Universität Zürich[UZH]

Institute of Computational Linguistics

A Resource for Natural Language Processing of Swiss German Dialects

Nora Hollenstein, Noëmi Aepli
Institute of Computational Linguistics, University of Zurich

Introduction

- Expansion of *NOAH's Corpus of Swiss German Dialects* consisting of various text genres
- More than 115'000 manually annotated tokens with Part-of-Speech tags
- Ensuring annotation consistency (*variation n-gram* method)
- Dialect-specific PoS-tagging evaluation
- Prototype system for dialect identification (character-based trigram approach)

Swiss German

- Swiss German is a low-resourced language and belongs to the Alemannic group of dialects.
- Swiss German is a dialect continuum whose dialects are very different from Standard German.
- It is used in spoken language and in informal written texts (emails, blogs, text messages, etc.).

Differences to Standard German

- Vocabulary: different genus for the same word
  Standard German: *das Radio*
  Swiss German: *der Radio*
- Verb tenses: no preterite form in Swiss German
  Standard German: *Ich las ein Buch.*
  Swiss German: *Ich ha es buech gläse.*
- Use of auxiliary verbs:
  Standard German: *Mir ist kalt.*
  Swiss German: *Ich ha chalt.*
- Verb order is more flexible in Swiss German
  Standard German: *Sie lies ihn gehen.*
  Swiss German: *Sie hät ihn gah lah.*
- Merged words in Swiss German
  Standard German: *gehen wir*
  Swiss German: *gömmer*

Part-of-Speech Tagging

- Training of 6 statistical PoS-Taggers
- Best results achieved with *BTagger*
- *BTagger* makes use of context information and emphasises the transition probability by learning sequences of tags.
- 10-fold cross validation over the complete corpus
- Most frequent errors:
  - Confusion of nouns (NN) and proper names (NE)
  - Confusion of articles (ART) and personal pronouns (PPER)
- Accuracy: 90.62%

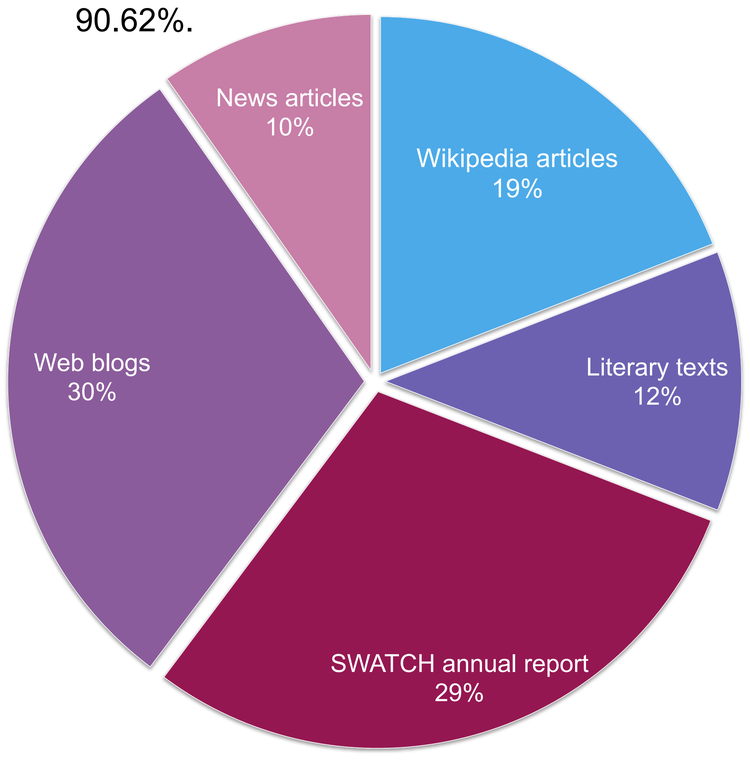
Annotation

- **Basic tagset**
  Stuttgart-Tübingen-Tagset (STTS), the standard for German.
- **Additional attributes**
  - Introduction of the tag *PTKINF* for infinitive particles:
    Standard German: *Ich gehe einkaufen.*
    Swiss German: *Ich go go (PTKINF) poschte.*
  - Adding of a "+"-sign to any PoS tag of a merged word:

PoS tag	Swiss German	Standard German	English
VAFIN+	*isches*	ist es	is it
KOUS+	*dasme*	dass man	that one
VMFIN+	*chame*	kann man	can one
PTKZU+	*zflügä*	zu fliegen	to fly
ADV+	*deetobe*	dort oben	up there

NOAH's Corpus of Swiss German Dialects

- **Download**: http://kitt.cl.uzh.ch/kitt/noah/
  Compilation of a corpus consisting of various text genres.
- Including dialects of most German-speaking regions of Switzerland.
- Manually annotated with Part-of-Speech tags.
- Training and evaluation of a statistical Part-of-Speech tagger, achieving an accuracy of 90.62%.



Corpus composition	Number of tokens	Tagging accuracy
Wikipedia articles	*22140*	*90.92%*
Literary texts (novels)	*13680*	*89.37%*
SWATCH annual report	*34048*	*88.82%*
Web blogs	*34839*	*88.10%*
Newspaper articles	*11271*	*87.17%*
Total	***115978***	***90.62%***

Dialect-specific POS-Tagging

Taking advantage of the fact that dialect information is available as metadata in our corpus, the PoS-tagger was trained for each dialect separately. We focused on the five dialects for which the largest amount of training data is available and evaluated these through a 10-fold cross-validation. Each model was trained on 4,000 tokens.

Dialect	Accuracy
Aarau	*85.73%*
Basel	*85.28%*
Bern	*87.85%*
Ostschweiz	*85.77%*
Zürich	*87.47%*

Dialect Identification

- **Goal**: building a dialect identification system for Swiss German texts.
- Implementation of a baseline system for five major dialects.
- **Development set**: 1470 sentences
- **Test set**: 250 sentences (50 per dialect)
- **Language model**: character-based trigram approach
- We trained a trigram language model for each dialect and scored each test sentence against every model.
- The predicted dialect was chosen based on the lowest perplexity.

Dialect	Precision	Recall	F-Score
Aarau	0.30	0.36	0.33
Basel	0.54	1.0	0.70
Bern	0.52	0.76	0.62
Ostschweiz	0.68	1.0	0.81
Zürich	0.74	1.0	0.85
Average	0.56	0.82	**0.66**

- **Future work**: more training data is required and taking into account the similarity of the dialects

Conclusion

- There is a need for more language processing tools for Swiss German.
- *NOAH's Corpus* is a basis for continuative research in Swiss German language processing.
- *NOAH's Corpus* is a foundation for downstream NLP applications such as dialect identification.