Multi-label Scandinavian Language Identification (SLIDE)

Mariia Fedorova*, Jonas Sebulon Frydenberg*, Victoria Handford*,
Victoria Ovedie Chruickshank Langø*, Solveig Helene Willoch,
Marthe Løken Midtgaard, Yves Scherrer, Petter Mæhlum, David Samuel
University of Oslo, Norway
{mariiaf,jonassf,vlhandfo,victocla,solvehw,marthemi,
yvessc,pettemae,davisamu}@ifi.uio.no

RESOURCEFUL-2025

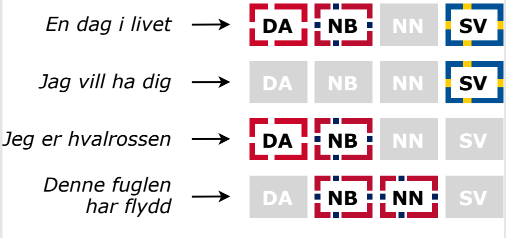
* Equal contribution.

Contents

Introduction

Why multi-label Scandinavian language identification?

Identifying closely related languages at sentence level is difficult:



Sentences valid in multiple Scandinavian languages are common: 5% of the test dataset and 16% of the sentences shorter than 6 words

Introduction

Why should we care about sentence-level LID?

- As modern language models are mostly pretrained on web crawls (Liu et al., 2019), (Touvron et al., 2023), texts of any length may occur in the pretraining data
- Code switching: a single text may contain sentences in different languages

Introduction

Why should we care about sentence-level LID?

- As modern language models are mostly pretrained on web crawls (Liu et al., 2019), (Touvron et al., 2023), texts of any length may occur in the pretraining data
- Code switching: a single text may contain sentences in different languages

Main contributions

- A multi-label evaluation dataset
- A suite of LID models
- A novel method of silver-labeling a LID dataset

Contents

Data

Why just another LID dataset?

most existing LID corpora rely on the source of a text: if a sentence is retrieved from a Danish newspaper, it is assumed to be only Danish. This approach doesn't work for similar languages (Goutte et al., 2016; Keleg and Magdy, 2023)

Data

Why just another LID dataset?

most existing LID corpora rely on the source of a text: if a sentence is retrieved from a Danish newspaper, it is assumed to be only Danish. This approach doesn't work for similar languages (Goutte et al., 2016; Keleg and Magdy, 2023)

Initial data sources

the Universal Dependencies 2.14 treebanks (Nivre et al., 2016, 2020) with their train/dev/test splits

Data

Manual inspection of dev and test splits

- sentences containing frequent words that unambiguously define a language (e.g. ‘ikkje’ is only valid in Nynorsk) were not subject to manual check

Data

Manual inspection of dev and test splits

- sentences containing frequent words that unambiguously define a language (e.g. ‘ikkje’ is only valid in Nynorsk) were not subject to manual check
- all annotators were native speakers who have received education in or about other Scandinavian languages

Data

Manual inspection of dev and test splits

- sentences containing frequent words that unambiguously define a language (e.g. ‘ikkje’ is only valid in Nynorsk) were not subject to manual check
- all annotators were native speakers who have received education in or about other Scandinavian languages

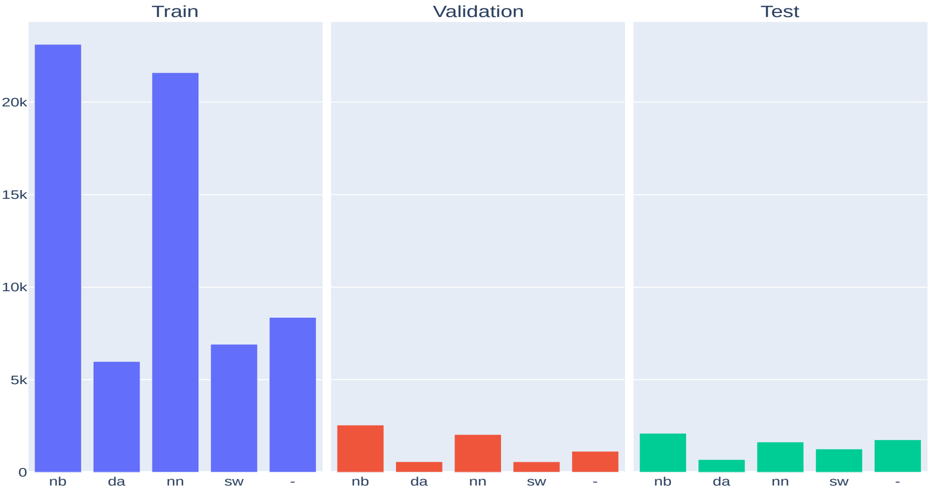
Machine translation silver-labeling of the train split

