Proceedings of SIGUL2022 @LREC2022, pages 129–136
Marseille, 24-25 June 2022

Machine Translation from Standard German to Alemannic Dialects

Louisa Lambrecht, Felix Schneider, Alexander Waibel
Interactive Systems Lab, Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
louisa.lambrecht@student.kit.edu, felix.schneider@partner.kit.edu,
waibel@kit.edu

Abstract

Machine translation has been researched using deep neural networks in recent years. These networks require lots of data to learn abstract representations of the input stored in continuous vectors. Dialect translation has become more important since the advent of social media. In particular, when dialect speakers and standard language speakers no longer understand each other, machine translation is of rising concern. Usually, dialect translation is a typical low-resourced language setting facing data scarcity problems. Additionally, spelling inconsistencies due to varying pronunciations and the lack of spelling rules complicate translation. This paper presents the best-performing approaches to handle these problems for Alemannic dialects. The results show that back-translation and conditioning on dialectal manifestations achieve the most remarkable enhancement over the baseline. Using back-translation, a significant gain of +4.5 over the strong transformer baseline of 37.3 BLEU points is accomplished. Differentiating between several Alemannic dialects instead of treating Alemannic as one dialect leads to substantial improvements: Multi-dialectal translation surpasses the baseline on the dialectal test sets. However, training individual models outperforms the multi-dialectal approach. There, improvements range from 7.5 to 10.6 BLEU points over the baseline depending on the dialect.

Keywords: machine translation, low-resource languages, dialect

1. Introduction

For almost a decade, neural networks have become an integral part of machine translation (MT) (Kalchbrenner and Blunsom, 2013). However, neural machine translation (NMT) struggles when only limited amounts of data are available. A typical low-resourced language setting is the translation of dialects. Though usually spoken, written dialect translation has gained more importance since the advent of social media in everyday life (Sajjad et al., 2020).

There are two main problems concerning dialect translation: firstly, data acquisition. Since dialects (even in written form) are primarily used in conversational settings, data is usually not publically available. Even less often is there actual parallel data. The second problem regards the language itself: dialects do not have uniform spelling rules. Many words have multiple spellings reflecting the varying pronunciations from region to region. That impairs the BLEU score (Papineni et al., 2002) checking for exact word matches. BLEU is the standard used metric to evaluate MT models. It is based on the amount of overlapping words and phrases (n-grams) between hypothesis and reference translation.

The Alemannic dialect is mostly spoken in Central Europe, i.e., southwestern Germany, German-speaking Switzerland, France (Alsace), Liechtenstein, and Austria (Vorarlberg). There are around 10 million people who speak Alemannic. The Alemannic language area can be divided into different regions. Figure 1 shows a map of the Alemannic language area and the Alemannic dialects spoken there.

Different language characteristics mark each region.

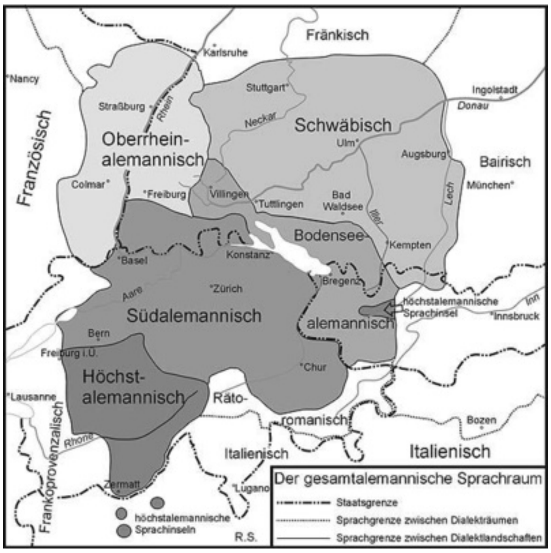


Figure 1: Alemannic language area in Central Europe (Schrambke, 2021)

Alemannic differs from Standard German in orthography, grammar and some vocabulary. For example, there are patterns in which orthography often changes (*st* → *scht* as in *Angst* → *Angscht* (fear) or prefix *ge* → *g* as in *gewöhnlich* → *gwöönlig* (usual, common)). Alemannic prefers perfect tense (more informal in Standard German) and passive voice over imperfect tense and active voice (Weinhold, 1863). Furthermore, the genitive is avoided in Alemannic and a small subset of the vocabulary is not derived from Standard German (e.g., *Grundbirne, Erdapfel, Häppere-Brägu,* *Häärpfel, Grompera, Gummel* all denote the potato - Standard German: *Kartoffel*) (Christen et al., 2013; Bühler, 2019).

