Czech Dataset for Cross-lingual Subjectivity Classification

Pavel Přibáň[1,2], and Josef Steinberger[1,2]

[1] Department of Computer Science and Engineering, Faculty of Applied Sciences,
[2] NTIS – New Technologies for the Information Society, Faculty of Applied Sciences,
University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic

E-mail: pribanp@kiv.zcu.cz Web: nlp.kiv.zcu.cz

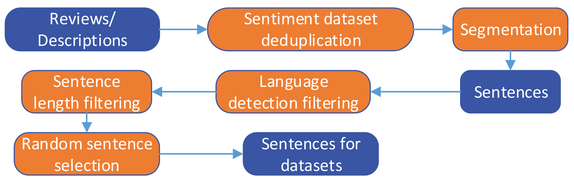
Motivation & Goal

- **Motivation**
  - No Czech dataset for SC
  - English dataset often used for evaluation
- **Goal**
  - Provide a reliable Czech dataset
  - Allows cross-lingual benchmark

Dataset Building

- **Cleaning and Obtaining Data**
  - Downloaded 4M reviews and 735k movie descriptions[a]
  - Remove sentiment 74k reviews contained in (Habernal et al. 2013)
  - Split text into sentences with UDPipe 2 (Straka, 2018)
  - Language filtering
  - Sentence length filtering – at least 6 tokens long.

⇒ **884k** description and **19M** review sentences
⇒ **20k** randomly selected for annotation from desc. and reviews



[a]from Czech Movie Database[b]

Dataset Annotation

- Manual annotation by 2 Czech native speakers
- **Four phases:**
  1. **Annotation of 100 sentences** with one of three labels: `subjective`, `objective` and `trash`
  2. **Conflict discussion and resolving**
     ⇒ labels extended by `unsure` and `question`
  3. **Annotation of 2,034 same sentences**
     ⇒ The Cohen's κ 0.68 for all five labels
     ⇒ 1,668 sentences only for objective and subjective, 0.83 Cohen's κ
  4. **Final annotations of almost 5000** sentences by each annotator
- At most 15 review sent. and 3 description sent. per movie

Subjectivity Classification

- Classify text (sentence) as **subjective** or **objective**

"*I liked the movie itself, but it didn't surprise me.*"

⇒ subjective

"*Maurice lives and works in the south of France.*"

⇒ objective

Annotation Statistics

Label	Reviews	Descriptions	Total
unsure	866 / 13.11%	457 / 8.62%	1 323
object.	**726 / 10.99%**	**4 464 / 84.22%**	**5 190**
subj.	**4 794 / 72.57%**	**208 / 3.92%**	**5 002**
quest.	114 / 1.73%	128 / 2.41%	242
trash	106 / 1.60%	44 / 0.83%	150
Total	6 606 / 100%	5 301 / 100%	11 907

Table 1: Annotation statistics for subjective and objective

- 11,907 annotated sentences
- Considerable % of sentences in reviews are not subjective (only 72.57% subjective)
- Relatively large part of non-objective sentences movie descriptions (84.22% of the sentences are objective)
- **Manual annotation is important for dataset quality**
- For easier comparison and evaluation of experiments we provide precisely balanced dataset of **10,000 sentences** (5k objective, 5k subjective)
  ⇒ called `Subj-CS`

Automatically Labeled Dataset

- Large dataset `Subj-CS-L` obtained in a distant supervised way
- Following idea from (Pang and Lee, 2004)
  - Review sentences ⇒ **subjective**
  - Description sentences ⇒ **objective**
- Automatically labeled 200k sentences
  - 100k subjective
  - 100k objective
  - At least six tokens
  - Potential unsupervised pre-training

Data for Experiments

- English subjectivity dataset from (Pang and Lee, 2004)

Dataset	Name	Subjective	Objective	Total
Subj-CS	cs-train	3 750	3 750	7 500
	cs-dev	250	250	500
	cs-test	1 000	1 000	2 000
		5 000	5 000	10 000
Subj-CS-L	cs-L-train	95 000	95 000	190 000
	cs-L-dev	5 000	5 000	10 000
		100 000	100 000	200 000
Subj-EN	en-train	3 764	3 736	7 500
	en-dev	231	269	500
	en-test	1 005	995	2 000
		5 000	5 000	10 000

Table 2: Datasets statistics.

