Machine Translation from Standard German to Alemannic Dialects

SIGUL 2022
Louisa Lambrecht, Felix Schneider, Alexander Waibel
Motivation

Why dialect translation?

- Many people speak dialect only
- Cultural heritage
Alemannic

Alemannic variants:
- Low Alemannic
- Upper Rhine Alemannic
- Lake Constance Alemannic
- High Alemannic
- Highest Alemannic
- Swabian

→ fluent transitions
Alemannic

Language characteristics:

- **K/ch-line**: Sundgau-Bodensee-Schranke (isogloss)
  - *Kind* vs. *Chind* (English: *child*)

- Vocabulary subset not shared with Standard German
  - *Kartoffel* vs. *Erdapfel, Grundbirne, Gummel* (English: *potato*)

- Grammar: passive voice and perfect tense
  - *er entdeckte* vs. *ist von ihm entdeckt worden* (English: *he discovered* vs. *has been discovered by him*)

- Grammar: avoiding genitive
  - *die Krone der Königin* vs. *die Krone von der Königin* (English: *the queen’s crown* vs. *the crown of the queen*)
Related Work

- Dialect translation mostly for Arabic
  - normalization vs. translation

- Swiss German normalisation

- Only one work translating into Alemannic dialects (Scherrer, 2012)

- Rule-based system with handwritten transformation rules
  - phonetic rules
  - word translation rules
  - syntactic rules

- Rules are georeferenced with probability maps
Data: Alemannic Wikipedia

- Parallel and monolingual corpus
- derived from the Alemannic Wikipedia
Data: Alemannic Wikipedia

- Parallel and monolingual corpus
- derived from the Alemannic Wikipedia

- 33,597 articles in the Alemannic Wikipedia (as at June 15, 2021)
- 5,462 articles are tagged with an Alemannic variant
- 29 tags
Data: Alemannic Wikipedia

- Parallel and monolingual corpus
- Derived from the Alemannic Wikipedia

- 33,597 articles in the Alemannic Wikipedia (as at June 15, 2021)
- 5,462 articles are tagged with an Alemannic variant
- 29 tags

→ use the dialect tags to split the corpora
Data: Challenges

- High diversity within the dialect
Data: Challenges

- High diversity within the dialect
  - e.g., use of accents
    Wia vielmols mit da Dialekta, düen d Üssproch un dr Wortschàtz vum Elsassischa schnall wachsla mit dr Geographie.

→ normalisation of data
Data: Challenges

- High diversity within the dialect
  - e.g., use of accents
    Wia vielmols mit da Dialekta, düen d Üssproch un dr Wortschàtz vum Elsassisch Schnall wachsla mit dr Geographie.
  - normalisation of data
- e.g., different spellings: hät, het, hot (English: has)
Data: splitting corpora
### Data: splitting corpora

#### Linguistic analysis

| Alemannic dialect         | gehört  |
|---------------------------|---------|
| Glarnerdeutsch            | gchört  |
| Walserdeutsch             | gchöört, ghöört |
| Alagnadeutsch             |         |
| Senslerdeutsch            | köört, ghöört, köre |
| Walliserdeutsch           | khöört, kheert |
| Issimedeutsch             | gheert  |
| Obwaldnerdeutsch          | gheerd  |
| Nidwaldnerdeutsch         | gheerid |
| Hochalemannisch           | gheert  |
| Schwyzterdeutsch          | ghöört  |
| Aargauerdeutsch           | ghööred, ghöört, ghört |
| Elsässisch                | ghert, ghere, gheera |
| Liechtensteinerisch       | ghörd, köört, ghört, khoört |
| Vorarlbergisch            | köört, ghört, khoört |
Data: splitting corpora

- Linguistic analysis
- Categorisation of Alemannic
Data: splitting corpora

- Linguistic analysis
- Categorisation of Alemannic
- Distribution of tagged articles

→ 8 Alemannic dialects
Data: splitting corpora

→ 8 Alemannic dialects

- Margravian (mg)
- Basel German (bd)
- Swabian (sw)
- High Alemannic (ha)
- Low Alemannic (na)
- Highest Alemannic (hoe)
- Alsatian (els)
- others (so)
Data: splitting corpora

- Train a classifier to identify the Alemannic dialect in which a Wikipedia article was written
- Divide articles into paragraphs of 6 sentences
- 22,277 data points
- 97.80% accuracy

→ apply to the parallel and monolingual corpus
→ filter monolingual corpus for the parallel
Data: splitting corpora

(a) tagged articles
(b) parallel corpus
(c) monolingual corpus

- mg
- bd
- sw
- ha
- na
- hoe
- els
- so
Data: back-translation

- Make use of the monolingual corpus
- Train
  - a back-translation model
  - and a German language model (on the German Wikipedia)
- BLEU: 55.31 with the LM weighted at 0.52
Data: back-translation

