universität innsbruck

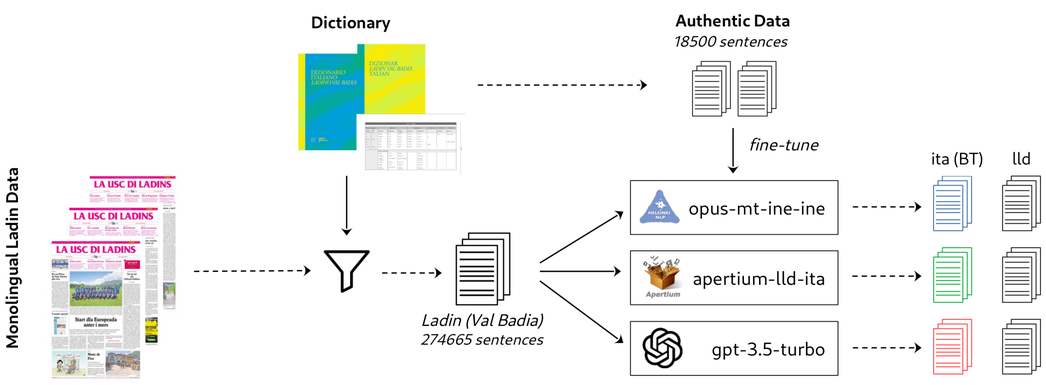
LoResMT @ ACL 2024

Rule-Based, Neural and LLM Back-Translation

Comparative Insights from a Variant of Ladin

Samuel Frontull and **Georg Moser** University of Innsbruck, Austria

Introduction



Back-translation is **a technique to augment the training data** for machine translation systems by translating monolingual data from the target language back to the source language. The resulting synthetic parallel data can then be used to train a translation model in the opposite direction.

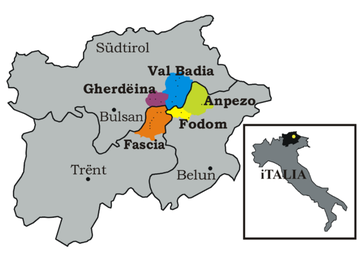
In our paper, we analyze **how different systems** used for back-translation **impact the models** trained for the language pair **Ladin-Italian**. For this, we use:

- the `Helsinki-NLP/opus-mt-ine-ine` multilingual **neural** model, which is fine-tuned with the small amount parallel data that is available, which we label `N1`,
- a **rule-based** system, `R1`, which we developed specifically for Ladin-Italian
- and the GPT-3.5 Turbo model, `L1`, which is a **large language model**.

All these systems have different strengths and weaknesses. Therefore, they are **likely to influence** the quality of the final models **in distinct ways**.

The Ladin Language

The Ladin language is a **Romance language** spoken in the nordest part of Italy, in the regions of South Tyrol, Trentino and Belluno. It has around **30,000 native speakers** that are spread over **5 valleys** with **each having its own variant** of the language, also in the written form. In our work, we restrict ourselves to the **variant of the Val Badia** valley.



https://commons.wikimedia.org/wiki/File:Ladin.png

The main source of monolingual data for Ladin is the weekly newspaper ***La Usc di Ladins*** from which we extracted **274,665 monolingual sentences** for the specific variant of Val Badia. There is also a digitised dictionary available which we use as basis for developing the rule-based machine translation system. Moreover, the dictionary also contains a set of **18k basic parallel sentences**. For **testing**, we collected parallel sentences from **three different domains**.

Testset 1

- Legal texts
- 424 sentences

Testset 2

- History of the region
- 833 sentences

Testset 3

- Fairy tale
- 1563 sentences

Translation Models

We used the `Helsinki-NLP/opus-mt-ine-ine` (`BM`) model as the base model for the experiments. In total, we trained the following 15 models:

- Model `N1`: `BM` fine-tuned with the available parallel data (18,139 sentences).
- Models `N2`/`R2`/`L2`: BM fine-tuned with authentic data and Ladin monolingual data back-translated to Italian using `N1`/`R1`/`L1` respectively.
- Models `N3`/`R3`/`L3`: This iteration extends the training base of `N2`/`R2`/`L2` by adding Italian monolingual data translated into Ladin utilizing `N2`/`R2`/`L2` respectively.
- Models `N4`/`R4`/`L4`: `BM` fine-tuned with the same training data as `N3`/`R3`/`L3` models, but with Ladin and Italian monolingual data backtranslated with `N3`/`R3`/`L3` model.
- Models `N5`/`R5`/`L5`: This iteration extends the training base of `N4`/`R4`/`L4` by adding also the forward-translations as training data.
- Models `A1`/`A2`: `A1` was trained on the combined training data used to train `N4`, `R4`, and `L4`. For `A2`, we additionally included the forward-translations.

