as precision, recall and F1 scores for each polarity label (positive, negative, neutral).

**Subtask 2 (SB2)**

For evaluating your system in SB2 respective commands are provided:

**Slot 1:**

java -cp ./A.jar absa16.Do Eval -prd pred.xml -gld teGld.xml -evs 1 -phs A -sbt SB2

**Slot 3:**

java -cp ./A.jar absa16.Do Eval -prd pred.xml -gld teGld.xml -evs 5 -phs B -sbt SB2

Again as in SB1 pred.xml should contain the gold annotations for the categories and targets in the same order as in gold annotations file (teGld.xml).

**Validation**

To check whether an xml file generated (for SB1 or SB2) by a system (e.g. pred.xml) is well formed and all the slots are filled with valid values you can run the command that is shown below. The first argument is the xml to be checked (pred.xml), the second is an xsd file (ABSA16.xsd) and the third the domain {lapt, rest, hote, phns, came}.

java -cp ./A.jar absa16.Do Validate ./pred.xml ./ABSA16.xsd lapt

The script validates the xml against the xsd and checks the slot values. For example, if slot 1 (category) is filled with a value that does not correspond to the E,A inventories[[6]](#footnote-6) of the domain a relevant message will be printed. Similarly, if slot 3 (polarity) is assigned a value not belonging to the set P = {positive, negative, neutral, conflict} a relevant message will also be printed.

1. More information about the task in http://alt.qcri.org/semeval2016/task5/ [↑](#footnote-ref-1)
2. OTE is not required in some domains (e.g. laptops). [↑](#footnote-ref-2)
3. A.jar is included in the package that is provided with this document. Java 1.8 is required. [↑](#footnote-ref-3)
4. http://alt.qcri.org/semeval2016/task5/ [↑](#footnote-ref-4)
5. A target is defined by its starting and ending offset. [↑](#footnote-ref-5)
6. For more information on E,A inventories see the guidelines that are provided in ABSA-16 site. [↑](#footnote-ref-6)