Lexicon-based data synthesis for Swiss German NLP

Barbara Kovačić
MaiNLP Research Lab
Ludwig-Maximilians-Universität München

Motivation

- **reasons to develop NLP tools for language variation**
  - language documentation and research
  - cultural preservation
  - more inclusive language technologies and applications
- **challenges**
  - no standard orthography
  - big regional differences
  - little to no data
- **possible solutions**
  - data synthesis techniques
  - transfer approaches from related languages

Data Synthesis

- **Definition:** techniques to increase the diversity of training data without collecting additional data[1]
- **Techniques:**
  - rule-based, e.g. EDA[2]
  - interpolation, e.g. MIXUP[3]
  - model-based, e.g. Backtranslation[4]

Transfer from High Resource Languages to Low Resource Languages[5]

- zero-shot learning
- annotation projection
- delexicalization
- relexicalization
- cross-lingual models

Differences between Standard and Swiss German[6][7]

E Aarm elai cha Bäärge verschiebe. Met allne viir Aarmi tailt s’Pokémon hammermässigi Schlääg uus.

One arm alone can move mountains. Using all four arms, this Pokémon fires off awesome punches.

Differences between Standard and Swiss German[6][7]

E Aarm elai cha Bäärge verschiebe. Met allne viir Aarmi tailt s'Pokémon hammermässigi Schlääg uus.

One arm alone can move mountains. Using all four arms, this Pokémon fires off awesome punches.

Me, writing my bachelor thesis about Swiss German NLP
MIT

Idea and Research Question

“How can data synthesis be effectively used to improve language models handling data including dialectal expressions?”

- enhance a Standard German dataset with Swiss German expressions by using a bilingual word list and inject Swiss German words in the dataset
- compare results of POS tagging a Swiss German test set with a language model, trained with a non-adapted Standard German dataset and a language model that has been trained with an enhanced Standard German dataset

Datasets

	Hamburg Dependency Treebank (HDT) [8]	**NOAH [7]**	**ArchiMob [9]**
Function	StG training dataset	Annotated gold standard test dataset for SwG	Bilingual word list
Language	Written	Written	Spoken
Source	StG sentences taken from technical news service "Heise"	SwG sentences, taken from a news paper, an annual report, novels, blogs and Alemannic Wikipedia	Transcriptions of oral history interviews in SwG
Content	StG word, POS tag, information about gender, number, etc.	SwG word, POS tag	SwG word, StG Version, POS tag
Dialects	n.A.	Aarau, Basel, Bern, Zurich and Eastern part of Switzerland	Zürich, Basel, Bern, Luzern

More Swiss German datasets:

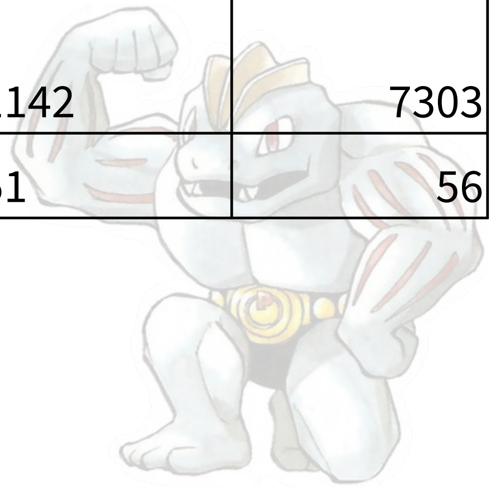
Blaschke et al (2023) [10]

NOAH

1. Extract data from XML files
2. Normalize Swiss-Specific STTS-Tags
    a. + Tags
    b. PTKINF tag
3. Restructure data in CoNLL file format according to HDT
4. Create one file with all genres and a file for each genre

NOAH

	NOAH-BLICK	NOAH-BLOGS	NOAH-SCHOBINGER	NOAH-SWATCH	NOAH-WIKI	NOAH-ALL
# of tokens	11256	34294	12855	33024	22136	113565
# of sentences	790	2937	1019	1415	1142	7303
# of tags	49	54	50	49	51	56



ArchiMob

1. Extract data from XML
2. Normalize "Schwyzerdütschi Dialektschrift"
3. Normalize POS tags
4. Remove duplicates
5. Align Swiss German variants to their Standard German equivalent

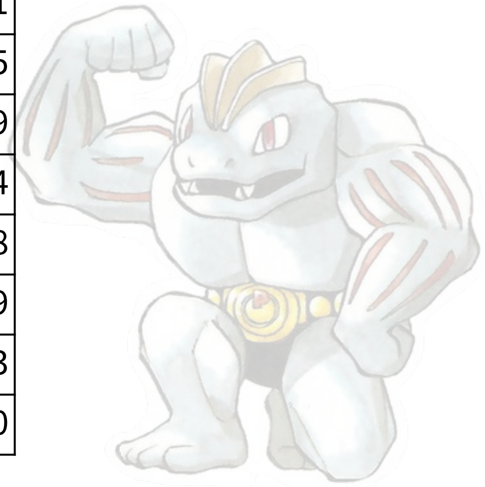
```
['Leute',
  'NN',
  [('Lüüt', 386),
   ('Lüt', 346),
   ('Liit', 22),
   ('Lit', 16),
   ('Leit', 4),
   ('Lüüte', 4),
   ('Lüüchte', 2),
   ('Lüte', 2),
   ('Lüütä', 1),
   ('Lüüter', 1),
   ('Liche', 1)]]
```

Result: bilingual word list with 41.013 StG words

A-HDT-ALL

- **Approach:** inject as many Swiss German words from the bilingual word list as possible into the HDT dataset
- **Result:** around 250.000 replacements spread over 20 different POS tags

