Contents

Chapter 1

Introduction

1.1 Motivation

The Romansh language is a Romance language spoken in Switzerland, primarily in the Canton of Grisons (henceforth Graubünden). Graubünden is the only canton in Switzerland with three official languages—German, Italian and Romansh. The number of Romansh speakers, 40,000, has been sinking in the last decades (Bundesamt für Statistik 2020). In order to protect Romansh from extinction, Graubünden braced the protection and the promotion of multilinguality within its borders in its constituion:

> Kanton und Gemeinden unterstützen und ergreifen die erforderlichen Massnahmen zur Erhaltung und Förderung der rätoromanischen und der italienischen Sprache[1]. (Art. 3 Abs. 2 der Bündner Verfassung[2])

Additionally, in 2006 a language law (*Sprachengesetz*) with the aim of further promoting and protecting the multilinguality of the canton:

> Dieses Gesetz bezweckt: ... e) die bedrohte Landessprache Rätoromanisch mit besonderen Massnahmen zu unterstützen[3] (Abs. 1 Art. 1 Bst. e des Sprachengesetz des Kantons Graubündens[4]);

Since 1997, the majority of all press releases published by the Canton Graubünden were released in these three langauges. This existence of such parallel documents in three langauges lends itself to the collection and the compliation of a trilingual parallel corpus. Of special interest is here the Romansh language, which, having such a low number of speakers, may be considered a "low resource language".

[1] The canton and the communities shall support and take the required measures to maintain and promote the Romansh language and the Italian language.

[2] `https://www.gr-lex.gr.ch/app/de/texts_of_law/110.100`

[3] The law of languages of the Canton Graubünden is meant to: e) to support the endangered national language Romansh.

[4] `https://www.gr-lex.gr.ch/app/de/texts_of_law/492.100#structured_documentingress_foundation_fn_4417_2_2_c`

1.2 Research Question and Goals

1.2.1 Research Questions

Jalili Sabet et al. 2020 were able to show that their algorithm for word alignment, which is similarity based and uses word embeddings to compute similarity, outperforms all the statistical baseline models.

But not only that the model outperforms the existing stastical models, its biggest advantage as propogated by Jalili Sabet et al. 2020 is that it requires no training data. Statistical models will only reach a threshold of good performance with enough training data (Jalili Sabet et al. 2020; Och and Ney 2000). With word embeddings, words in just one single sentence can be aligned with high precision, without the need of a large set of sentence pairs for training a word alignment model. However, all of this works persuming we already have trained a multingual language model, whose learned embeddings we can leverage for this task. There exist some language models that were trained on multi-lingual data. mBERT was trained on 104 languages[5], LASER was trained on 93 languages(Artetxe and Schwenk 2019) and XLM-RoBERTa base was trained on 100 languages (Conneau et al. 2020).

Multilingual language models were shown to perform also well on unseen languages, dubbed as "zero-shot setting". Although the LASER model was pretrained on 93 languges, it obtained strong results for sentence embeddings in 112 languages (Artetxe and Schwenk 2019). It was also shown that mBERT performs well on unseen languages in a variety of tasks such as Named Entity Recognition (NER) and Part of Speech (POS) tagging (Pires, Schlinger, and Garrette 2019).

There is, then, good reason to believe that similarity based word alignment using multilingual word embeddings would work also for the case of German–Romansh or Italian–Romansh. Especially since vocabulary overlaps between unseen and seen languages favor performance in zero-shot settings (Pires, Schlinger, and Garrette 2019), and since Romansh displays a high similarity with other seen Romance languages, e.g., Italian, French, Spanish. English also has a large portion of Romance-based vocabulary.

1.2.2 Goals

My goals for this thesis are twofold:

- Test whether similarity based word alignment using multilingual word embeddings will perform on par with statistical word alignment models on the uneen language Romansh;
- Collect the press releases of the canton Graubünden, published in German, Romansh and Italian, and compile a parallel trilingual corpus.

To test the quality of the word alignments, I will create a gold standard and manually annotate word alignment for German-Romansh sentence pairs.

After finishing my work, I will make my gold standard and the corpus I compiled available for further research by future students.

[5] https://github.com/google-research/bert/blob/master/multilingual.md

1.3 Structure

In the course of the following pages I will first give a short introduction to the Romansh language (Chapter ??), then describe how I collected the data and aligned the documents (Chapter ??) and how I further aligned the sentences to create sentence pairs (Chapter ??). In Chapter ?? I will shortly explain the mechanism behind word alignment methods. Finally, I will explain how and according to which guidelines I created the gold standard (Chapter ??) and show the results my experiments comparing different word aligning systems (Chapter ??).

Throughout this work, I went to effort to not become too technical in details, always writing to an imaginary fellow Linguistics student, such that this work, if it ever falls in the hands a future student, will be comprhensible and readable. I hope that it will be read by and inspire future students, the way I was inspired by master's theses written before me.

1.4 GitHub repository

The code I wrote and the data I collected in the course of this work is available on my GitHub repository. Please contact me in order to gain access to it.

Glossary

Graubünden The Canton of Grisons. 1, 4

Acronyms

AER Average Error Rate. 32, 40, 42, 43, 45, 52

NER Named Entity Recognition. 2

POS Part of Speech. 2

List of Tables

List of Figures

Bibliography

Artetxe, Mikel and Holger Schwenk (Sept. 2019). "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". In: *Transactions of the Association for Computational Linguistics* 7, pp. 597–610. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00288. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00288/1923278/tacl_a_00288.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00288.

Bundesamt für Statistik (2020). *Hauptsprachen in der Schweiz - 2020*. URL: https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/sprachen-religionen/sprachen.assetdetail.21344032.html (visited on 06/17/2022).

Conneau, Alexis et al. (July 2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

Jalili Sabet, Masoud et al. (Nov. 2020). "SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 1627–1643. DOI: 10.18653/v1/2020.findings-emnlp.147. URL: https://aclanthology.org/2020.findings-emnlp.147.

Och, Franz Josef and Hermann Ney (Oct. 2000). "Improved Statistical Alignment Models". In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, pp. 440–447. DOI: 10.3115/1075218.1075274. URL: https://aclanthology.org/P00-1056.

Pires, Telmo, Eva Schlinger, and Dan Garrette (July 2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493.