This paper describes the most promising approaches using back-translation and a more fine-grained differentiation of dialects to handle Alemannic dialect translation and the problem of inconsistent orthography. Section 2 gives a short overview over related work concerning low-resourced MT, dialect translation in general and Alemannic (mostly Swiss German) dialect translation. In Section 3, the corpora, a dialect classifier, and the experiments are described. Section 5 presents the evaluation results as well as some examples. A more fine-grained differentiation between Alemannic dialects using the dialect classifier proved highly efficient in combination with back-translation. The dialects Margravian, Basel German, and Swabian were examined in more detail. The first two achieved their best results in separate models while the lowest-resourced dialect, Swabian, profited from a multilingual setting. Due to the limited size of the Swabian test set, this effect should not be overestimated, though.

2. Related Work

Methods for improving (low-resourced) NMT in general are byte pair encoding (BPE) (Sennrich et al., 2016b; Gage, 1994), transfer learning (Zoph et al., 2016), back-translation (Sennrich et al., 2016a), and multilingual MT (Dong et al., 2015; Luong et al., 2015; Ha et al., 2016; Johnson et al., 2017). Translating dialects has been a topic for several languages, e.g., Arabic (Baniata et al., 2018; Tachicart and Bouzoubaa, 2014; Salloum and Habash, 2013), Chinese (Wan et al., 2020; Huang et al., 2016), and Indian languages (Chakraborty et al., 2018). Most of the dialect translation research focuses on the translation into the standard language or vice-versa.

Concerning the Alemannic dialect, there are mainly works focusing on Swiss German rather than the full range of the Alemannic dialects. Most of them translate (or normalize) from Swiss German into Standard German. Many works applied rule-based approaches or statistical machine translation (Samardzic et al., 2015; Garner et al., 2014; Scherrer and Ljubešić, 2016). Two more recent works, that employ (at least partially) NMT are (Honnet et al., 2018) and (Arabskyy et al., 2021). Honnet et al. combine character-based neural machine translation with phrase-based statistical machine translation to translate from written Swiss German to Standard German. Arabskyy et al. propose a hybrid system that combines automatic speech recognition (ASR), a lexicon, an acoustic model, and a neural language model to recognizes Swiss German speech data and translate it to Standard German text.

The only work translating into Alemannic or Swiss German is a rule-based system that generates sentences in multiple Swiss German dialects using hand-written transformation rules (Scherrer, 2012). Most of these rules are georeferenced as they utilize probability maps to determine the dialectal differences. Scherrer also describes the challenge of evaluation: due to minimal changes in the dialectal orthography the exact word matching implemented in the BLEU metric often fails. This problem has also been detected for morphologically rich languages like Hindi, Finnish, and German (Chauhan et al., 2021; Niehues et al., 2016). Therefore, Scherrer utilizes the longest common subsequence ratio (LCSR) (Melamed, 1995) that calculates the proportion of identical letters between candidate and reference translations. However the score comparing hypothesis to reference was hardly different from the one comparing hypothesis to source text (83.30% vs. 82.77%).

3. Methodology

This section first describes the existing parallel corpus and the collection of a monolingual corpus from the Alemannic Wikipedia[1]. Secondly, the training of a dialect classifier is presented using additional dialect information extracted from the Wikipedia dump. This classifier was used to split the corpora into smaller dialectal corpora. Then, general preprocessing steps applied to both corpora are listed. In the end, the baseline used for comparison is described.

3.1. Data

The Alemannic Wikipedia is, like any language Wikipedia, an encyclopedia that relies on a community of volunteers who collaborate to write and maintain articles in Alemannic. Some of the Alemannic articles are direct translations of the Standard German correspondence. In 2019 prior to this work, Ann-Kathrin Habig sentence-aligned these articles manually with their Standard German equivalent. Thus, the parallel corpus of 16 438 sentences emerged.

