Contents

1 Introduction **2**
- 1.1 Motivation . . . 2
- 1.2 Research Question and Goals . . . 2

2 Romansh **4**

3 Web scraping **5**

4 Document alignment **6**

5 Sentence alignment **7**

6 Word alignment **8**

7 Gold standard **9**
- 7.1 Introduction . . . 9
- 7.2 Sure and Possible Alignments . . . 9
- 7.3 Evaluation Metrics . . . 10
- 7.4 Gold standard for German-Romansh . . . 10
  - 7.4.1 Annotation tool . . . 10
  - 7.4.2 Guidelines . . . 10
  - 7.4.3 General priniciples . . . 11
  - 7.4.4 Examples . . . 12
- 7.5 Flaws . . . 14

8 Evaluation **16**

9 Contributions **17**

10 Summary **18**

Bibliography **18**

Chapter 1

Introduction

1.1 Motivation

The Romansh language is a Romance language spoken in Switzerland, primarily in the Canton of Grisons (henceforth *Graubünden*), by around 60,000 speakers. Graubünden is the only canton in Switzerland with three official languages – German, Italian and Romansh.

When traveling by train, the announcements are heard in German, Romansh or Italian according to which part of the canton one is currently at. It is enough to travel to the next valley to suddenly be greeted on the street in a different language. Newspapers, radio and television exist in all three languages, but also official documents, laws and press releases are published trinlingually. While I was resident in Graubünden, I was fascinated by this multilinguality and it was my wish to somehow capture it and make it available to others. This is why I decided to build a multilingual corpus, a parallel collection of sentences in German, Romansh and Italian, in which the sentences are translations of each other.

Having such a low number of speakers makes it a so-called low resource language. Having so little speakers means there is also little data, be it corpora or research data. Most of the reasearch in NLP focuses on high resource languages.

1.2 Research Question and Goals

Jalili Sabet et al. 2020 were able to show that their algorithm for word alignment outperforms all the statistical baseline models. Contrary to statistical models, their model uses vectors of word representations learned by a neural net (also commonly known as word embeddings) and, by using some sort of similarity measurement (e.g., cosine similarity), aligns the most similar words in the source and the target sentence.

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 (TODO: cite numbers from SimAlign). Using word embeddings can be used to align words in just a single sentence with high precision. Of course, all of this works persuming we already have a trained model whose learned embeddings we can use for this task. There exist some language models that were trained on multi-

lingual data. mBERT was trained on 104 languages and LASER was trained on 93 languages. But will word embeddings based word alignment will work in zero-shot settings? That is, can the embeddings learned by a multilingual language model be used for word alignment for a language that wasn't included in the training data?

Chapter 2

Romansh

In this chapter, I will provide a short context about Romansh.

In 1873, an Italian linguist by the name of Grziadio Ascoli pointed out to a shared number of chracterizing phenomena in a number of Romance dialects spoken and named this group of dialects "Ladino". Since 1883, influenced by Theodor Gartner's publication *Raetoromanishce Grammatik* on this group of dialects, this name (German *Rätoromanisch*, English "Raeto-Romance") became associated with them. Raeto-Romance is spoken in three separated areas: in parts of the Swiss canton of Graubünden, in the Dolomotic Alps in northern Italy (Südtirol) and around the drainage basin of the Tagliamento river, between Venice and Trieste (Haiman and Benincà 1992, p. 1).

There have been long discussions in Romance linguistics about whether Raeto-Romance can be seen as a unity of dialects, or whether such a unity is merely a linguistic construct, lacking a sociolinguistical-historical basis. This dispute, referred to as the *questione ladina* "the Ladin question" (Liver 1999). This question is not of importance to this thesis and will not bother us in the course of it. It is, however, important to remember that names and definitions posed by academics do not always correspond to the feelings of the speakers and their own sense of identity.

The term Roamansh is a collective name referring to the Raeto-Romance dialects spoken in Switzerland and is recognized as a single language (Haiman and Benincà 1992). They are currently spoken by around 40,000 people (Bundesamt für Statistik 2020). This number has been diminishing constantly – 30 years ago there were 50,000 speakers (Haiman and Benincà 1992).

Romansh was officially acknowledged as a fourth official language in Switzerland (besides German, French and Italian) in a federal referendum that took place in 1938, in the eve of the Second World War, with a whopping majority of 92% Yes votes. It has been shown that this referendum played in the hands of the raeto-romanians in Graubünden to promote their nationalistic political postulate, but was also instrumentalised by the Swiss federal government to counter Mussolini's pretensions to enquoteItalian terrotires in Switzerland (referred to as the Italian irredentism) (Valär 2012).