Results

In the following table, we report the BLEU scores of the different models on the three testsets, in both translation directions:

		Italian → Ladin (Val Badia)			Ladin (Val Badia) → Italian		
		Testset 1	Testset 2	Testset 3	Testset 1	Testset 2	Testset 3
NMT `opus-mt-ine-ine`	`BM`	0.08	0.55	0.05	8.17	8.07	2.29
`BM` fine-tuned with authentic data	`N1`	10.22	10.14	12.76	12.65	11.49	11.83
+ *lvb* monolingual BT with `N1`	`N2`	19.09	18.05	16.50	13.01	12.40	13.23
+ *ita* monolingual BT with `N2`	`N3`	19.54	19.45	16.66	21.98	19.37	15.01
+ *lvb* and *ita* monolingual BT with `N3`	`N4`	19.61	19.16	16.40	22.90	21.12	16.17
+ *lvb* and *ita* monolingual FT with `N3`	`N5`	20.24	19.39	15.56	21.49	20.53	15.10
RBMT `apertium-lld-ita`	`R1`	4.94	4.50	3.19	11.38	11.60	8.48
`BM` fine-tuned with authentic data							
+ *lvb* monolingual BT with `R1`	`R2`	19.18	16.96	15.21	14.43	13.27	13.99
+ *ita* monolingual BT with `R2`	`R3`	19.86	17.70	15.04	22.17	19.27	15.89
+ *lvb* and *ita* monolingual BT with `R3`	`R4`	20.93	19.32	16.65	21.36	20.27	16.34
+ *lvb* and *ita* monolingual FT with `R3`	`R5`	19.97	18.65	16.61	22.50	20.37	16.36
LLM `gpt-3.5-turbo-0125`	`L1`	5.54	3.84	1.16	26.77	21.17	10.37
`BM` fine-tuned with authentic data							
+ *lvb* monolingual BT with `L1`	`L2`	22.09	19.71	14.16	12.93	12.21	13.22
+ *ita* monolingual BT with `L2`	`L3`	21.59	19.96	14.23	22.69	20.37	15.26
+ *lvb* and *ita* monolingual BT with `L3`	`L4`	20.82	19.87	16.55	23.01	21.38	15.12
+ *lvb* and *ita* monolingual FT with `L3`	`L5`	20.93	19.38	15.84	23.11	20.86	15.19
ALL `BM` fine-tuned with authentic data							
+ *lvb* and *ita* monolingual BT with `N3`, `R3`, `L3`	`A1`	19.83	19.94	16.54	23.58	21.30	15.32
+ *lvb* and *ita* monolingual FT with `N3`, `R3`, `L3`	`A2`	20.81	19.71	16.36	24.12	22.24	15.98

We observed only minimal differences in BLEU scores. However, the following example shows that the models behave differently and that **characteristic errors persist**.

Ladin → Italian

English	South Tyrol is an autonomous region of Italy in which about 500,000 people live, among them the Ladins
Ladin	Südtirol é na regiun autonoma dla Talia olache al vir incër 500 mile persones, danter chëstes ince i ladins.
`N4`	Southampton è una regione autonoma dell'Italia dove vivono circa 500 000 persone, tra cui queste anche i ladini.
`R4`	Sudtirolo è una regione autonoma in Italia dove vive circa 500 miglia persone, fra queste anche i ladini.
`L4`	Il Sudafrica è una regione autonoma dell'Italia in cui vivono circa 500 000 persone, tra cui i ladini.
Italian	Il Sudtirolo è una regione autonoma dell'Italia in cui vivono circa 500 mila persone, tra queste anche i ladini.

Italian → Ladin

English	It seems unbelievable to me, and yet, sometimes, life confronts us with extraordinary situations that we never expected.
Italian	Mi sembra incredibile, eppure, a volte, la vita ci pone di fronte a situazioni straordinarie che non ci saremmo mai aspettati.
`N4`	Al me pé incherscimun, y bel, avisa, la vita nes mët dant situaziuns straordinares che ne s'esson mai aspeté
`R4`	Al me sá nia da crëie, impó, de iadi, la vita nes mët dant a situaziuns straordinares ch'i ne fosson mai aspetá
`L4`	Al me pé nia da crëie, y scëmpl, gonot, la vita nes mët dant colaboraziuns straordinares che an ne s'ess mai aspeté
Ladin	Al me pé nia da crëie, impó nes mët la vita datrai dant situaziuns straordinares ch'i ne s'esson mai aspeté.

ISTITUT LADIN MICURÁ DE RÜ

REGIONE AUTONOMA TRENTINO-ALTO ADIGE
AUTONOME REGION TRENTINO-SÜDTIROL
REGION AUTONÓMA TRENTIN-SÜDTIROL

This research was carried out within a project funded by *Regione Autonoma Trentino-Alto Adige/Südtirol*, in collaboration with the Ladin Cultural Institute *Micurà de Rü*.

www.uibk.ac.at