<DCSE> DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Ing. Jakub Šmíd

Master's Thesis

Computer Science and Engineering
Software Engineering (NLP)
2022/2023

Thesis supervisor:
Ing. Pavel Přibáň

Cross–lingual Aspect–Based Sentiment Analysis

Abstract

This thesis focuses on cross-lingual aspect-based sentiment analysis (ABSA), an understudied area in contrast to monolingual ABSA. This thesis proposes two methods that can be used with prompting and traditional-based fine-tuning. The first method is a sequence-to-sequence method that solves multiple ABSA tasks simultaneously and outperforms previous state-of-the-art results on benchmark datasets in multiple languages. Prompting improves the performance of the T5 model and its multilingual version significantly, which resulted in the best overall results among the tested models. The best cross-lingual results are also promising. The second method classifies the sentiment polarity of aspect terms and categories, establishing new state-of-the-art results in multiple languages and achieving excellent cross-lingual results, often within 2% of monolingual results. In addition, this thesis presents a newly annotated Czech dataset for ABSA.

Introduction

This thesis focuses on cross-lingual aspect-based sentiment analysis (ABSA) and explores state-of-the-art transfer learning methods for addressing ABSA and cross-lingual ABSA. The main goal is to propose effective methods for solving multiple ABSA tasks and evaluate their performance across multiple languages. Additionally, the thesis investigates zero-shot cross-lingual settings, where the model is not fine-tuned on any data from the target language.

Aspect–Based Sentiment Analysis

Aspect-based sentiment analysis is a natural language processing task that aims to identify the sentiment of each aspect of a product or service. Multiple sentiment elements are involved in ABSA.
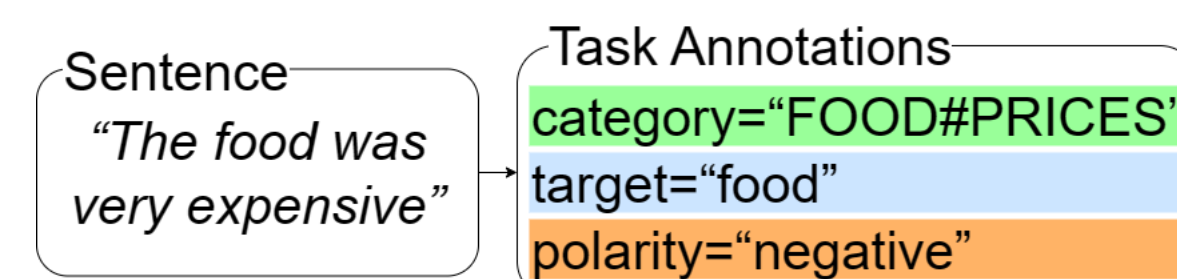

Aspect-based sentiment analysis.

Cross-lingual ABSA aims to transfer knowledge from one language to another.

Proposed Solution

This thesis proposes a sequence-to-sequence method that addresses three ABSA tasks simultaneously: target-aspect-sentiment detection (TASD), aspect term extraction (ATE), and aspect category detection (ACD). ATE identifies aspect terms, ACD detects aspect categories, and TASD identifies all (aspect term, aspect category, sentiment polarity) triplets. The method transforms each sentiment triplet (c, a, p) into a natural language phrase *"$P_c(c)$ is $P_p(p)$, given the expression: $P_a(a)$"*, where $P_z(z)$ is a projection function that maps the sentiment element z from its original format to a natural language form.

The thesis also proposes a method for classifying the sentiment polarity of aspect terms and categories.

