CLUZH at VarDial GDI 2017: Testing a Variety of Machine Learning Tools for the Classification of Swiss German Dialects

Simon Clematide and Peter Makarov

A Supervised Classification Problem

Classes	BE	%	BS	%	LU	%	ZH	%	Total
Training	3889	**27**	3411	24	3214	22	3964	**27**	14478
Test	906	25	939	26	916	25	877	*24*	3638
Δ		-2		+2		+3		-3	

- Roughly balanced data sets, but differently biased for train and test set.
- Train and test set not iid (disjoint speakers)
- Transcriber bias: test set transcriber for LU not seen in training

Transcription Examples

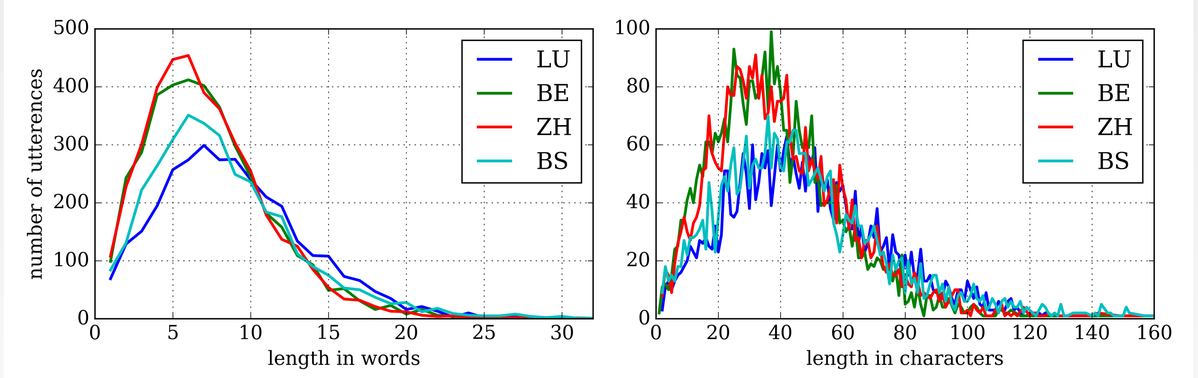
(BE) *a* *das* *chamì* *natüürlich* *aus* *nìd* *erinnere*
(de) an das kann ich mich natürlich alles nicht erinnern
(en) 'of cause I can't remember all this'

(BS) *a* *dr* *a* *a* *das* *mag* *mi* *äu* *erinnere*
(de) an dr an an das kann ich mich auch erinnern
(en) 'I can remember him . . . this too'

(LU) *bì* *disen* *ìsch* *es* *plòff* *gsii* *ùnd* *gwaut*
(de) bei diesen ist es Bluff gewesen und Gewalt
(en) 'for these it was bluff and violence'

(ZH) *als* *schwiizer* *gfüült*
(de) als Schweizer gefühlt
(en) 'felt as a Swiss'

Number of Words per Utterance



Challenges

- Short sequences (14% have less than 4 words)
- Many infrequent tokens due to writing variations: out of 14,065 word types, 9,372 appear once, 2,032 twice, 929 three times
- Many diacritics for phonetic properties

Run 1: Naive Bayes (NB)

- Our baseline approach using character n-grams ($2 \geq n \geq 6$) with add-one smoothing (scikit-learn)
- Very strong validation set results, but rather weak on test set

Run 2: Conditional Random Fields (CRF)

- Linear-chain CRF: each utterance is a sequence of words (wapiti)
- Rich feature set for each word including word forms, prefix/suffix pairs (1-3 characters), n-grams ($1 \geq n \geq 6$), consonant/vowel word shapes (contribute 1 percentage point of accuracy)
- Optimal number of epochs (35) precomputed by cross-validation
- Not the best results on internal validation set, but it generalized better on test set than NB

Run 3: Ensemble (NB, CRF, SVM)

- Linear SVM with the same n-gram feature model as NB (scikit-learn)
- Majority voting scheme (defaulting to NB in case of ties)

Our Official Results (10 Participants)

Run	F1 (weighted)	F1 (macro)	Accuracy
Baseline			25.80
1st System MAZA	**66.24**	**66.34**	**68.06**
2nd System CECL	66.11	66.25	66.36
Our Run 1 (NB)	61.56	61.65	63.50
Our Run 2 (CRF)	(**3rd**) **65.31**	(**3rd**) **65.38**	67.07
Our Run 3 (ENS)	65.27	65.34	(**2nd**) **67.34**

Detailed Evaluation

	Run 1 (NB)				**Run 2** (CRF)				**Run 3** (ENSEMBLE)			
Σ	BE	BS	LU	ZH	BE	BS	LU	ZH	BE	BS	LU	ZH
Precision	73	*64*	66	57	73	66	73	61	71	69	**76**	61
Recall	66	66	*30*	92	69	74	34	92	73	73	32	**93**
F1	70	65	*42*	70	71	70	47	**74**	72	71	45	**74**

- ZH too greedy; LU difficult to identify; BE/BA more balanced
- Consistent with results of other participants

Experiments with LSTMs

Character-based Long Short-Term Memory Networks (keras) with a hidden layer of 90 neurons. Prediction using softmax over hidden values computed after seeing last character.

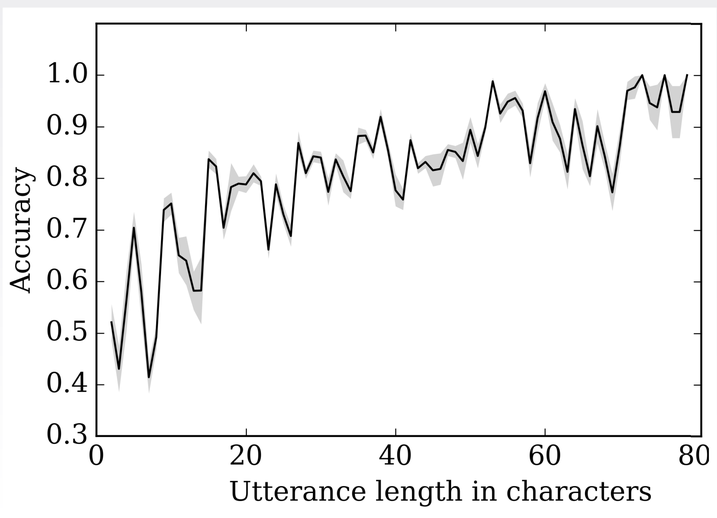
- ± Data augmentation of training set by splitting long utterances: works!
- ± Character embeddings: 2/3 of input size: might hurt. . .
- ± Replacement of diacritic characters: might work. . .

Internal Evaluation (10% of Training Data)

	Accuracy	F1 (macro)	F1 (weighted)
NB	85.43	85.36	85.44
CRF	85.01	85.02	85.01
SVM	82.39	82.36	82.39
ENSEMBLE (NB, CRF, SVM)	**85.50**	**85.42**	**85.50**
LSTM	83.49	83.30	83.46

Our best LSTM approach could not beat the other approaches.

Effect of Utterance Length (LSTMs)



As expected
the longer the sequence, the better the decision

Acknowledgments

Peter Makarov is supported by European Research Council Grant No. 338875.

VarDial Workshop EACL 2017, Valencia

Contact

University of Zurich[UZH]

Institute of Computational Linguistics
http://www.cl.uzh.ch

Simon Clematide
Andreastrasse 15, CH-8050 Zurich
simon.clematide@cl.uzh.ch