Experiments

- Monolingual experiments
  - Set a baseline for newly created dataset
  - Verify the dataset quality
- Zero-shot cross-lingual classification
  - Test usability of the dataset as cross-lingual benchmark
  - Test the ability of multilingual models to transfer knowledge between Czech and English

Cross-lingual Results

Model	EN → CS		Monoling. (cs-train)
	en-dev	cs-test	cs-test
mBERT	95.38 ± 0.22	86.18 ± 0.33	91.23 ± 0.21
XLM-R-Large	97.60 ± 0.18	**90.75 ± 0.32**	93.56 ± 0.13

Table 4: Accuracy results for experiments from English to Czech.

Cross-lingual Results II

Model	CS → EN (cs-train)		CS → EN (cs-L-train)		Monolingual (en-train)
	cs-dev	en-test	en-dev	en-test	en-test
mBERT	92.11 ± 0.38	88.99 ± 0.94	85.80 ± 0.89	85.53 ± 0.98	95.87 ± 0.13
XLM-R-Large	94.40 ± 0.36	**92.86 ± 0.44**	93.35 ± 0.22	**90.98 ± 0.26**	97.28 ± 0.07

Table 7: Accuracy results for cross-lingual experiments from Czech to English.

Model	Joint (cs-train + en-train)		Monolingual (cs-train)	Monolingual (en-train)
	cs-test	en-test	cs-test	en-test
mBERT	91.12 ± 0.24	95.69 ± 0.22	91.23 ± 0.21	95.87 ± 0.13
XLM-R-Large	**93.85 ± 0.15**	**96.95 ± 0.12**	93.56 ± 0.13	97.28 ± 0.07

Table 8: Accuracy results for models jointly trained on English and Czech.

Transformers Models

Model	#Params	Vocab	#Langs
Czech Electra	13M	30k	1
Czert-B	110M	30k	1
RobeCzech	125M	52k	1
BERT	110M	30k	1
mBERT	177M	120k	104
XLM-R-Large	559M	250k	100

Table 3: A comparison of used models.

Dataset & Source Code

https://github.com/pauli31/czech-subjectivity-dataset

Monolingual Results

Model	Subj-CS (cs-train) cs-test	Subj-CS-L (cs-L-train) cs-test
Czech Electra	91.85 ± 0.27	91.21 ± 0.08
Czert-B	92.85 ± 0.20	91.79 ± 0.07*
RobeCzech	93.29 ± 0.19*	91.63 ± 0.08
mBERT	91.23 ± 0.21	91.14 ± 0.11
XLM-R-Large	**93.56 ± 0.13**	**91.96 ± 0.10**

Table 5: Czech monolingual results as average accuracy.

Model	en-test	en-10-fold
BERT	96.55 ± 0.16	96.87 ± 0.25
mBERT	95.87 ± 0.13	96.03 ± 0.24
XLM-R-Large	**97.28 ± 0.07**	**97.34 ± 0.21**
(Wang et al., 2021)†	97.40 ± 0.10*	-
(Nandi et al., 2021)	-	97.30
(Zhao et al., 2015)	-	95.50
(Amplayo et al., 2018)	-	94.80
(Khodak et al., 2018)	-	94.70
(Reimers and Gurevych, 2019)	-	94.52

Table 6: English monolingual results as average accuracy.

Conclusion

We introduce the first Czech subjectivity dataset `Subj-CS` that consists of 10k manually annotated subjective and objective sentences from movie reviews and descriptions.

We perform a series of monolingual experiments with five pre-trained BERT-like models to obtain the baseline results for the newly created Czech dataset and we are able to achieve 93.5%6 of accuracy with the XLM-R-Large model. We also perform monolingual experiments for the existing English subjectivity dataset with three models obtaining 97.28% of accuracy, which is on par with the current state-of-the-art results for this dataset. Finally, we conduct zero-shot cross-lingual subjectivity classification to verify the usability of our dataset as the cross-lingual benchmark for pre-trained multilingual models that allow transfer learning.

Our experiments confirm that we provide a dataset of relatively high quality and it can be used as an evaluation benchmark to test the ability of pre-trained models to transfer knowledge between Czech and English.