- Make use of the monolingual corpus
- Train
  - a back-translation model
  - and a German language model (on the German Wikipedia)
- BLEU: 55.31 with the LM weighted at 0.52

Data availability:

- Parallel corpus: 16,438 sentences
- Monolingual corpus: 390,561 sentences
Experiments
Baseline

- 10% test data that reflects the distribution of Alemannic variants
- Transformer architecture
  - default architecture but only 2 attention heads
  - high dropout rates: 0.3, 0.1 (attention), 0.3 (activation), 0.3 (decoder layers)
- BLEU: 37.29

|       | mg  | bd  | sw  | ha  | na  | hoe | els | so  | total |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| Baseline | 43.43 | 32.80 | 13.04 | 28.25 | 25.06 | 4.97 | 27.10 | 3.80 | **37.29** |
1. E2e training on back-translated data

- Use back-translated monolingual data
- Transformer model
- 390k sentences instead of 16k

→ mediocre results
1. **E2e training on back-translated data**

- Use back-translated monolingual data
- Transformer model
- 390k sentences instead of 16k
  → mediocre results
- Fine-tune on the parallel training data
  → strong improvements
2. E2e training for selected Alemannic variants

Alemannic dialects: Margravian, Basel German, Swabian
2. E2e training for selected Alemannic variants

- Alemannic dialects: Margravian, Basel German, Swabian

- Train a model for each Alemannic variant
- Train on back-translated monolingual data, fine-tune on parallel data
- Evaluation on the dialect-specific test set
2. E2e training for selected Alemannic variants

- Alemannic dialects: Margravian, Basel German, Swabian
- Train a model for each Alemannic variant
- Train on back-translated monolingual data, fine-tune on parallel data
- Evaluation on the dialect-specific test set
- Problem: Swabian test set contains only 15 sentences
3. Multi-dialectal translation

- Use all Alemannic variants except for Low Alemannic, Highest Alemannic and „others“
  → small data sets and heterogenous data
3. Multi-dialectal translation

- Use all Alemannic variants except for Low Alemannic, Highest Alemannic and „others“
  - small data sets and heterogenous data

- Train a multi-lingual transformer
  - one encoder for Standard German input
  - one decoder each for the Alemannic variants
3. Multi-dialectal translation

- Use all Alemannic variants except for Low Alemannic, Highest Alemannic and "others"
  - small data sets and heterogenous data

- Train a multi-lingual transformer
  - one encoder for Standard German input
  - one decoder *each* for the Alemannic variants

- Evaluation on the dialect-specific test sets
BLEU scores

|                        | mg  | bd  | sw  | ha  | na  | hoe | els | so  | total |
|------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| Baseline               | 43.43 | 32.80 | 13.04 | 28.25 | 25.06 | 4.97 | 27.10 | 3.80 | 37.29 |
| +back-translation      | 48.56 | 38.01 | 12.85 | 26.53 | 23.48 | 4.62 | 45.82 | 5.39 | 41.84 |
| separate dialects (mg)| 50.87 | 18.77 | 10.72 | 20.89 | 25.36 | 4.58 | 29.17 | 3.11 | 35.54 |
| separate dialects (bd)| 19.93 | 42.95 | 13.17 | 25.17 | 16.20 | 4.81 | 22.05 | 5.96 | 29.27 |
| separate dialects (sw)| 12.68 | 10.98 | 23.63 | 12.08 | 10.11 | 6.06 | 17.00 | 8.93 | 12.11 |
| multilingual (mg)      | 44.82 | 16.68 | 11.39 | 19.99 | 22.73 | 6.27 | 29.85 | 3.16 | 31.51 |
| multilingual (bd)      | 18.08 | 39.30 | 10.44 | 22.35 | 13.18 | **6.55** | 19.05 | 5.96 | 26.57 |
| multilingual (sw)      | 9.12  | 8.81  | **31.25** | 9.50  | 9.03  | 4.38 | 13.85 | 3.67 | 9.29  |
### BLEU scores

Splitting into Alemannic variants shows great improvements.

|                      | mg  | bd  | sw  | ha  | na  | hoe | els | so  | total |
|----------------------|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| Baseline             | 43.43 | 32.80 | 13.04 | 28.25 | 25.06 | 4.97 | 27.10 | 3.80 | 37.29 |
| +back-translation    | 48.56 | 38.01 | 12.85 | 26.53 | 23.48 | 4.62 | 45.82 | 5.39 | 41.84 |
| separate dialects (mg) | **50.87** | 18.77 | 10.72 | 20.89 | **25.36** | 4.58 | 29.17 | 3.11 | 35.54 |
| separate dialects (bd) | 19.93 | **42.95** | 13.17 | 25.17 | 16.20 | 4.81 | 22.05 | 5.96 | 29.27 |
| separate dialects (sw) | 12.68 | 10.98 | **23.63** | 12.08 | 10.11 | 6.06 | 17.00 | **8.93** | 12.11 |
| multilingual (mg)    | **44.82** | 16.68 | 11.39 | 19.99 | 22.73 | 6.27 | 29.85 | 3.16 | 31.51 |
| multilingual (bd)    | 18.08 | **39.30** | 10.44 | 22.35 | **13.18** | **6.55** | 19.05 | 5.96 | 26.57 |
| multilingual (sw)    | 9.12 | 8.81 | **31.25** | 9.50 | 9.03 | 4.38 | 13.85 | 3.67 | 9.29  |
## BLEU scores