Chapter 3

Web scraping

Chapter 4

Document alignment

Chapter 5

Sentence alignment

Chapter 6

Word alignment

Chapter 7

Gold standard

7.1 Introduction

In order to measure the quality of words alignments, a model's performance is measured on a test set which is a gold standard created by human annotators. For the gold standard to be of good quality and consistent with itself, annotators have to follow strict guidelines. These guidelines address issues of ambiguity in word alignments. (Koehn 2009, p. 115).

Some problematic cases that might occur are function words (TODO) that have no clear equivalent in the other language. Koehn 2009 gives as an example the German-English sentence pair: *John wohnt hier nicht John does not live here*. What German word should the English word *does* be aligned to? Three different choices can be made:

1. The word should remain unaligned since it has to clear equivalent in German.
2. The word *does* is connected with *live*; it contains the number and tense information which is in German contained in one word *wohnt*, so it should be aligned to *wohnt*, together with *live*.
3. *does* is part of the negation; without it, the sentence would not contain this word. Therefore, *does* should be aligned with *nicht* (the German negation).

7.2 Sure and Possible Alignments

An approach for solving problematic cases is the distinction between *sure* (s) and *possible* (p) alignments (Och and Ney 2000), which are also sometimes referred as fuzzy alignments (Clematide et al. 2018). Generally, these labels allow to distinguish between ambiguous and unambiguous links. Ambiguous links are labeled *possible* and unambiguous links are labeled *sure* (Lambert et al. 2005). The *possible* label was conceived to be used especially for aligning words within idiomatic expressions, free translations and missing function words (Och and Ney 2000). This distinction also has an impact on the way the evaluation metrics are computed (more on that later).

There seems to be no clear global definition about which alignments should be considered as umabiguous and marked as *sure* and which should be considered ambiguous marked as *possible*. For some created gold standards, no distinciton between *sure* and *possible* alignments was made

(Clematide et al. 2018). In another case, annotators were asked to first label all alignments as *sure* and then refine their alignments with confidence labels (Holmqvist and Ahrenberg 2011). In the creation of the English-Icelandic gold standard in Steingrímsson, Loftsson, and Way 2021, annotators used only *sure* links. Their annotations were then combined, with all 1-1 alignments both annotators agreed upon (i.e., the intersection of their annotations) makred as *sure* and differences all other alignments made by either one or both were marked as *possible* (Steingrímsson, Loftsson, and Way 2021).

7.3 Evaluation Metrics

TODO: move to results/evaluation part Four types of measures have become standard for evaluating word alignment. Three of them – precision, recall and F-measure – are well known in Information Retrieval metrics Mihalcea and Pedersen 2003. The fourth, alignment error rate (AER) one was introduced by Och and Ney 2000.

7.4 Gold standard for German-Romansh

In order to measure the performance of both models, the embedding based model (SimAlign) and the stastitical model (fast_align), on the language pair German-Romansh a gold standard is needed. Since no such gold standard exists, I took upon myself to create one. Although I am not a speaker of Romansh, my experience as a trained linguist, as well as my knowledge in related languages (Latin, Italian, French), allows me to confidently tackle this task. Additionally, whenever I was in doubt, I referred to the online dictionary Pledari grond, which also offers a grammar overview. (TODO: add more grammar references)

7.4.1 Annotation tool

I used the tool *AlignMan* which was originally programmed for creating the gold standard for English-Icelandic by Steingrímsson, Loftsson, and Way 2021. It is quite easy to use and its code is readable. I also had to make some small changes to the code. For instance, the sentences to be aligned, while loaded into the database, were read in opposite order, such that the source language became the target language and vice versa. I fixed this issue, so that source (German) and target (Romansh) languages stay the same accross all applications.

As mentioned above, the tool does not allow labeling of links with *Sure* and *Possible*. Instead, AlignMan treats the union of 1-1 alignments made by two annotators as *Sure* alignments and all other alignments as *Possible*. This means, each annotator is expected to only annotate *Sure* alignments, which also applied to me while annotating the German-Romansh gold standard.

7.4.2 Guidelines

As mentioned above, clear guidelines need to be defined for creating the gold standard in order to ensure quality and consistency. I shall now proceed to describe the guidelines I used for my annotation of the word alignments for the gold standard.

