5th SwissText & 16th KONVENS Joint Conference 2020,
June 23-25 2020

Idiap Submission to Swiss German Language Detection Shared Task

[1]Shantipriya Parida, [1,2]Esau Villatoro Tello, [1]Qingran Zhan,
[1]Petr Motlicek, and [3]Sajit Kumar

[1]Idiap Research Institute, Martigny, Switzerland
[2]Universidad Autónoma Metropolitana Unidad Cuajimalpa, Mexico
[3]Indian Institute of Technology, Kharagpur, India

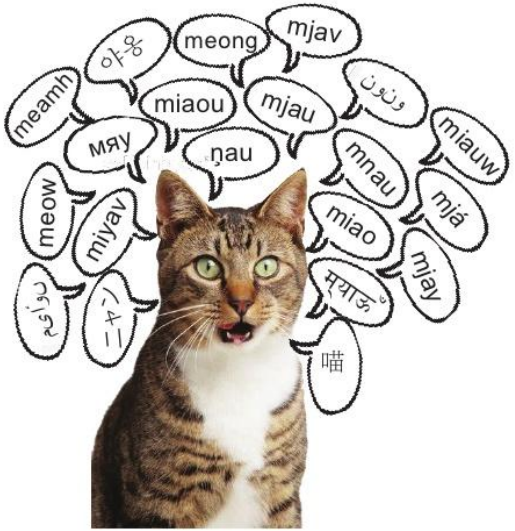
idiap RESEARCH INSTITUTE

UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Cuajimalpa

Contents

Introduction

- It is considerably challenging to detect languages that have similar origins or dialects (e.g. German dialect identification, Indo-Aryan language identification)
- It may not be possible to distinguish related dialects with very similar phoneme and grapheme inventories for some languages.

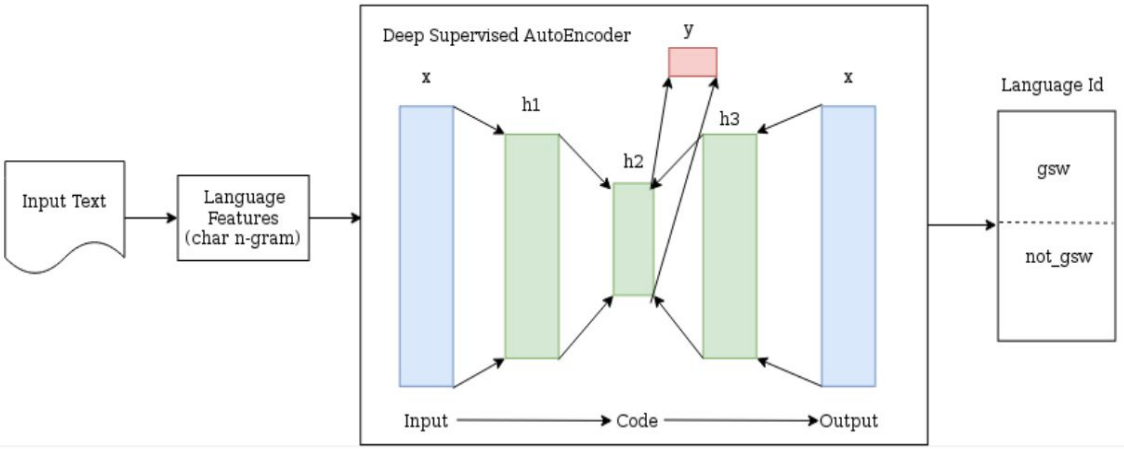


Introduction

- The Swiss German language detection shared task organized by the 2020 Germeval evaluation campaign.
- Participants asked to build a system that can automatically identify a snippet of text in Swiss German.
- Organizers supplied 2,000 Swiss German tweets and encouraged to use additional resource for training and 5,374 tweets to be classified.

Method Description

- We used character n-gram for extracting features from the input text.
- Extracted features are input to the deep supervised autoencoder (SAE).
- A SAE is an autoencoder with the addition of a supervised loss on the representation layer.
- Bayesian optimizer used for selecting the optimal hyperparameters.



Proposed model architecture

Dataset

We have created our training dataset by mixing data from the following:

- Swiss German language - NOAH, SwissCrawl, Swiss German Training Tweets
- Other languages - DSL, Ling10

DSL Dataset: Discriminating between Similar Language (DSL) contains 13 different languages based on 6 different language group. We used DSLCCv2.0 in our experiment.

Ling10 Dataset: It contains 190,000 sentences categorized into 10 languages (English, French, Portuguese, Chinese Mandarin, Russian, Hebrew, Polish, Japanese, Italian, Dutch).