- Splitting into Alemannic variants shows great improvements.
- E2e training using back-translated data works best for Alemannic as a whole.

|                                   | mg  | bd  | sw  | ha  | na  | hoe | els | so  | total |
|-----------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| Baseline                          | 43.43 | 32.80 | 13.04 | 28.25 | 25.06 | 4.97 | 27.10 | 3.80 | 37.29  |
| +back-translation                 | 48.56 | 38.01 | 12.85 | 26.53 | 23.48 | 4.62 | 45.82 | 5.39 | 41.84  |
| separate dialects (mg)            | 50.87 | 18.77 | 10.72 | 20.89 | 25.36 | 4.58 | 29.17 | 3.11 | 35.54  |
| separate dialects (bd)            | 19.93 | 42.95 | 13.17 | 25.17 | 16.20 | 4.81 | 22.05 | 5.96 | 29.27  |
| separate dialects (sw)            | 12.68 | 10.98 | 23.63 | 12.08 | 10.11 | 6.06 | 17.00 | 8.93 | 12.11  |
| multilingual (mg)                 | 44.82 | 16.68 | 11.39 | 19.99 | 22.73 | 6.27 | 29.85 | 3.16 | 31.51  |
| multilingual (bd)                 | 18.08 | 39.30 | 10.44 | 22.35 | 13.18 | 6.55 | 19.05 | 5.96 | 26.57  |
| multilingual (sw)                 | 9.12  | 8.81  | 31.25 | 9.50  | 9.03  | 4.38 | 13.85 | 3.67 | 9.29   |
Splitting into Alemannic variants shows great improvements.

E2e training using back-translated data works best for Alemannic as a whole.

Several test sets are too small to produce reliable results (sw, na, hoe, els, so).

### BLEU scores

|                                | mg  | bd  | sw  | ha  | na  | hoe | els | so  | total |
|--------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| **Baseline**                   | 43.43 | 32.80 | 13.04 | 28.25 | 25.06 | 4.97 | 27.10 | 3.80 | **37.29** |
| **+back-translation**          | 48.56 | 38.01 | 12.85 | 26.53 | 23.48 | 4.62 | 45.82 | 5.39 | **41.84** |
| separate dialects (mg)         | 50.87 | 18.77 | 10.72 | 20.89 | 25.36 | 4.58 | 29.17 | 3.11 | **35.54** |
| separate dialects (bd)         | 19.93 | 42.95 | 13.17 | 25.17 | 16.20 | 4.81 | 22.05 | 5.96 | **29.27** |
| separate dialects (sw)         | 12.68 | 10.98 | 23.63 | 12.08 | 10.11 | 6.06 | 17.00 | 8.93 | **12.11** |
| multilingual (mg)              | 44.82 | 16.68 | 11.39 | 19.99 | 22.73 | 6.27 | 29.85 | 3.16 | **31.51** |
| multilingual (bd)              | 18.08 | 39.30 | 10.44 | 22.35 | 13.18 | 6.55 | 19.05 | 5.96 | **26.57** |
| multilingual (sw)              | 9.12  | 8.81  | **31.25** | 9.50 | 9.03 | 4.38 | 13.85 | 3.67 | **9.29** |
Differentiating Alemannic variants

Target (mg):
S ältst bekannt Dokumänt, wo Aiche als Ort gnännt wird, chunnt uss em Johr 1275. Dört, wo hütt di ehemoligi Gmeind isch, hät mer scho 1160 e Kapälle dokumentiert.

Hypothesis (mg):
S eltscht bekannt Dokumänt, wu Aiche as Ort gnännt, stammt us em Johr 1275. E Kapelle im Biet vum hitige Ort isch scho 1160 dokumäntiert wore.

Hypothesis (sw):
S eldeshcde bekannde Dokument, wo Aiche als Ort zom erschte Mol gnennt, stammt us-em Johr 1275 em Gebiet vom heidiga Ort isch scho 1160 dokumentlicht worra.
Conclusion
Outlook

- Data preprocessing to remove spelling inconsistencies
  - e.g., Swiss-German dictionary of variations in speech and writing
  - e.g., word alignments and choosing the most frequent spelling

- Train the multi-dialectal model longer

- Use the German version of BERT/RoBERTa, i.e., GottBERT, for transfer learning

- Integrate speech data
Thank you!