Both methods can be used with prompting and traditional fine-tuning. Prompting encourages a pre-trained model to make specific predictions by providing a prompt specifying the task to be done.
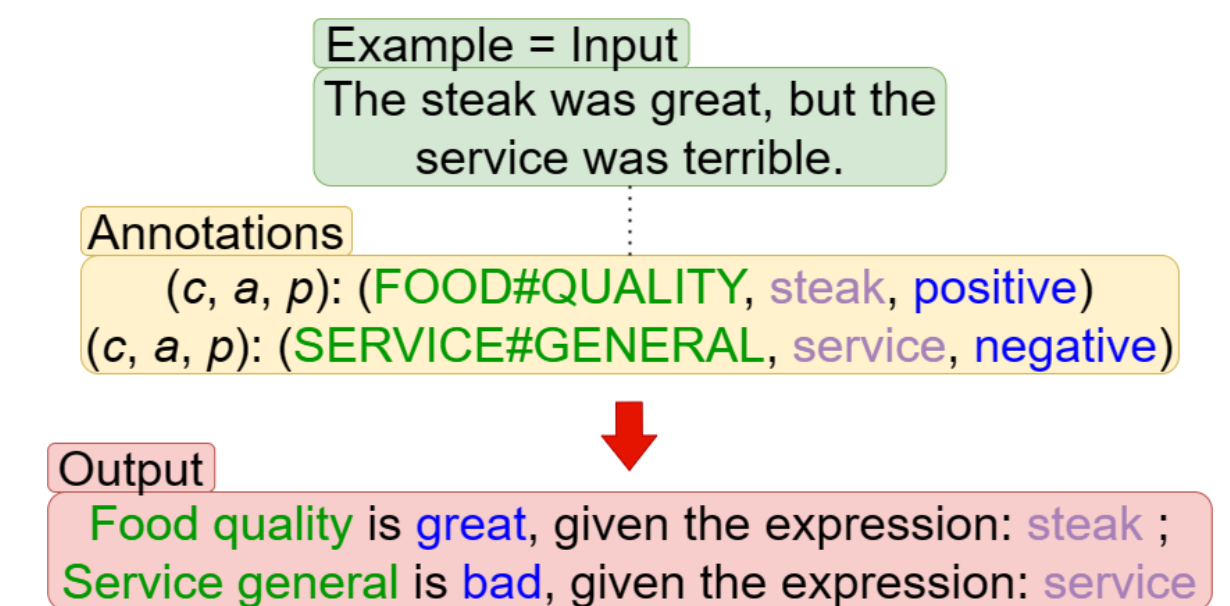
Label construction.

Achieved Results

The proposed methods were evaluated using multiple Transformer-based models, outperforming previous state-of-the-art results on benchmark datasets in multiple languages.

In the sequence-to-sequence method, the T5 model and its multilingual version (mT5) achieve the best overall results when used with prompting, significantly improving their performance. The cross-lingual results are promising, often within 5 to 15% of the monolingual results, depending on the specific task and the combination of source and target languages.

The XLM-RoBERTa model achieves the best results for sentiment polarity classification. The cross-lingual results are excellent, often within 2% of the monolingual results.

The table shows the best achieved micro F1 scores in percentages for each task and language combination, with the monolingual results for a given target language in parentheses.

Lang	ACD	ATE	TASD	Polarity
cs-en	79.46 (87.20)	73.80 (86.85)	58.22 (73.59)	91.06 (95.01)
es-en	80.76 (87.20)	72.50 (86.85)	58.60 (73.59)	93.08 (95.01)
fr-en	84.73 (87.20)	77.70 (86.85)	64.95 (73.59)	93.91 (95.01)
nl-en	79.67 (87.20)	67.14 (86.85)	53.86 (73.59)	92.77 (95.01)
ru-en	81.00 (87.20)	76.40 (86.85)	62.50 (73.59)	93.60 (95.01)
tr-en	77.33 (87.20)	64.46 (86.85)	52.87 (73.59)	92.30 (95.01)
en-cs	80.21 (85.45)	69.20 (84.80)	52.57 (67.30)	89.99 (88.59)
en-es	78.60 (85.53)	71.95 (80.66)	56.61 (68.03)	93.00 (94.29)
en-fr	78.19 (80.77)	74.44 (80.38)	53.21 (60.56)	86.67 (90.31)
en-nl	78.37 (81.43)	64.94 (81.16)	54.70 (64.17)	91.32 (91.55)
en-ru	83.26 (86.79)	66.98 (81.88)	55.22 (68.61)	90.44 (91.30)
en-tr	82.15 (83.31)	57.50 (71.95)	42.86 (54.40)	93.43 (93.78)

Conclusion

This thesis proposes methods for solving ABSA tasks that can be used with prompting and traditional fine-tuning. The sequence-to-sequence method can address multiple ABSA tasks simultaneously. Our proposed methods achieve new state-of-the-art results on benchmark datasets in multiple languages. Prompting is especially effective for the sequence-to-sequence method when used with the T5 model and its multilingual version. The cross-lingual results are promising, often within a few per cent of monolingual results. Additionally, we introduce a new Czech dataset for ABSA. Overall, this thesis significantly contributes to the cross-lingual ABSA field.

FACULTY OF APPLIED SCIENCES
UNIVERSITY OF WEST BOHEMIA