Charles University in Prague, Faculty of Mathematics and Physics

Manual for Annotation

Ondřej Fiala

2015

1 Preface

This text is intended as a guide for annotators and should lead them through a huge number of text lying everywhere around. Purpose of this text is not restrictive. Annotating data in sentiment analysis field is still an objective thing. This document should help to unite annotators.

Text is applicable on all datasets released with this document. Obtained data are from different domains, but their structure and purpose is similar.

Main focus is on Czech language, but there is a very small number of language specific material. It can be used for all languages from similar language groups.[1]

2 Introduction

This document is intended for annotators of data from the sentiment analysis field. It is a part of natural language processing (NLP) area and it belongs to linguistics. Another related areas are opinion detection, information retrieval and data extraction.

Annotator have to look for sentences with sentiment and find evaluative phrases and target of this phrase. *Evaluative phrase* is a group or words expressing author's opinion on some object. *Target* is that object. The goal of sentiment analysis is to identify all these entities in a text. Aspect is a part of object, or it is object's property. E.g. battery of a notebook and ease of use.

[1] One of the few limitations is requirement for continuity of tagged text chunks. This can be overcome by adding identification numbers to all related tags, or limiting number of tags in one sentence. E.g. one target tag in one sentence.

Main advice is simple, **be as objective as possible**. Don't judge author complaining about local post office, described fancy colours or interesting language structure. Don't judge very product from own perspective, of course.

All examples contain root node `<text>Content text</text>`. This is just for practical reason. It is separating one example from another. In practise, reviews contains much more text and this node can be missing or replaced with another root node.

3 Purpose of annotation

Annotated text is used for unsupervised algorithms to learn principles of language. It can also help in searching for special cases and constructing system based on rules. These are the most common cases for these datasets.

There exist already big number of approaches for detecting targets of evaluative phrases. The most simple ones are just listing all occurring nouns. The most advanced approaches are focusing on part of speech and expressions with hazy meaning. For example pronouns. The big amount of work si dedicated to resolve these pronouns.

Resolution of connection between targets and phrases is usually done simply by their distance. The more sophisticated approach use morphological parsing for this relationship.

4 Conditions influencing annotator

The main idea that should fulfil annotators mind is the purpose of the final annotated data. What they will be used for and what text features are usually hard to detect.

It ic very helpful to have a look at the data. Annotator should first annotate some reviews. Then discus these results with this document or other annotators preferably. Few days break is a good time to occasionally thing about polarity of text that annotator is accessing every day. Then return to practise work and real annotation follows.

Properties are changing over the time. Especially in electronics domain. Annotation should not rate product by himself or from modern point of view. He should keep authors opinion.

Annotator can emerge very strange language structures and phenomena. Some reviews are barely understandable. Others are containing only one word. This is the real language used on the Internet and annotation helps to understand this language. Any language construction should not set annotator fly off the handle.[2]

5 Tag structure

Final structure with all possible elements annotated follows. As we can see, tags are not overlapping an it is recommended practise.

[2] It is reasonable to note down all interesting language constructions for a future work and new ideas.

```
<review>
  <text>Tablet má <rate>nadstandardní</rate> <target>mechanické
     provedení</target>.
  </text>
  <summary>11,6" ASUS Transformer Trio bych sice neoznačil jako
     revoluční zařízení, originalitu mu však nelze upřít.
  <summary>
  <positive_summary>kvalitní IPS displej</positive_summary>
  <negative_summary>krátká výdrž na jedno nabití</negative_summary>
</review>
```

6 Tags

There is a list of all tags used for annotation. Datasets contains also other tags, e.g. *url*, *title*, *text*, *h1*. This tags have its own clear purpose and are not connected with sentiment analysis.

target Object of evaluative phrase.

positive Phrase with positive sentiment.

negative Phrase with negative sentiment.

summary Summary of the review text.

positive_summary The most positive aspects.

negative_summary The most negative aspects.

6.1 Target

Usually object of the sentence is also target of evaluative phrase. It is not intended to limit targets to nouns only. Everything that author have opinion about and exactly states it is a target. Annotating pronouns is not common in systems without context resolving.

Gold rule for finding target is **annotated evaluative phrase implies annotation of target**. It is possible, that target is not part of this phrase. When annotator is in doubt about potential target, this rule helps to annotate it.