‘En dag i livet’

- Bokmål: En dag i livet
- Nynorsk: Ein dag i livet
- Danish: En dag i livet
- Swedish: En dag i livet

NorMistral-11b (Samuel et al., 2024), further fine-tuned on Tatoeba evaluation set (Tiedemann, 2020) in all translation directions between Bokmål, Danish, Nynorsk and Swedish

Data



Number of sentences per language (kept as in the original treebanks)

Contents

Evaluation

Loose accuracy

- a prediction is considered correct if intersection between predictions and gold labels is not empty

  a model that always predicts all four languages would get 100%

Strict accuracy

- exact match between the predicted and gold labels sets

  a model that always predicts all four languages would get as many % as much share all-four instances is in the data

Per-language F1-scores

- a true positive prediction happens if and only if the respective language is present both in the set of predicted labels and in the set of gold labels

Contents

Our approach: BERTs (SLIDE-xs, SLIDE-s, SLIDE-base)

Base model selection

Model	**Loose accuracy**	**Exact-match accuracy**	**Macro F_1**
XLM-RoBERTa-base (Conneau et al., 2020)	96.8	94.6	95.4
DistilBERT-base (Sanh et al., 2019)	96.5	94.5	95.2
ScandiBERT(Snæbjarnarson et al., 2023)	97.6	95.9	96.6
NorBERT3-base (Samuel et al., 2023)	**98.6**	**96.4**	**97.0**

Base model selection We made our choice based on the validation data split, the metrics in this table, given in percent, are for the test split. F_1 is per-language exact match. NorBERT3 refers to the same model as SLIDE.

Our approach: BERTs (SLIDE-xs, SLIDE-s, SLIDE-base)

Data augmentation and normalization

Basic idea: ensure that *'Oslo är Norges huvudstad'* will not be labeled as Norwegian

- Punctuation augmentation to prevent our models from relying too much on punctuation specific for a language

Our approach: BERTs (SLIDE-xs, SLIDE-s, SLIDE-base)

Data augmentation and normalization

Basic idea: ensure that *'Oslo är Norges huvudstad'* will not be labeled as Norwegian

- Punctuation augmentation to prevent our models from relying too much on punctuation specific for a language
- Regular expression normalization - normalize URLs, email addresses, and numbers into the following special symbols: $\langle$URL$\rangle$, $\langle$mail$\rangle$ and $\langle$num$\rangle$

Our approach: BERTs (SLIDE-xs, SLIDE-s, SLIDE-base)

Data augmentation and normalization

Basic idea: ensure that *'Oslo är Norges huvudstad'* will not be labeled as Norwegian

- Punctuation augmentation to prevent our models from relying too much on punctuation specific for a language
- Regular expression normalization - normalize URLs, email addresses, and numbers into the following special symbols: $\langle$URL$\rangle$, $\langle$mail$\rangle$ and $\langle$num$\rangle$
- Alphabet variations - adding Swedish sentences containing the Danish–Norwegian letters and Danish and Norwegian sentences containing the Swedish letters (e.g., in proper names and in the context of quotations)

Our approach: BERTs (SLIDE-xs, SLIDE-s, SLIDE-base)

Data augmentation and normalization

Basic idea: ensure that *'Oslo är Norges huvudstad'* will not be labeled as Norwegian

- Punctuation augmentation to prevent our models from relying too much on punctuation specific for a language
- Regular expression normalization - normalize URLs, email addresses, and numbers into the following special symbols: ⟨URL⟩, ⟨mail⟩ and ⟨num⟩
- Alphabet variations - adding Swedish sentences containing the Danish–Norwegian letters and Danish and Norwegian sentences containing the Swedish letters (e.g., in proper names and in the context of quotations)
- Named entity swaps - (NER) on the training data using spaCy; randomly swap the recognized entities with other entities from the same category

Our approach: BERTs (SLIDE-xs, SLIDE-s, SLIDE-base)

Data augmentation

Alterations	**Loose accuracy**	**Exact-match accuracy**
Augmentation + Regex normalization	98.6	**96.4**
Augmentation	98.4	96.3
Regex normalization	98.4	96.2
NER	**98.7**	95.5
Base	98.3	96.2

Ablation study Impact of data augmentation and regular expression normalization on SLIDE-base measured by test set performance. "Augmentation" refers to punctuation and alphabet augmentation, "Regex" refers to regular expression normalization, "NER" refers to named entity swaps and "Base" is neither of the above.