Group Name	Language	Id
South Eastern Slavic	Bulgarian Macedonian	bg mk
South Western Slavic	Bosnian Croatian Serbian	bs hr sr
West-Slavic	Czech Slovak	cz sk
Ibero-Romance(Spanish)	Peninsular Spain Argentinian Spanish	es-ES es-AR
Ibero-Romance(Portuguese)	Brazilian Portuguese European Portuguese	pt-BR pt-PT
Astronesian	Indonesian Malay	id my

DSL Language Group. Similar languages with their language code.

Experimental Setup

- We made three settings (S1, S2, and S3) combing Swiss German and Other language texts.

Setting	Datasets and Language	Distribution	Distribution (Overall)	Training	Dev	Test
S1	NOAH (*Swiss-German*) SwissCrawl (*Swiss-German*) SwissTextTrain (*Swiss-German*) DSL (*not Swiss-German*) Ling10 (*not Swiss-German*)	7,327 (8%) 40,697 (40%) 1,976 (2 %) 25,000 (25 %) 25,000 (25 %)	50% Swiss-German 50% not Swiss-German	80,000	20,000	5,374
S2	NOAH (*Swiss-German*) SwissCrawl (*Swiss-German*) SwissTextTrain (*Swiss-German*) DSL (*not Swiss-German*) Ling10 (*not Swiss-German*)	7,327 (5%) 81,841 (55 %) 1,976 (1 %) 25,000 (17 %) 33,856 (22 %)	61% Swiss-German 39% not Swiss German	130,000	20,000	5,374
S3	NOAH (*Swiss-German*) SwissCrawl (*Swiss-German*) SwissTextTrain (*Swiss-German*) DSL (*not Swiss-German*) Ling10 (*not Swiss-German*)	7,327 (4 %) 81,841 (41 %) 1,976 (1 %) 50,000 (25 %) 58,856 (29 %)	46% Swiss-German 54% not Swiss-German	180,000	20,000	5,374

Dataset Statistics. The training-development-test set distribution for each of setting (S1, S2 and S3). The distribution is based on the number of sentences selected from the datasets.

Experimental Setup

- The SAE model configuration for training and search space hyperparameter range are shown in the table.

Hyperparameter	Range
number of layer	1-5
learning rate	10^{-5} - 10^{-2}
weight decay	10^{-6} - 10^{-3}
activation functions	‘relu’, ‘sigma’

Search space hyper parameter range.

Parameter	Value
char n-gram range	1-3
number of target	2
embedding dimension	300
supervision	‘clf’ (classification)
converge threshold	0.00001
number of epochs	500

SAE model configuration used for training

Evaluation

Setting	Prec (gsw)	Rec (gsw)	F1 (gsw)	Avg. Prec	AUROC
S1	0.649	0.997	0.786	0.871	0.924
S2	0.673	0.997	0.804	0.911	0.946
S3	**0.775**	**0.998**	**0.872**	**0.965**	**0.975**

Performance of setting S1, S2, and S3.

Evaluation (all submissions)

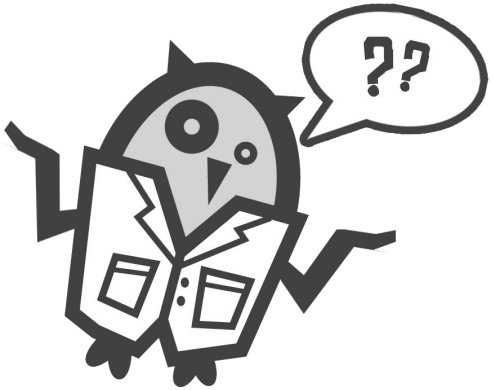
System	Precision	Recall	F1
Mohammad reza Banaei (MB)	**0.984**	0.979	**0.982**
jj-cl-uzh	0.945	0.993	0.968
IDIAP	0.775	**0.998**	0.872

Precision, Recall, and F1 scores for the positive class (GSW) of all submissions

Conclusion

- Supervised autoencoder with Bayesian Optimization for the language detection task found effectively for discriminating between very close languages or dialects.
- Future work involves:
  - Verify our model with other language detection datasets.
  - Explore other supervised autoencoder model with variational autoencoder for language detection task.

Q&A



Contact information:

- Shantipriya Parida:
  - Email: shantipriya.parida@idiap.ch
  - Twitter: @Shantipriyapar3

- Esaú Villatoro Tello:
  - Email: evillatoro@correo.cua.uam.mx / esau.villatoro@idiap.ch
  - Twitter: @EsauVT

Thank You