6.2 Evaluative phrase

Positive and negative phrases. They usually express emotions and authors opinion. They are not measuring length or other unit. Only if it is exactly expressed. Phrase "car is long" is neutral, but phrase "big nice screen" can mark "big" as positive.

6.3 Summary

This tag, with other summaries usually comes from review structure. It refreshes important point from a review and it can be the last paragraph in a review.

6.4 Polarity summary

The most important aspect of described object. Usually it exists with evaluative phrase and it is occurring multiple times in one group.

7 Note about annotation summary

Summary and polarity summaries (positive summary and negative summary) are different entities. It is possible, that polarity summaries are embedded in summary, but it is not the preferred way. It is better to not include any polarity into summary.

8 Examples

Including necessary words into evaluative phrase is a good practise. Especially words weakening or intensifying this phrase belong there.

```
<text>Na můj vkus je <target>klávesnice</target> <negative>možná příliš
   hlučná</negative>.
</text>
```

Very common example is a conjugation. It is supposed do tag all entities separately.

```
<text>Zápory: <negative_summary>váha</negative_summary> a
   <negative_summary>cena</negative_summary>
</text>
```

Comma separated list of targets implies a big number of separate tags. Note, that there is tagged positive summary and not a plain text of a review!

```
<text>Klady: <positive_summary>mobilita</positive_summary>,
   <positive_summary>výkon</positive_summary>, výdrž
   baterie</positive_summary>, <positive_summary>image</positive_summary>
</text>
<text>Klady: <positive_summary>rozpoznávání textu</positive_summary>,
   <positive_summary>rozměry a hmotnost (ve srovnání s
   notebooky)</positive_summary>
</text>
```

Following example is not the same case. Keyboard buttons forms one rated group.

```
<text>Zápory: <negative_summary>absence samostatných kláves Home, End,
   PgUp, PgDn</negative_summary>
</text>
```

Next example talks about bad manipulation with both parts of a product.

```
<text>Zápory: <negative_summary>horší manipulace s bateriemi a
   MultiBay</negative_summary>
</text>
```

Colours can be a problem. Sometimes it is recommended to tag just word "barva", because it is an aspect of a object. It have to be tagged, if there is only a name of the colour and this colour is evaluated.

```
<text><negative>Hnusná</negative> <target>šedá</target>.</text>
<text>seda <target>barva</target></text>
```

Obviously, all annotated text so far is a good example. It is also not definitive. Discussion between annotator can change final rules.

8.1 Borderline cases

It is usually hard to differentiate same aspect described once with a noun and once with an adjective. This should not be a surprise> Target can transform into evaluative phrase in the next sentence.

```
<text>Jeho <target>funkce</target> se mi velmi líbí. Jedná se o velmi
   <positive>funkční</positive> <target>tlačítko</target>.
</text>
```

It is hard to annotate authors expectations and prejudices. Sentence can be slightly negative, but it is not reviewing a product. These expressions don't need to be tagged.

```
<negative_summary>Čekal jsem že bude silnější  Čekal jsem že bude
   poslacený <target>Jack konektor</target>  Ty zápory to je uplný prd.
</negative_summary>
```

One aspect can have its own aspects. E.g. notebook has a display and this display has measurable sharpness. Resolving these connections is very hard. Annotator can decide about tagging this features. Following example breaks one of the main rules. It is not a big error this time.

```
<text><positive>Kvalitni</positive>
   <target>display</target>,<positive>vyborna</positive> ostrost
</text>
```

Example of the last, the most tricky phrase. Author used plus sign and damaged suggested system of conjugations. It is possible, that it is a mathematical equation with missing brackets.

```
<text>Klady: <positive_summary>poměr cena / výkon +
   výbava</positive_summary>
</text>
```

9 Tools

Used structure is very easy to establish. It have only a few different tags, so annotator can modify his preferred annotation tool and enjoy the comfort he is used to.

Attachment to this work is a definition of macros for text editor called PSPad `http://www.pspad.com/`.[3] Macros don't have to be installed, it is required to copy them into program's directory. List of keyboard shortcuts is located together with macros.

[3] PSPad is an old and popular freeware editor for programmers written by Jan Fiala