Contents

Results

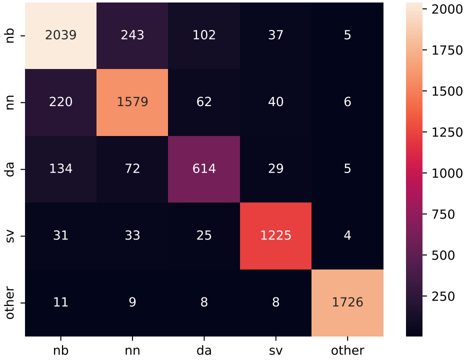
Results

Model	Loose accuracy	Exact-match accuracy	NB F_1	DA F_1	NN F_1	SV F_1	Other F_1	Runtime ms/sample
BASELINES								
gpt2-lang-ident	61.2	58.9	47.0	24.0	36.9	83.6	86.2	52.07
FastText-176*	80.5	77.7	72.6	66.0	55.7	92.7	93.5	**0.01**
NLLB-218*	95.3	91.6	93.0	85.9	89.0	96.8	93.6	0.08
NB-Nordic-LID*	83.3	80.6	85.0	67.0	84.8	89.7	70.2	0.02
OpenLID*	94.2	90.2	91.5	82.6	88.7	95.7	93.3	0.08
GlotLID*	97.2	93.4	93.5	89.5	89.4	97.9	98.1	0.51
Heliport (HeLI-OTS)	96.5	92.6	90.9	89.0	91.2	97.6	97.2	0.02
OUR MODELS								
SLIDE-x-small (15M)	97.8	95.7	97.5	90.4	96.2	98.0	98.7	13.22
SLIDE-small (40M)	98.1	95.7	97.7	89.9	96.3	98.0	99.1	19.70
SLIDE-base (123M)	**98.6**	**96.4**	**98.1**	**92.0**	**97.1**	**98.6**	**99.4**	38.41

* shows which baselines use FastText. Heliport is the only multilabel one

Contents

Discussion



SLIDE-base confusion matrix

Sources of Norwegian errors: spelling variations, ambiguity

- ‘høyre’ - Nynorsk (‘hear’), Bokmål (‘right’);
- ‘I alle år har vi fått høyre at med dagens forbruk er det olje nok for mange ti år.’ (Nynorsk) misclassified as Bokmål
- ‘I den nye designen er høgre og venstre spalte på framsida til nettavisa fjerna.’ (Bokmål, Nynorsk) misclassified as Nynorsk only

Discussion

Conclusion

- the dataset is released on https://github.com/ltgoslo/slide; the models will be made public on HuggingFace soon

Discussion

Conclusion

- the dataset is released on https://github.com/ltgoslo/slide; the models will be made public on HuggingFace soon
- using machine translation for creating a silver multi-label training dataset from a single-label one has proved to be efficient

Discussion

Conclusion

- the dataset is released on https://github.com/ltgoslo/slide; the models will be made public on HuggingFace soon
- using machine translation for creating a silver multi-label training dataset from a single-label one has proved to be efficient
- no clear answer how much data preprocessing/data augmentation makes the model most robust

Discussion

Conclusion

- the dataset is released on https://github.com/ltgoslo/slide; the models will be made public on HuggingFace soon
- using machine translation for creating a silver multi-label training dataset from a single-label one has proved to be efficient
- no clear answer how much data preprocessing/data augmentation makes the model most robust
- future work and the right way to solve the task: multilabel with machine translation GlotLID's dataset (3.9M samples for Norwegian Bokmål only); tokenize in a Scandinavian-friendly way; train FastText embeddings; train a multilabel classifier on top of it. An open question is how many 'other' data is needed

Acknowledgments

We would like to thank Helene Bøsei Olsen and Karoline Sætrum for their work on annotating the initial version of the test set.

Contents

Appendix

GlotLID is a strongest baseline and faster than a BERT: but...

