University of Hamburg

Masters Thesis

Neural Machine Translation for Dialects of Low-Resource Languages by Incorporating Linguistic Information to Create Synthetic Data

Author: Christian Schuler (6449321)

First Supervisor: Dr. Sina Ahmadi
Second Supervisor: Dr. Seid Muhie Yimam
Examiner: Prof. Dr. Chris Biemann

A thesis submitted in fulfillment of the requirements for the Master of Science Informatik

of the

Language Technology
Faculty of Informatics
Fachbereich Informatik

June 21, 2024

UNIVERSITY OF HAMBURG

Abstract

Faculty of Informatics
Fachbereich Informatik

Master of Science Informatik

Neural Machine Translation for Dialects of Low-Resource Languages by Incorporating Linguistic Information to Create Synthetic Data

by Christian Schuler

This work attempts to advance neural machine translation of dialects from low-resource languages in a meaningful way by utilizing synthetic text data, created via incorporation of linguistic information.

Natural language processing, Neural Machine Translation, Low-resourced language, Dialectal varieties, Linguistic information

Contents

List of Figures

List of Tables

List of Abbreviations

Chapter 1

Evaluation

1.1 Variables and Parameter

1.1.1 Language Variety

The degree of data scarcity vastly differs between language varieties (refer to recent alam paper) which in turn limits the quality and amount of data and linguistic rules that can be derived from it. The choice of language variety can be expected to have the greatest impact on the performance of the entire pipeline / this approach in it's entirety.

1.1.2 Data Quality

Naive denotes data that has been collected via means such as opustools which encompasses a range of different text corpora of varying degrees of quality. This data contains a lot of noise. Sentences that a ill-aligned, sentences that are of low quality, and even text from very different languages. (Show Examples: Corpora in which the aligned sentence is the same in both languages AND sometimes Chinese or Bengali characters labelled to be German or Bavarian text...)

Clean is the data that has gone through a rudimentary round of preprocessing such as detecting the language based on the script in which the characters are written and estimating the validity of sentence alignments by comparing their length (reasoning that a sentence more than x times the length of another sentence, can hardly be considered to be -well-aligned-.

Informed is all those data that has been labelled and evaluated by human native speaker of the corresponding language (for aligned text: translators)

1.1.3 Feature Validity

Guess is similar to the above -naive- as it is based on automatic functions and a data-driven approach which can be applied without access to native speakers, experts, or linguistic literature to draw from. As the name indicates, these rules are very basic and might be considered close to guessing the correct replacement of a word or sub-word unit.

Reason is an improved version of the rules from above in which multiple quality assuring measures are taken. Such as preventing the replacement of single characters with an empty string (without taking the context into account) which results in entire texts missing a set of characters.

Authentic is an approach of using replacement rules that have been derived from descriptions in scientific literature by expert linguists.

1.1.4 Perturbation Type

Lex denotes lexicographic replacements of entire words based on bilingual word lists.

Mor denotes morphological replacements of sub-word units based on rules derived by processing bilingual word lists.

All denotes the combination of both previous replacements by applying morphological ones after the lexicographic ones.

1.2 German and Bavarian

TABLE 1.1: Evaluation Metrics for Bavarian against the German reference

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	none	none	PERT	16.745	28.9858	97.9605
naive	none	none	NLLB	4.4749	14.3248	119.8208

Table 1.1 shows the differences between the Standard German and the Bavarian variant in terms of word-level (PERT) and in terms of translation model performance (or robustness to the dialect) by using the results of the German-to-English translations as reference (NLLB).

TABLE 1.2: Evaluation Metrics for Bavarian-German (standardized as part of preprocessing)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	PERT	13.787	38.4879	75.925
naive	guess	mor	PERT	6.7303	19.5736	85.1412
naive	guess	all	PERT	6.6107	17.5	86.3544
naive	guess	lex	NLLB	4.322	19.6195	127.2325
naive	guess	mor	NLLB	0.4404	10.9578	181.0172
naive	guess	all	NLLB	0.424	12.2833	223.8087

Table 1.2 shows how standardized text (Bavarian text that has been perturbed to resemble Standard German text) performs compared to the unaltered Standard German text. Again (PERT) indicates experiments of modifying text data, while (NLLB) indicates the translation to English compared to the Standard German text translated into English.

Table 1.3 shows how translating English text into German and then, in turn, perturbing this text into a more Bavarian text, compares to the original Bavarian sentences that have already been aligned with the English input sentences.

Table 1.4 shows how dialectized text (Standard German text that has been perturbed to resemble Bavarian text) performs compared to the unaltered Standard

Table 1.3: Evaluation Metrics for English-Bavarian (dialectized as part of postprocessing)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	all	POST	5.092	15.8652	83.6179
naive	guess	mor	POST	5.1327	17.119	83.5147
naive	guess	lex	POST	29.1958	59.5913	54.0825

Table 1.4: Evaluation Metrics for German-Bavarian (dialectized)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	PERT	51.5588	72.5826	24.3256
naive	guess	mor	PERT	10.2236	18.9192	70.7583
naive	guess	all	PERT	10.2056	17.3224	70.7903
naive	guess	lex	NLLB	38.2932	51.2928	65.3969
naive	guess	mor	NLLB	1.0201	9.3358	152.2364
naive	guess	all	NLLB	0.7376	8.2818	161.4027

German text. Again (PERT) indicates experiments of modifying text data, while (NLLB) indicates the translation to English compared to the Standard German text translated into English.