Additional monolingual data was gathered in this work. As of June 15, 2021, the Alemannic Wikipedia consisted of 25 032 articles (and 8 564 forwarding articles coming along). The monolingual corpus was created from the entire Alemannic Wikipedia dump. Forwarding articles and short articles containing less than 50 words were filtered from the Wikipedia dump. The sentences present in the parallel and this monolingual data were deleted from the monolingual corpus to keep both corpora independent. Due to changes between 2019 and 2021 in the Alemannic Wikipedia, 10% of the parallel sentences could not be identified in the monolingual corpus. This was considered a reasonable amount to keep as the sentences had to have considerably changed that they were not recognized anymore. The monolingual corpus held 522 018 sentences by then.

[1] `https://als.wikipedia.org/`

dialect	#articles	parallel	mono
Markgräflerisch (mg)	852	8 253	128 825
Basel German (bd)	1 002	5 613	88 169
Swabian (sw)	873	128	23 683
High Alemannic (ha)	499	1 722	104 205
Low Alemannic (na)	145	243	6 952
Highest Alemannic (hoe)	56	43	5 615
Alsatian (els)	1 896	107	29 358
others* (so)	139	32	3 754
not classified	n/a	297	131 457
sum	5 462	16 438	522 018

Table 1: Number of tagged articles in the Alemannic Wikipedia and sentences per dialect in the two corpora. *: others consists of “Liechtensteinerisch” and “Vorarlbergisch”

3.2. Dialect Classifier

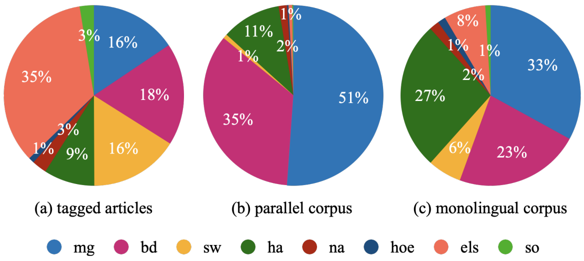
Authors submitting an article to the Alemannic Wikipedia have the option of tagging the article with their local dialect. 5 462 articles in the Wikipedia dump included dialect tags. 29 such dialect tags were extracted from the data. Some tags were present in only one or two articles, e.g., “Nidwaldnerdeutsch”, “Issimedeutsch”, others have several hundred associated articles, e.g., Swabian, Basel German, Alsatian. A rough linguistic analysis of the data based on frequently occurring words like *Einwohner* (inhabitant), *größte* (biggest, largest, greatest), *können* (can) and *haben* (have) conveyed similarities between the dialects. The dialect tags were grouped according to this linguistic analysis and the systematics of Alemannic dialects. The goal was to identify a rather rough clustering, i.e., few classes of dialects, but keeping the extent of inconsistencies within a dialect class minor. Furthermore, the classes should be balanced to prevent a bias to a certain dialect. Table 1 shows the identified classes (column 1), and the number of corresponding tagged articles (column 2).

Since most of the monolingual data did not have any dialect information, we trained a classifier with the extracted tagged data to identify the dialects of the remaining 19 570 articles in the monolingual corpus. The tagged articles were sliced into paragraphs of six sentences or at most 250 tokens to generate more data. These were added the corresponding label. This yielded 22 277 data points. The classifier was trained by fine-tuning the pre-trained RoBERTa (Liu et al., 2019) base model. Fine-tuning RoBERTa for a classification task was done according to the suggested design and hyperparameter choices[2] by Fairseq (Ott et al., 2019). After ten epochs of training, the classifier reached an accuracy of 97.80%. Table 2 shows the confusion matrix for the independent test set.

p* l*	mg	bd	sw	ha	na	hoe	els	so	sum
mg	370	1	0	3	0	2	1	0	377
bd	1	382	0	1	0	0	0	0	384
sw	2	0	454	0	0	0	1	1	458
ha	2	10	0	333	0	1	0	0	346
na	10	0	2	2	63	1	1	0	79
hoe	0	0	0	3	0	22	0	1	26
els	1	0	0	1	0	0	506	0	508
so	0	0	0	0	0	0	1	49	50
sum	386	393	456	343	63	26	510	51	2228

Table 2: Confusion matrix of the dialect classifier. *: l=label, p=prediction

The untagged data was classified by slicing the articles into paragraphs as well. These were classified, and only if there was a majority on the labels of the paragraphs, the article received this label. The other articles remained unclassified and were removed from the monolingual corpus. Table 1 also shows the statistics for the corpora after classification (column 3 and 4). In the end, the monolingual corpus held 390 561 classified sentences. The distribution of dialects in the corpora and of the original Wikipedia dialect tags differs greatly as Figure 2 illustrates.