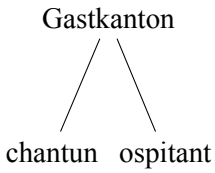
A motto cited often for annotating word alignments is "Align as small segments as possible, and as long segments as necessary" (Vronis and Langlais 2000, cited in Ahrenberg 2007). A variation of this is found in Clematide et al. 2018: "as few words as possible and as many words as necessary that carry the same meaning should be aligned.", referring to Lambert et al. 2005.

In the following sections I will list some general principles as well as more specific principles involving German and Romansh.

7.4.3 General priniciples

Principle I. Use only *Sure* alignments. Since the annotating tool I was using does not provide the use of confidence labels (cf. section 7.4.1), I only aligned words which would be considered *Sure* alignments, i.e., they are unambiguous (cf. section 7.2).

Principle II. Prefer 1-1 alignments over 1-n alignments or n-n alignments. Since all alignments are seen as *Sure* alignments, 1-n alignments should be avoided, unless a single word in the source sentence lexically corresponds to several words in the target sentence (see TODO principle sth.) This means alignments of phrases should be avoided. This is also due to the fact that we are testing models for automatic word, and not phrase alignments.



Words that are repeated in one language, but not in the other, should only be linked once, leaving the repetition unaligned.

Principle III. Lexical alignments should always be preferred over all other alignments (part-of-speech alignments or morphosyntactical alignments). This means alignments should describe first and foremost lexical correspondences, i.e., they have the same lexical meaning (but not necessarily share the same grammatical function or the same part-of-speech). Only words that are translations of each other also outside of the specific context of the sentence pair at hand should be aligned. This is in line with Clematide et al. 2018. In cases of paraphrasing during translations, words should remain unaligned (TODO: example?)

- only sure alignments
- prefer 1-1 alignments over 1-n alignments
- align words, not phrases
- only align words that are translations of each other also outside of context
- POS doesn't matter: German often prefers a nominal style, Romansh prefers a verbal style – expect some noun-verb alignments.

German	Romansh	
Beratungsstelle	*post* ***da*** *cussegliaziun*	"consultation point"
Gebäudeversicherung	*Assicuranza* ***d'****edifizis*	"building insurance"
Webseite	*pagina* ***d'****internet*	"web site"
Kindermasken	*mascrinas* ***per*** *uffants*	"children masks"
Brandversicherung	*assicuranza* ***cunter*** *fieu*	"fire insurance"
Gastkanton	*chantun ospitant*	"hosting canton"

Table 7.1: Translation examples of German compounds into Romansh

7.4.4 Examples

I will now supply some examples to illustrate the above principles.

Compound words

Compounding is the formation of new lexemes by adjoining two or more lexemes (Bauer 1988). In German, compounds are productive and prominent means of word formation in German (Clematide et al. 2018). In a sample of 4,500 types examined by Clematide et al. 2018, 80% of German nouns were compounds. Romansh, in comparison, uses prepositions (usually *da*) for linking nouns, with one noun modifying the other (Tscharner and Denoth n.d.). Other prepositions that can be found for linking words are *cunter* and *per*. [1] In other cases, German compounds might be translated to Romansh using an adjective + noun, e.g., German *Gastkanton* was translated to *chantun ospitant* "hosting canton". See table 7.1 for examples.

German compounds will be aligned to their equivalent lexical words, but not to function words, resulting in a 1-n alignment: *Webseite ~pagina [d'] internet*, *Gebäudeversicherung ~Assicuranza [d'] edifizis*. This is also inline with principles I, II and III in Clematide et al. 2018.

German preterite vs. Romansh perfect

In the corpus at hand, two tenses are used in German for referring to past events: the preterite and the perfect. The German preterite is a synthetic verb form, i.e., it is made up of a single conjugated form. Some examples are *nahm* (infinitive *nehmen* "take") or *wurde* (infinitive *werden* "become"). The German perfect is an analytic construction made up of an auxiliary verb (*haben* "have" or *sein* "be") and the past participle, e.g., *Die Präsidentenkonferenz* ***hat*** *nun* ***entschieden*** "The conference has decided".

In contranst to German, Romansh only has one tense referring to past events: the perfect. It is an analytic construction made, in a similar fashion as in German, of an auxiliary *habere* "have" for transitive verbs or *esse* "be" for intransitive verbs and the past participle (Bossong 1998, p. 189). The German sentence given above (*Die Präsidentenkonferenz hat nun entschieden*) was translated as *La conferenza da las presidentas e dals presidents* ***ha*** *usse* ***decidi***. *ha* is the auxiliary and *decidi*

[1] Typologically, this is inline with other Romance languages such as French, which uses prepositions (*de*, *en* and *à*) for linking two nouns, e.g., *une robe de soie* "a silk dress" (Price 2008)[510].

