

Contents

Abstract	i
Acknowledgements	i
1 Introduction	1
1.1 Motivation	1
1.2 Research Question and Goals	2
1.2.1 Research Questions	2
1.2.2 Goals	2
1.3 Structure	3
1.4 GitHub repository	3
2 Romansh	3
2.1 Rhaeto-Romance	3
2.2 Romansh	4
2.3 Rumantsch Grischun	5
2.3.1 Lia Rumantscha	5
2.3.2 Rumantsch Grischun	5
2.3.3 Properties	5
2.3.4 Today	6
3 Compiling the Corpus	7
3.1 Introduction	7
3.2 Collecting the Data	7
3.3 Web Scraping	7
3.4 Building the Corpus	9
3.4.1 HTML Parsing	9
3.4.2 Document Alignment	10
3.5 Manual alignment of unlinked documents	13
3.6 SQLite database	13
3.7 Summary	13
4 Sentence Alignment	15
4.1 Introduction	15

4.1.1	Formal definition	15
4.2	Method Overview	16
4.2.1	Length Based	16
4.2.2	Partial Similarity Based	16
4.2.3	Translation based	17
4.2.4	Hybrid models	17
4.2.5	Summary	17
4.3	More Recent methods	18
4.3.1	Bleualign	18
4.3.2	Vecalign	19
4.4	Sentence alignment pipeline	19
4.4.1	Tool of choice	19
4.4.2	Pipeline	20
4.4.3	Sentence segmentation	20
4.4.4	Aligning language pairs	21
4.4.5	Filtering and tokenizing	22
4.5	Results	22
5	Word Alignment	24
5.1	Introduction	24
5.2	Overview of Methods	25
5.2.1	IBM Model 1	25
5.2.2	Higher IBM Models	26
5.3	Word Embeddings	27
5.3.1	Excursion: Words	27
5.3.2	Word Embeddings	28
5.3.3	Word Similarity	29
5.3.4	Multilingual Word Embeddings	29
5.3.5	Summary	30
5.4	Similarity Based Word Alignment	30
5.4.1	Method	30
5.4.2	Summary	32
6	Gold standard	33
6.1	Introduction	33
6.2	Sure and Possible Alignments	33
6.3	Evaluation Metrics	34
6.4	Gold standard for German-Romansh	34
6.4.1	Annotation tool	34
6.4.2	Guidelines	35
6.4.3	General principles	35
6.4.4	Examples	36
6.5	Flaws	38

7	Results	40
7.1	Evaluation Metrics	40
7.2	Baseline Systems	41
7.2.1	fast_align	41
7.2.2	eflomal	41
7.2.3	Performance	41
7.3	SimAlign	42
7.3.1	Performance	42
7.4	Discussion	42
7.4.1	General Problems with Evaluation	43
7.5	Summary	45
8	Concluding Words	47
8.1	Goals	47
8.2	Corpus Compilation	47
8.3	Gold Standard	47
8.4	Evaluation	48
8.5	Future	48
	List of Tables	50
	List of Figures	51
	Bibliography	52
A	Algnment Examples	57

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 constitution:

Kanton und Gemeinden unterstützen und ergreifen die erforderlichen Massnahmen zur Erhaltung und Förderung der rätoromanischen und der italienischen Sprache¹. (Art. 3 Abs. 2 der Bündner Verfassung²)

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 unterstützen³ (Abs. 1 Art. 1 Bst. e des Sprachengesetz des Kantons Graubündens⁴);

Since 1997, the majority of all press releases published by the Canton Graubünden were released in these three languages. This existence of such parallel documents in three languages lends itself to the collection and the compilation 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”.

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

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

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

⁴https://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 statistical models, its biggest advantage as propagated 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 presuming we already have trained a multilingual 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 languages⁵, 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 languages, 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 unseen 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

2.1	Examples for choosing the forms for Rumanstch Grischun, based on liver1999 . . .	6
4.1	Parallel corpus in numbers, as of July 20, 2022. “Source” refers to the language on the left and “target” to the language on the right, and not necessarily to the actual source language of the translation.	22
6.1	Translation examples of German compounds into Romansh	36
7.1	Evaluation metrics for word alignments with the baseline models for different dataset sizes. “Dataset Size” refers to the number of sentence pairs.	42
7.2	Evaluation metrics for word alignments using SimAlign, with different embeddings and word/sub-word level. Best result per embedding type in bold.	43
7.3	Comparison of the best performance of each of the three methods. The best value in each column is in bold.	43

List of Figures

2.1	Distribution of Rhaeto-Romance, taken from haiman1992	4
3.1	Directory tree of corpus_builder	8
3.2	Directory scheme for saving the HTML files	8
3.3	Portion of automatically aligned press releases up to 2009	11
4.1	Sentence alignment pipeline	19
5.1	Word alignment example	24
5.2	Similarity matrix	31
5.3	Alignment matrix	31
5.4	The resulting word alignment	31
6.1	Aligning German perfect to Romansh perfect	37
6.2	Alignment of German preterite to Romansh perfect	37
6.3	Aligning German present participles to Romansh relative clauses	38
7.1	Comparing precision between the systems for different dataset sizes.	44
7.2	Comparing recall between the systems for different dataset sizes.	44
7.3	Comparing AER between the systems for different dataset sizes.	45
A.1	Word alignment example 1	57
A.2	Word alignment example 2	58
A.3	Word alignment example 3	58
A.4	Word alignment example 4	59
A.5	Word alignment example 5	59

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/tac1_a_00288. eprint: https://direct.mit.edu/tac1/article-pdf/doi/10.1162/tac1_a_00288/1923278/tac1_a_00288.pdf. URL: https://doi.org/10.1162/tac1%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>.