Category	**Tag**	**# of injections**
Article	ART	115.761
Noun	NN, NE	84.795
Adposition	APPRART, APPR	13.739
Adverb	ADV	12.994
Pronoun	PRELS, PDAT, PPER, PIS, PDS, PPOSAT	11.798
Verb	VMFIN, VAFIN, VVFIN, VVINF, VVPP	7.019
Adjective	ADJA, ADJD	4.553
Conjunction	KOUS	4.270



A-HDT-ALL

Als	*Bischpill*	*erscheinen*	*überwiegend*
	Beispiel		
Web-Inhalte	*met*	*tüpische*	*Theeme*
	mit	typischen	Themen
fü	*de*	*elteri*	*Generazioon.*
für	die	älteren	Generationen.

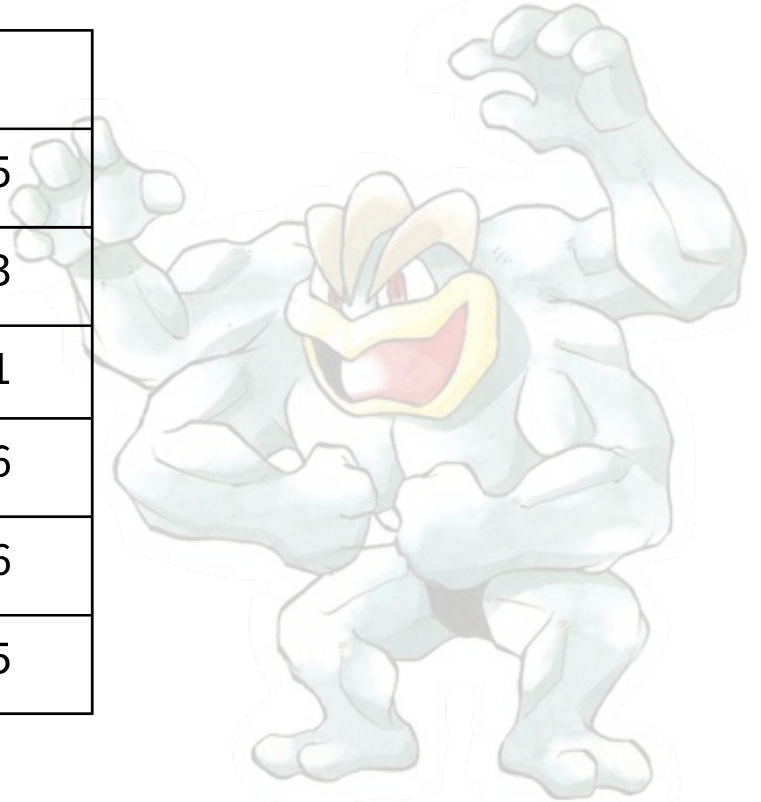
POS Tagger: MaChAmp[11]

“Massive Choice, Ample Tasks”

- toolkit based on multi-task learning
- allows multiple datasets and multi-task setups
- offers a wide range of NLP tasks
- supports initialization and fine-tuning of contextualized embeddings from Hugging Face
- **default:** mBERT
- **for sequence labeling:** greedy decoding approach using a softmax output layer on contextual embeddings

Performance per Genre

accuracy	**Baseline**	**A-HDT-ALL**	
NOAH-BLICK	0.67	0.82	0.15
NOAH-BLOGS	0.60	0.73	0.13
NOAH-SCHOBINGER	0.60	0.81	0.21
NOAH-SWATCH	0.67	0.83	0.16
NOAH-WIKI	0.68	0.83	0.16
NOAH-ALL	0.64	0.79	0.15



Outlook

- further data analysis
- try different combinations of replaced POS tags
- find a more suitable bilingual word list
- apply methods to include spelling variations
- test it on the complete HDT dataset
- use German “dbmdz” BERT
- …

Contact

Barbara Kovačić

Computational Linguistics and Phonetics student
Student Assistant at MaiNLP Lab
Ludwig-Maximilians-University, Munich

✉ Barbara.Kovacic@campus.lmu.de

@FrauKovacic

barbarakovacic

Sources

[1] Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. *arXiv preprint arXiv:2105.03075*.

[2] Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

[3] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. Proceedings of ICLR.

[4] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics. https://aclanthology.org/P16-1009/

Sources

[5] Zampieri, M., Nakov, P., & Scherrer, Y. (2020). Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, *26*(6), 595-612.

[6] Clyne, M. (1991). German as a pluricentric language. In M. Clyne (Ed.), *Pluricentric Languages: Differing Norms in Different Nations* (pp. 117-148). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110888140.117

[7] Hollenstein, N., & Aepli, N. (2014, August). Compilation of a Swiss German dialect corpus and its application to PoS tagging. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects* (pp. 85-94).

[8] Foth, K., Köhn, A., Beuck, N., & Menzel, W. (2014). Because size does matter: The hamburg dependency treebank.

[9] Scherrer, Y., Samardžić, T., & Glaser, E. (2019). ArchiMob: ein multidialektales Korpus schweizerdeutscher Spontansprache. *Linguistik online*, *98*(5), 425-454.

Sources

[10] Blaschke, V., Schuetze, H., & Plank, B. A Survey of Corpora for Germanic Low-Resource Languages and Dialects. In *The 24rd Nordic Conference on Computational Linguistics*.

[11] van der Goot, R., Üstün, A., Ramponi, A., Sharaf, I., & Plank, B. (2020). Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. *arXiv preprint arXiv:2005.14672*.