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

Figure 2: Distribution of dialects in (a) tagged articls, (b) the parallel corpus, and (c) the monolingual corpus

3.3. Preprocessing

Both corpora were split in training, validation and test data. Due to the limited size of the corpora only 10% was used as test data. The remaining 90% were also split 90:10 between training and validation data. All sets represent the dialectal classes in size according to their distribution over the entire corpus. That leads to small test sets ($<$ 25 sentences) in the dialects that are underrepresented in the parallel corpus, i.e., Swabian, Low Alemannic, Highest Alemannic, and Alsatian

As preprocessing, the data was normalized (accent removal), tokenized (sacremoses), and byte pair encoding was applied (subword-nmt). The byte pair encoding was learned on the German and Alemannic parallel training sets limited to 8 000 BPE codes producing a joint dictionary of 8 340 subwords. These codes were applied to the train/validation/test sets and used in the baseline and the further experiments.

[2] https://github.com/pytorch/fairseq/blob/main/examples/roberta/README.custom_classification.md

3.4. Baseline

As a baseline, a transformer model (Vaswani et al., 2017) was trained on the parallel corpus. Embedding dimensions for the baseline and in the other experiments were chosen as proposed by the authors. Merely the number of layers and attention heads was reduced to 4 and 2/4 in some experiments. All trained models were set higher dropout rates as suggested by Araabi and Monz (2020).

4. Experiments

This section presents three experiments to overcome the challenges of data scarcity and inconsistent orthography in the Alemannic dialects. The first experiment adds the monolingual corpus by using back-translation. Both other experiments are based on the classified split corpora training separate models for three chosen dialects first and secondly combining several dialects in a multilingual model.

4.1. Back-translation

The model that was used to translate the Alemannic monolingual corpus into Standard German was trained on the parallel training data and combined with a Standard German language model (LM). This LM was trained on the German Wikipedia and weighted at 0.52. Together the models reached a BLEU score of 55.3 producing acceptable translations.

The parallel corpus's test set is used to assess the model's performance despite the size discrepancy (351.5k training vs. 1 644 test sentences). That ensures correct measurement of translation quality despite the imperfect synthetic data and enables comparability with the baseline.

Since the amount of synthetic data is significantly higher than the number of sentences in the parallel corpus (16.4k vs. 390.6k sentences), the learning opportunities are increased. On the other hand, the quality of this data is certainly lower than that of the parallel corpus.

The back-translated monolingual corpus was split into 10% validation and 90% training data. A transformer model was trained on this data first. Afterwards, the model was fine-tuned on the parallel corpus. Note that the distribution of dialectal classes in the monolingual corpus differs from that of the parallel corpus.

4.2. Individual Models for the Dialects

In order to reduce the spelling possibilities based on the clustering of Alemannic dialects, three end-to-end transformer models were trained for the dialects Margravian ("Markgräflerisch"), Basel German ("Baseldeutsch"), and Swabian ("Schwäbisch"). Margravian was selected since it has the most extensive dialectal corpus. Basel German with its slightly smaller corpus is at the border between High and Low Alemannic and, therefore, interesting as it might still hold many ambiguities. Swabian was chosen due to its unique position among the dialectal variants. Its spelling differs more clearly from the other Alemannic variants. All three dialectal variants have in common that they have their own tag in the Alemannic Wikipedia, which might be an advantage considering the number of inconsistent spellings.

The end-to-end models for the three Alemannic dialects were trained with the same transformer architecture. Dropout rates were slightly increased compared to the baseline. The trainings were stopped early to prevent overfitting. Afterwards, the models were fine-tuned on their respective dialectal parallel training data.