- >2000 languages is a bottleneck for the runtime
- single-labeling
- solution: train an own MLP on top of GlotLID embeddings
- still a problem: OOV words; data normalization/augmentation did not help (probably also because of OOV)

Appendix

Results

Model	Loose accuracy	Exact-match accuracy	NB F_1	DA F_1	NN F_1	SV F_1	Other F_1	Runtime ms/sample
BASELINES								
GlotLID*	97.2	93.4	93.5	89.5	89.4	97.9	98.1	0.51
Heliport (HeLI-OTS)	96.5	92.6	90.9	89.0	91.2	97.6	97.2	0.02
OUR MODELS								
SLIDE-fast	95.7	93.4	94.5	90.2	92.4	97.5	96.4	0.16
SLIDE-x-small	97.8	95.7	97.5	90.4	96.2	98.0	98.7	13.22
SLIDE-small	98.1	95.7	97.7	89.9	96.3	98.0	99.1	19.70
SLIDE-base	**98.6**	**96.4**	**98.1**	**92.0**	**97.1**	**98.6**	**99.4**	38.41

Did we overfit to UD?

Model	3K test split	15K test split
SLIDE-base	92.7	95.3
SLIDE-fast	85.4	88.5
GlotLID	**93.0**	**95.7**

As (Haas and Derczynski, 2021) is a single-label dataset, we consider a prediction to be correct, if one of the predicted languages is correct.

Did we overfit to UD?

Model	3K test split	15K test split
SLIDE-base	92.7	95.3
SLIDE-fast	85.4	88.5
GlotLID	**93.0**	**95.7**

As (Haas and Derczynski, 2021) is a single-label dataset, we consider a prediction to be correct, if one of the predicted languages is correct.

- no ‘other’ languages, except of Icelandic and Faroese

Did we overfit to UD?

Model	3K test split	15K test split
SLIDE-base	92.7	95.3
SLIDE-fast	85.4	88.5
GlotLID	**93.0**	**95.7**

As (Haas and Derczynski, 2021) is a single-label dataset, we consider a prediction to be correct, if one of the predicted languages is correct.

- no 'other' languages, except of Icelandic and Faroese
- lower-cased and stripped out of numbers, punctuation signs and some accented character

Did we overfit to UD?

Model	3K test split	15K test split
SLIDE-base	92.7	95.3
SLIDE-fast	85.4	88.5
GlotLID	**93.0**	**95.7**

As (Haas and Derczynski, 2021) is a single-label dataset, we consider a prediction to be correct, if one of the predicted languages is correct.

- no ‘other’ languages, except of Icelandic and Faroese
- lower-cased and stripped out of numbers, punctuation signs and some accented character
- ‘ou di be t aatm ne enwadi’ - Swedish, ‘atahualpa yupanqui’ - Danish, ‘tromssan ruijan-suomalainen yhdistys’ - Nynorsk

Appendix

Confusion with ‘other’

- proper names (‘kruvi: *Karl Marx*’) (50% of ‘other’ instances misclassified as Scandinavian)
- English (30% of ‘other’ instances misclassified as Scandinavian)
- loanwords: server med pastasalat med bakte grønsaker og tsatsiki til
- ‘Va shiaulteyr *er* *ny* skeabey harrish boayrd.’ (Manx)

References I

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Goutte, C., Léger, S., Malmasi, S., and Zampieri, M. (2016). Discriminating similar languages: Evaluations and explorations. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1800–1807, Portorož, Slovenia. European Language Resources Association (ELRA).

References II

Haas, R. and Derczynski, L. (2021). Discriminating between similar Nordic languages. In Zampieri, M., Nakov, P., Ljubešić, N., Tiedemann, J., Scherrer, Y., and Jauhiainen, T., editors, *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 67–75, Kiyv, Ukraine. Association for Computational Linguistics.

Keleg, A. and Magdy, W. (2023). Arabic dialect identification under scrutiny: Limitations of single-label classification. In Sawaf, H., El-Beltagy, S., Zaghouani, W., Magdy, W., Abdelali, A., Tomeh, N., Abu Farha, I., Habash, N., Khalifa, S., Keleg, A., Haddad, H., Zitouni, I., Mrini, K., and Almatham, R., editors, *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

References III

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

References IV

Samuel, D., Kutuzov, A., Touileb, S., Velldal, E., Øvrelid, L., Rønningstad, E., Sigdel, E., and Palatkina, A. (2023). NorBench – a benchmark for Norwegian language models. In Alumäe, T. and Fishel, M., editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Samuel, D., Mikhailov, V., Velldal, E., Øvrelid, L., Charpentier, L. G. G., and Kutuzov, A. (2024). Small languages, big models: A study of continual training on languages of norway.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

References V

Snæbjarnarson, V., Simonsen, A., Glavaš, G., and Vulić, I. (2023). Transfer to a low-resource language via close relatives: The case study on Faroese. In Alumäe, T. and Fishel, M., editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Tiedemann, J. (2020). The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

References VI

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.