

IN5550, Spring 2021 Home Exam

Task Description

Targeted Sentiment Analysis

Area chairs: Jeremy Barnes

Introduction

This document introduces one of the tasks for the Spring 2021 Home Exam for IN5550: Targeted Sentiment Analysis for Norwegian. For general instructions regarding the home exam, see the information at the semester page for the course:

<https://www.uio.no/studier/emner/matnat/ifi/IN5550/v21/exam.html>

The task in short

Fine-grained Sentiment Analysis (SA), sometimes referred to as Opinion Analysis/Mining, is the task of identifying opinions in text and analyzing them in terms of their polar expressions, targets, and holders. In this task we will focus on targeted SA, i.e. the identification of the target of opinion along with the polarity with which it is associated in the text (positive/negative). In the example below, for instance, the target of the opinion is *disken* ‘the disk’ and it is ascribed a positive polarity by the surrounding context.

- (1) *Denne disken_{POS} er svært stillegående*
This disk is very quiet-going
‘This disk runs very quietly’

All data and pre-code needed to work on this assignment is available from:

https://github.com/uio-no/in5550/2021/tree/master/exam/targeted_sa

Data

We will be working with the recently released NoReC_{fine}, a dataset for fine-grained sentiment analysis in Norwegian. The texts in the dataset have been annotated with respect to polar expressions, targets and holders of opinion but we will here be focusing on identification of targets and their polarity only. The underlying texts are taken from a corpus of professionally authored reviews from multiple news-sources and across a wide variety of domains, including literature, games, music, products, movies and more. Table 1 presents the dataset and its annotated targets. The dataset is distributed with pre-defined train, development and test splits.

	# Examples				
	Train	Dev.	Test	Total	Avg. len.
Sents.	5914	1151	895	7960	16.8
Targets	3339	629	511	4479	2.0

Table 1: Number of sentences and annotated targets across the data splits.

Data format

The task repository contains data that has been converted from the native json-format of NoReC_{fine} to the conll-format assumed for this task: each line is a token and label, separated by a tab, and sentences are separated by a new line. The labels are BIO + polarity (Positive, Negative) for a total of 5 labels (B-targ-Positive, I-targ-Positive, B-targ-Negative, I-targ-Negative, O).

```
# sent_id = 501595-13-04
Munken      B-targ-Positive
Bistro      I-targ-Positive
er          O
en          O
hyggelig    O
nabolagsrestaurant O
for         O
hverdagslige O
og          O
uformelle   O
anledninger O
.
```

Modeling

The main objective of the home exam is to train a neural system to perform targeted sentiment analysis for Norwegian text. In order to complete the task you should follow these steps:

Baseline model You can base your work on PyTorch pre-code for a baseline model and evaluate this on the development and test data. This is a simple bi-LSTM model that leaves room for a number of possible improvements.

Experimental evaluation You should experimentally evaluate the effect of at least three different changes to your basic system. Some possible directions for further experimentation are provided below, but you are also free to come up with experimental directions of your own. Evaluation of changes to your systems should be performed on the development set.

Held-out testing The best configuration of your system following experimentation should be evaluated on the test set.

Write a report Your experiments should be described in a report following the exam template detailing your experiments and findings.

Evaluation

The models will be evaluated on two different metrics: proportional F_1 and binary F_1 . Binary Overlap counts any overlapping predicted and gold span as correct. Proportional Overlap instead assigns precision as the ratio of overlap with the predicted span and recall as the ratio of overlap with the gold span, which reduces to token-level F_1 . Proportional F_1 is therefore a stricter measure than Binary F_1 . You will have scripts available to calculate these scores.

Possible directions for experimentation

You can explore a number of directions we suggest below, but you're encouraged to come up with other ideas for yourself.

1. Instead of using the collapsed BIO+polarity labels, first predict the BIO labels and then predict the polarity of the extracted targets.
2. Impact of different architectures:
 - LSTM vs. GRU vs. Transformer
 - Include a character-level RNN/CNN and concatenate the output to the token embeddings.
 - Depth of model (2-layer, 3-layer, etc)
 - Include POS or Dependency information (by embedding labels and concatenating to token embeddings)
3. Effect of using pretrained models from NorLM¹ (ELMo or BERT), or Multilingual Bert², or XLM-Roberta.³
4. Perform a small error analysis (confusion matrix, the most common errors).

Recommended reading

1. **Neural network methods for natural language processing.** Goldberg, Y., 2017.⁴
2. **A Fine-Grained Sentiment Dataset for Norwegian.** Øvreliid, L., Mæhlum, P., Barnes, J. & Velldal, E., 2020.⁵
3. **Open Domain Targeted Sentiment** Margaret Mitchell, Jacqueline Aguilar, Theresa Wilson, Benjamin Van Durme⁶

¹<http://wiki.nlpl.eu/Vectors/norlm>

²<https://github.com/google-research/bert/blob/master/multilingual.md>

³https://huggingface.co/transformers/model_doc/xlmroberta.html

⁴<https://www.morganclaypool.com/doi/10.2200/S00762ED1V01Y201703HLT037>

⁵<https://arxiv.org/pdf/1911.12722.pdf>

⁶<https://www.aclweb.org/anthology/D13-1171.pdf>

4. **Neural Networks for Open Domain Targeted Sentiment.** Meishan Zhang, Yue Zhang, and Duy-Tin Vo ⁷

⁷<https://www.aclweb.org/anthology/D15-1073.pdf>