4.3. Multi-dialectal Model

As mentioned in Section 2, many low-resourced language settings profit from integrating other (closely related) languages into a multilingual setting. In theory, shared embeddings and hidden representations soften the data sparsity problem and enable zero-shot translations (Zoph et al., 2016; Artetxe and Schwenk, 2019). Therefore, a multi-dialectal translation model was trained with five of the eight Alemannic dialects (mg, bd, sw, ha, els). The other dialects were not included due to their small corpus size and heterogeneous nature found in the linguistic analysis.

The multilingual transformer was trained to translate from German into the specified dialects. One encoder was used to encode Standard German input and one decoder each for decoding the Alemannic variants. The embeddings were not shared across the dialects. The multilingual transformer training was terminated after 103 epochs. Fine-tuning was performed for ten epochs. We also trained models with shared embeddings and shared decoders. However, these setups did not yield as good results as using one decoder for each output dialect.

5. Evaluation

The evaluation was done with sacrebleu[3] (Post, 2018) after generating translations with Fairseq's generation tool that also takes care of BPE removal and detokenization. All translations are generated with the parameters beam=5 (default) and no-repeat-ngram-size=3.

The results of the baseline and the experiments are listed in Table 3. The table shows the BLEU scores on the entire parallel test set (column *total*) and additionally the scores for the dialectal test sets. The dialectal test sets are subsets of the parallel test set and hold the test sentences of the respective dialect, i.e., column *mg* shows the BLEU scores for the Margravian test sentences that are part of the entire test set (*total*).

The baseline and the model incorporating back-translated monolingual data should be evaluated on the entire parallel test set (column *total*). In contrast, the

[3] sacrebleu configuration: `BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.14`

	mg	bd	sw	ha	na	hoe	els	so	total
baseline	**43.4**	**32.8**	**13.0**	28.3	25.1	5.0	27.1	3.8	**37.3**
with back-translation	48.6	38.0	12.9	26.5	23.5	4.6	45.8	5.4	**41.8**
separate dialect (mg)	**50.9**	18.8	10.7	20.9	25.4	4.6	29.2	3.1	35.5
separate dialect (bd)	19.9	**43.0**	13.2	25.2	16.2	4.8	22.1	6.0	29.3
separate dialect (sw)	12.7	11.0	**23.6**	12.1	10.1	6.1	17.0	8.9	12.1
multilingual (mg)	**44.8**	16.7	11.4	20.0	22.7	6.3	29.9	3.2	31.5
multilingual (bd)	18.1	**39.3**	10.4	22.4	13.2	6.6	19.1	6.0	26.6
multilingual (sw)	9.1	8.8	**31.3**	9.5	9.0	4.4	13.9	3.7	9.3

Table 3: BLEU scores of the different experiments: relevant test sets for comparison in bold, best results underlined.

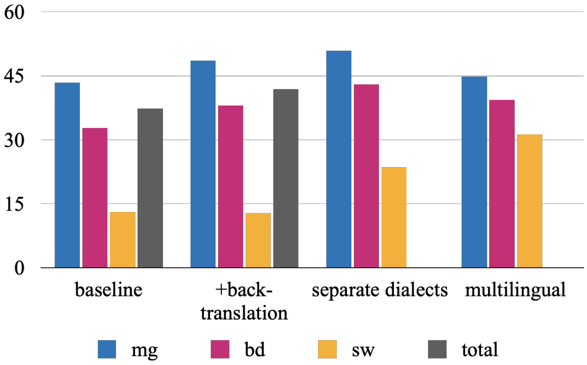


Figure 3: BLEU scores of the different experiments in total and on the relevant dialectal test sets

dialectal models and the multilingual model should not be evaluated on the whole test set as they are designed for a specific dialect. Therefore, the results of the corresponding relevant test sets are highlighted in bold font in Table 3. The scores on the other dialectal subsets were included for comparison. In addition, that might disclose some correlations among the dialects.

The baseline trained only on the parallel data achieves a BLEU score of 37.3 on the independent test set. Naturally, the dialectal variants with higher data proportions (mg, bd) perform better than the others.