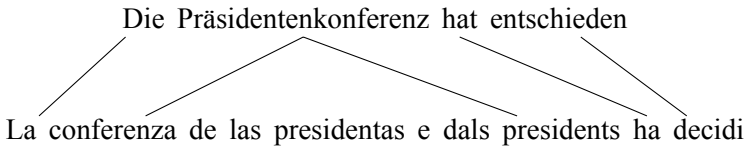


Figure 7.1: Aligning German perfect to Romansh perfect

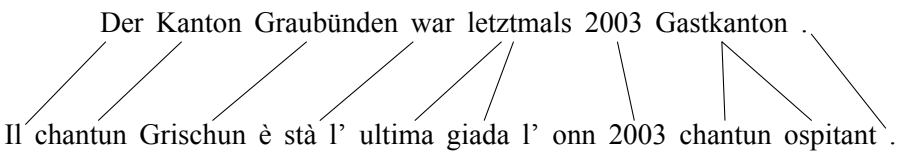


Figure 7.2: Alignment of German preterite to Romansh perfect

is the past participle. This poses no real problem since we can link the German auxiliary to the Romansh auxiliary and the German participle to the Romansh participle.

However, a German preterite is always translated using the Romansh perfect. For example, in the sentence *Der Kanton Graubünden war letzsmals 2003 Gastkanton* "The last time the Canton of Grisons was a host canton was in 2003" the verb *war* "was" is translated as *è stà*. This theoretically results in a 1-2 link. However, since the verb *è* here only carries grammatical information of tense and number, but no real lexical information, it should reamin unaligned.

The German perfect should be aligned to the Romansh perfect using a 1-1 alignment; auxiliary to auxiliary and participle to participle. **The German preterite should also be aligned using a 1-1 alignment to the Romansh participle, leaving the auxiliary unaligned and avoiding a 1-2 alignment.**

German present participle

German present participles (known in German as *Partizip I*) are translated to Romansh using relative clauses. Moreover, adjectives (and participles in the function of adjectives), can be nominalized, meaning they become the head of a noun phrase and there is no need for an actual noun. A good example for that in the corpus is the German noun phrase *nichtarbeitslose Stellensuchende* (cf. ex. 1), which was translated as a noun phrase with a relative clause: *persunas che tschertgan ine plazza che n'èn betg dischoccupadas* "persons who look for a job who are not unemployed".

(1) nicht-arbeit-s-los-e Stellen-such-end-e
not-work-gen-less-Pl job-search-pres.part-pl
"People looking for jobs who are not unemployed"

In this case, these two phrases should not be aligned as phrases, but only the content words

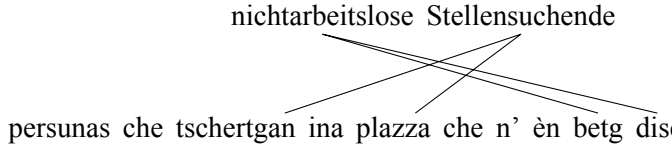


Figure 7.3: Aligning German present participles to Romansh relative clauses

which lexically correspond to each other: *nichtarbeitslose ~ betg dischoccupadas*; *Stellensuchende ~ tschertgan [ina] plazza*.

Double negation

Negation in Romansh is built using two particles: *na* and *betg* to negate verbs or *nagin-* to negate nouns. Since we prefer 1-1 alignments, the German negations *nicht* (for verbs) and *kein-* for nouns should be aligned only to the second Romansh particle (*betg*/*nagin-*), leaving Romansh *na* unaligned. Granted, this is also in favor of the SimAlign output, but it is also linguistically motivated: when negating the imperative form, *na* can be omitted required TODO:cite Grammatica per l'instrucziun dal rumantsch grischun.

Articles and prepositions

German articles inflect in case, which expresses some syntactic relations between nouns. Romansh often uses preopsitions for expressing the same relations. For instance *Zustimmung der Person* "the person's agreement" is translated as *consentiment da la persuna*. I align the German article *der* with Romansh *da*, leaving *la* unaligned. Except for my preference for 1-1 alignments, the motivation for this is that it is the preposition *da* that expresses the genitival relations between the nouns.

Separable verbs

German uses many verbs to which an adverb or a preopisition is affixed in order to delimit the verb's meaning (or sometimes completely change its meaning). In such cases, both the verb and its affix should be aligned to the corresponding Romansh verb, resulting in a 2-1 alignment.