5.1. Results

The model incorporating back-translated monolingual data reaches a BLEU score of 41.8 after fine-tuning. It shows an increase of performance in comparison to the baseline in the dominant dialects but decreases in most of the other dialects. Alsatian is a strong outlier. However, a large number of the Alsatian articles seem to focus on municipalities in Alsace. These articles are so similar to each other that they could be generated automatically. This would certainly create a strong bias within the Alsatian dialect.

Differentiating more fine-grained between Alemannic dialects showed improvements in both respective experiments in all three examined dialects. In the dialects Margravian and Basel German, the separate dialect models dominated. According to the corpus size, the model for Margravian achieved its best result after 300 epochs of training, the Basel German model trained 236 epochs, and the Swabian 162 epochs. For Margravian the BLEU score is improved by 7.5 points to 50.9 while the Basel German model surpasses the baseline by 10.2 BLEU points on the respective dialectal test set. The multilingual model also improves upon the baseline. The best model was reached after five epochs of fine-tuning. Its results show that mainly the lowest-resourced language, Swabian, benefits from the multilingual setting. Translating into Swabian the multilingual model surpasses the baseline by 18.3 BLEU points. Figure 3 summarizes the results of the experiments for the considered dialects.

5.2. Example

Table 4 lists the hypotheses of the different experiments for a sentence in Margravian. As the baseline's BLEU scores are very high from the beginning, translation quality is high in all hypotheses and differences between the experiments are minor. The hypotheses specific to Margravian agree in orthography for a great part. The baseline and the model using back-translated data are influenced by other dialectal orthography and their hypotheses show more differences. The translations of Standard German *Dokument* (*Dokumänt*; document) and *Jahr* (*Johr*; year) show how spelling is altered in Alemannic to match pronunciation. *aus* (out) is an example of one pronunciation having multiple spellings (*us, uss*) in the same Alemannic dialect (compare target and Margravian hypotheses) while *älteste* (oldest) has multiple spellings and pronunciations, e.g., *ältst* (in the target) and *eltscht* (in the dialectal hypotheses). Finally, there are some changes in the choice of words in Alemannic, e.g. *genannt* (*gnännt, gnennt*; call) instead of *erwähnt* (*erwäänt*; mention), *kommt* (*chunnt*; come) instead of *stammt* (date back). These lexical differences and paraphrasing proved most difficult for all models as most of the data contains only simpler reorderings due to changes in tense and case.

In contrast, Table 5 shows the hypotheses for the same sentence in the different dialects. The Margravian and Basel German hypotheses were produced with the individual dialectal models while the Swabian hypothesis was produced with the multilingual model. This table demonstrates the orthographic differences between the

Alemannic dialects. For example, *älteste* (oldest) had multiple spellings and pronunciations in Margravian alone. However, the baseline and the model with the back-translated monolingual corpus were trained with the full range of Alemannic and produce other valid translations and pronunciations of *älteste*. The dialects Basel German and Swabian add even more, e.g., the characteristic Swiss German *i* in the end of adjectives is preferred by the model with the back-translated monolingual corpus (Table 4) and the Swabian model (Table 5) produces "softer" pronunciations by choosing *d* over *t* as in *eldeschde* (and also *bekannde* (known)).

6. Discussion

Assessing translation quality using BLEU scores has become the predominant method. Compared to human evaluation it is less costly and less subjective. However, BLEU as an evaluation method has its drawbacks when it comes to morphologically rich languages. The high range of spelling possibilities can be viewed in the same way: there are several correct ways of expressing (or spelling) certain content. Usually, no more than one reference translation is available. That can diminish the BLEU scores for such languages. The examples shown in Table 4 demonstrate that translation quality is high concerning grammar, legibility, and correctness. However, concerning the separate dialect model's hypothesis and the target, five unigrams are incorrect - three of them differ in just one letter (*wu/wo, as/als, us/uss*). That can have tremendous effects on the BLEU score, and human evaluation might be an adequate alternative in this setting.

Nevertheless, some of the reported BLEU scores are relatively high. Note that the Alemannic dialects and Standard German are highly related. In contrast to spoken Alemannic, most written Alemannic texts (except Highest Alemannic) are intelligible for Standard German speakers without dialect background. BLEU scores reported in related work translating from Alemannic/Swiss German into Standard German are at a similar level. They range from 36 (Honnet et al., 2018) to 46 (Arabskyy et al., 2021), and 75 BLEU points (Garner et al., 2014).

The gain of 4.5 BLEU points by using back-translation is in the expected range. Splitting the data into smaller dialectal groups lead to respectable improvements. It was surprising that the multilingual model could not reach up to the individual dialect models (concerning Margravian and Basel German). Perhaps the multilingual model could benefit from other Germanic languages with larger corpora or transfer learning on the encoder side.

The BLEU scores found for the other dialects (apart from Margravian, Basel German, and Swabian) show some interesting correlations: All models perform considerably worse on the Highest Alemannic dialects and the data grouped in "others" than the other dialectal test sets. This supports the subjective impression that these dialects differ greatly from the other Alemannic data and endorses the decision of excluding this data from the multilingual setting. Similarly, the BLEU scores emphasize the differences to Swabian. Swabian does not only receive low scores in the baseline/with back-translation models but the Swabian models also perform very poor on the other dialectal test sets. However, the Swabian data was limited. That might inhibit Swabian models from performing well in general. Thus, the tremendous improvement by the multilingual model on the Swabian data (+18.3 BLEU points) also has to be interpreted with care as the Swabian test set contains less than 20 sentences.

Model/Language	Example
English	The oldest known document that mentions Aichen as a village dates back to 1275.
Standard German	Das älteste bekannte Dokument, das Aichen als Ort erwähnt, stammt aus dem Jahre 1275.
Alemannic Target (mg)	S ältst bekannt Dokumänt, wo Aiche als Ort gnännt wird, chunnt uss em Johr 1275.
Baseline	S älteschte Dokumänt, wo Aiche als Ort erwäänt, stammt us em Johr 1275.
with back-translation	S ältischti bekannti Dokument, wo Aiche als Ort erwähnt, stammt us em Johr 1275.
separate dialect (mg)	S eltscht bekannt Dokumänt, wu Aiche as Ort gnännt, stammt us em Johr 1275.
multilingual (mg)	S eltscht bekannt Dokumänt, s Aiche as Ort gnännt, stammt us em Johr 1275.

Table 4: Example of a Margravian (mg) sentence translated by the models of the different experiments

Model/Language	Example
English	The oldest known document that mentions Aichen as a village dates back to 1275.
Standard German	Das älteste bekannte Dokument, das Aichen als Ort erwähnt, stammt aus dem Jahre 1275.
Alemannic Target (mg)	S ältst bekannt Dokumänt, wo Aiche als Ort gnännt wird, chunnt uss em Johr 1275.
Margravian	S eltscht bekannt Dokumänt, wu Aiche as Ort gnännt, stammt us em Johr 1275.
Basel German	S eltiste bekannte Dokumänt, wo Aiche as Ort erwäänt, stammt us em Joor 1275.
Swabian	S eldeschde bekannde Dokument, wo Aiche als Ort zom erschte Mol gnennt, stammt us-em Johr 1275.

Table 5: Example of a Margravian (mg) sentence translated into different dialects

7. Conclusion

This work presents several experiments to improve machine translation of low-resourced languages on the example of the Alemannic dialects. Dialect translation has two primary problems: few parallel resources are available, and the colloquial nature of dialects often leads to inconsistent orthography. Using back-translation the parallel corpus of approximately 16k sentences could be expanded with a monolingual corpus holding 390k sentences. Tackling the problem of spelling inconsistencies does not have a definite course of action. Splitting the data into dialect groups and thus splitting the problem over several "languages" was rewarding. There are still spelling inconsistencies within these dialect groups, but the number certainly decreases. Individual models were trained for three Alemannic dialects on the corresponding subsets of the Alemannic monolingual data. Fine-tuning was performed with the analogous subset of the parallel corpus. BLEU scores on the dialectal test sets outperform the baseline by 7-10 BLEU points. A multi-dialectal model was trained on five Alemannic dialects. Its BLEU scores outperform the baseline on the dialectal test sets, but mainly the lowest-resourced dialect profited from the multilingual setting. The models trained for the separate Alemannic dialects achieved the best results. They produce high quality translations that account for the diversity of the Alemannic dialect by differentiating between Alemannic variants. Thus, the results propose a solid approach to deal with the problems of inconsistent orthography in dialects.