7.5 Flaws

I shall now discuss the quality of my gold standard and some flaws it has.

The most obvious flaw is the fact that I created the gold standard alone. With more than one annotator, more intricate annotating schemes can be used in order to ensure higher quality, consistency and harmony. For instance the annotators' agreement can be measured using the so-called inner-annotator agreement (Holmqvist and Ahrenberg 2011). Further, the intersection of the annotators' *Sure* alignment can be used to build the final *Sure* alignments set and the reunion of the *Possible* alignments can be used to create the final *Possible* alignments set Mihalcea and Pedersen 2003. A third annotator can also revise and resolve conflicts between two annotaors Mihalcea and Pedersen 2003. When several annotators work on the same task, they can also discuss conflicts and resolve them using a majority vote (Melamed 1998).

All of these possible schemes cannot be realized in my case.

Another flaw is the missing confidence labels (*Sure* and *Possible*), which may influence the evaluation scores. Doing without *Possible* links and using only *Sure* links is however precedented (Clematide et al. 2018; Mihalcea and Pedersen 2003) and hence defensible.

In order to test my own consistency, I have re-annotated the first 100 sentences in the sample. TODO: results

Despite of the flaws mentioned, I am certain that gold standard is of high quality and consistency, due to the fact that I was also the one to define the guidelines.

Chapter 8

Evaluation

Chapter 9

Contributions

Chapter 10

Summary

Bibliography

Ahrenberg, Lars (2007). *LinES 1.0 Annotation: Format, Contents and Guidelines*. Tech. rep.

Bauer, Laurie (1988). *Introducing Linguistic Morphology*. Edinburgh University Press.

Bossong, Georg (1998). *Die Romanischen Sprachen: Eine vergleichende Einführung*. Hamburg: Helmut Buske Verlag.

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.

Clematide, Simon et al. (2018). "A multilingual gold standard for translation spotting of German compounds and their corresponding multiword units in English, French, Italian and Spanish". In: *Multiword Units in Machine Translation and Translation Technology*. Ed. by Ruslan Mitkov et al. John Benjamins, pp. 125–145. doi: https://doi.org/10.1075/cilt.341.

Haiman, John and Paola Benincà (1992). *The Rhaeto-Romance Languages*. Londn and New York: Routledge. isbn: 0-415-04194-5.

Holmqvist, Maria and Lars Ahrenberg (May 2011). "A Gold Standard for English-Swedish Word Alignment". In: *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*. Riga, Latvia: Northern European Association for Language Technology (NEALT), pp. 106–113. url: https://aclanthology.org/W11-4615.

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.

Koehn, Philipp (2009). *Statistical Machine Translation*. Cambridge University Press.

Lambert, Patrik et al. (2005). "Guidelines for Word Alignment Evaluation and Manual Alignment". In: *Language Resource and Evalution* 39, pp. 267–285. doi: 10.1007/s10579-005-4822-5.

Liver, Ricarda (1999). *Rätoromanisch: Eine Einführung in das Bündnerromanische*. Tübingen: Narr.

Melamed, I. Dan (1998). "Annotation Style Guide for the Blinker Project". In: *CoRR* cmp-lg/9805004. url: http://arxiv.org/abs/cmp-lg/9805004.

Mihalcea, Rada and Ted Pedersen (2003). "An Evaluation Exercise for Word Alignment". In: *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pp. 1–10. url: https://aclanthology.org/W03-0301.

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.

Price, Glanville (2008). *A Comprehensive French Grammar*. Blackwell Publishing.

Steingrímsson, Steinþór, Hrafn Loftsson, and Andy Way (May 2021). “CombAlign: a Tool for Obtaining High-Quality Word Alignments”. In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, pp. 64–73. url: https://aclanthology.org/2021.nodalida-main.7.

Tscharner, Gion and Duri Denoth (n.d.). *Grammatikteil des Vallader / Grammatica valladar*. url: http://www.udg.ch/dicziunari/files/grammatica_vallader.pdf.

Valär, Rico Franc (2012). “Geschichte und Gegenwart des Rätoromanischen in Graubünden und im Rheintal”. In: ed. by Gerhard Wanner and Georg Jäger. Chur: Desertina. Chap. Wie die Anerkennung des Rätoromanischen die Schweiz einte. Einige Hintergründe zur Volksabstimmung vom 20. Februar 1938, pp. 101–116.

Vronis, Jean and Philippe Langlais (2000). *Evaluation of parallel text alignment systems - The ARCADE project*.