8. Acknowledgments

The compilation of the parallel corpus was carried out by research assistant Ann-Kathrin Habig in 2019.

9. References

Araabi, A. and Monz, C. (2020). Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Arabskyy, Y., Agarwal, A., Dey, S., and Koller, O. (2021). Dialectal speech recognition and translation of swiss german speech to standard german text: Microsoft's submission to swisstext 2021. Arxiv.

Artetxe, M. and Schwenk, H. (2019). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Baniata, L. H., Park, S., and Park, S.-B. (2018). A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational intelligence and neuroscience*, 2018.

Bühler, R. (2019). Sprachalltag ii: Sprachatlas–digitalisierung–nachhaltigkeit und das arno-ruoff-archiv am ludwig-uhland-institut für empirische kulturwissenschaft der universität tübingen. *Linguistik online*, 98(5):411–423.

Chakraborty, S., Sinha, A., and Nath, S. (2018). A bengali-sylheti rule-based dialect translation system: Proposal and preliminary system. In *Proceedings of the International Conference on Computing and Communication Systems*, pages 451–460. Springer.

Chauhan, S., Daniel, P., Mishra, A., and Kumar, A. (2021). Adableu: A modified bleu score for morphologically rich languages. *IETE Journal of Research*, 0(0):1–12.

Christen, H., Glaser, E., and Friedli, M. (2013). *Kleiner Sprachatlas der deutschen Schweiz*. Huber.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732.

Gage, P. (1994). A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Garner, P. N., Imseng, D., and Meyer, T. (2014). Automatic speech recognition and translation of a Swiss German dialect: Walliserdeutsch. In *Proc. Interspeech 2014*, pages 2118–2122.

Ha, T.-L., Niehues, J., and Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Honnet, P.-E., Popescu-Belis, A., Musat, C., and Baeriswyl, M. (2018). Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Huang, G., Gorin, A., Gauvain, J.-L., and Lamel, L. (2016). Machine translation based data augmentation for cantonese keyword spotting. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6020–6024. IEEE.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709.

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*.

Luong, M.-T., Le, Q., Sutskever, I., Vinyals, O., and Kaiser, L. (2015). Multi-task sequence to sequence learning. *Proceedings of ICLR, San Juan, Puerto Rico*, 11.

Melamed, I. D. (1995). Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. *arXiv preprint cmp-lg/9505044*.

Niehues, J., Peter, J.-T., Guillou, L., Huck, M., Sennrich, R., Bojar, O., Kocmi, T., Burlot, F., Skadina, I., and Deksne, D. (2016). Intermediate report: Morphologically rich languages.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Sajjad, H., Abdelali, A., Durrani, N., and Dalvi, F. (2020). Arabench: Benchmarking dialectal arabic-english machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107.

Salloum, W. and Habash, N. (2013). Dialectal arabic to english machine translation: Pivoting through modern standard arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358.

Samardzic, T., Scherrer, Y., and Glaser, E. (2015). Normalising orthographic and dialectal variants for the automatic processing of swiss german. In *Proceedings of the 7th Language and Technology Conference*.

Scherrer, Y. and Ljubešić, N. (2016). Automatic normalisation of the swiss german archimob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*.

Scherrer, Y. (2012). *Generating Swiss German sentences from Standard German: a multi-dialectal approach*. Ph.D. thesis, University of Geneva.

Schrambke, R. (2021). Die gliederung des alemannischen sprachraums. [Online; accessed August 23, 2021].

Sennrich, R., Haddow, B., and Birch, A. (2016a). 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.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Tachicart, R. and Bouzoubaa, K. (2014). A hybrid approach to translate moroccan arabic dialect. In *2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14)*, pages 1–5. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wan, Y., Yang, B., Wong, D. F., Chao, L. S., Du, H., and Ao, B. C. (2020). Unsupervised neural dialect translation with commonality and diversity modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9130–9137.

Weinhold, K. (1863). *Alemannische Grammatik*. Ferd. Dümmler's Verlagsbuchhandlung Harrwitz und Gossmann.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.