Contents

Glossary

Graubünden The Canton of Grisons. 1, 5, 9, 56

HTML Hypertext Markup Language. A language containing display instructions for web browsers and the format in which web pages are usually saved . 10

JSON JavaScript Object Notation. A format for organizing data in a hierarchical form. 11

Standeskanzlei State Chancellery of Grisons. 9, 56

URL Uniform Resource Locator. A reference to an internet resource, a web address. 10

Acronyms

AER average error rate. 38, 48, 50, 51, 52, 54, 61

EM exepctation-maximization. 31

gen genitive. 45

HTML Hypertext Markup Language. 10, 11, 15

JSON JavaScript Object Notation. 11, 12, 13, 15, 56, 58, 79

NER named entity recognition. 2

NMT neural machine translation. 24, 27

part participle. 45

pl plural. 45

POS part of speech. 2, 43, 55

pres present. 45

SMT statistical machine translation. 23

URL Uniform Resource Locator. 27, 56

List of Tables

List of Figures

List of Listings

Bibliography

Brown, Peter F. et al. (1993). “The Mathematics of Statistical Machine Translation: Parameter Estimation”. In: *Computational Linguistics* 19.2, pp. 263–311. URL: `https://aclanthology.org/J93-2003`.

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`.

Appendix C

Aligning Romansh to Italian

Due to the nature of my research question, I virtually ignored in the course of this work the issue of word alignments using embeddings (i.e., SimAlign) between Romansh and Italian. Therefore, I would like to curtly attend this issue in this appendix part.

Romansh and Italian share many similarities. Both of them are Romance languages and some researchers even consider Romansh to be a part of the Italian dialect continuum (see Section 2.1).

Since 1-to-many alignments and differing word order are more challenging to model than 1-to-1 alignments and similar or identical word order—word order or 1-to-many alignments are not modeled by IBM Model 1, but only by higher models (Brown et al. 1993)—one might expect that it should be easier to word-align languages that are more similar in structure, word order and grammar. That is, word-aligning Romansh to Italian should be easier than aligning Romansh to German due to the higher similarity between the former languages. Further, when dealing with unseen languages, as in the case of Romansh, multilingual language models have been shown to favor language similarity and vocabulary overlaps (Pires, Schlinger, and Garrette 2019). All this gives rise to the assumption that word alignment for Romansh–Italian might perform better.

I randomly hand-picked a few examples[1] and compared SimAlign's performance on the pairs Romansh-Italian and Romansh-German in order to unempircally[2] test this notion.

C.1 Examples

Figure C.1 is an example for a word alignment that works perfectly both with Italian and with German. In Figure C.2[3], word alignment works well with Italian and German exactly

[1] The only precondition was that the sentences be short; Visualization for longer sentences leaves something to be desired.

[2] Obviously, a gold standard for Romansh-Italian would be needed.

[3] Apologies for the somewhat unreadable edges in Romansh–German

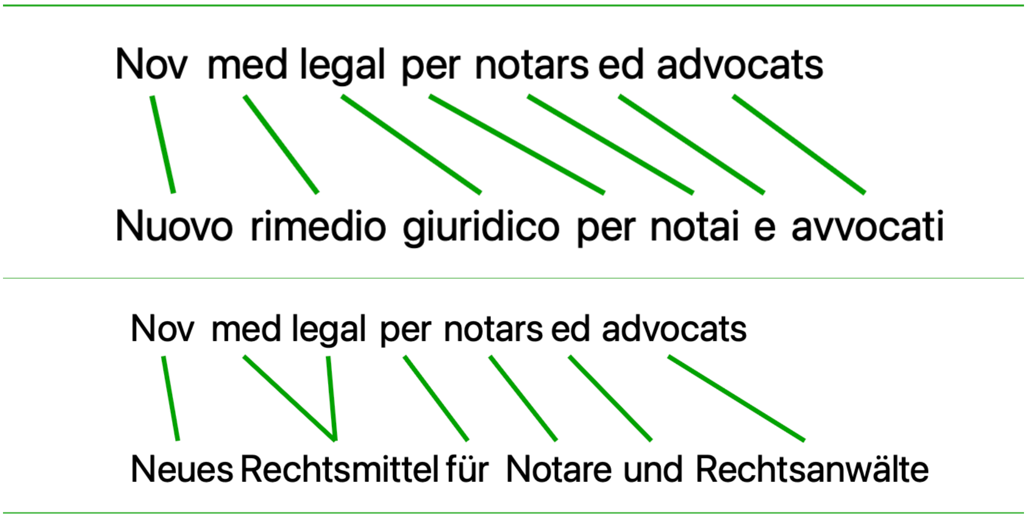
Nov med legal per notars ed advocats
Nuovo rimedio giuridico per notai e avvocati
Nov med legal per notars ed advocats
Neues Rechtsmittel für Notare und Rechtsanwälte

Figure C.1: Word alignment example Romansh–Italian and Romansh–German

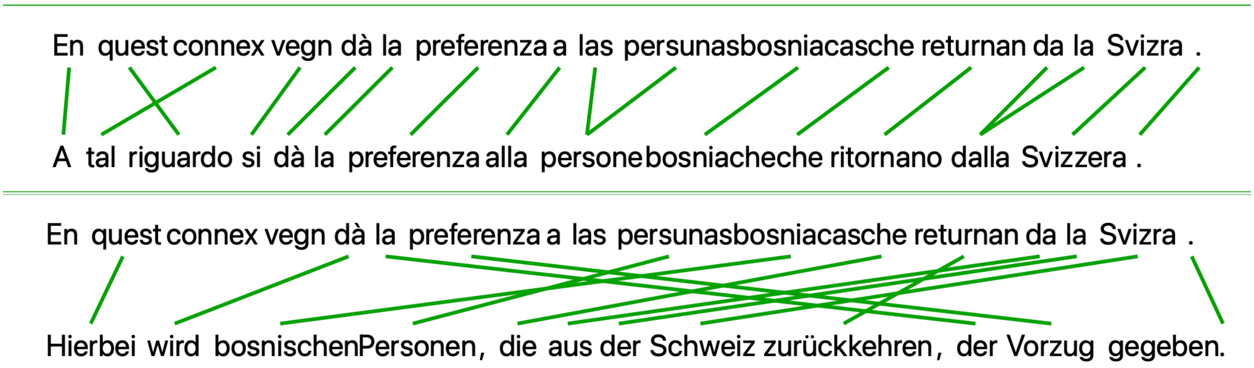
En quest connex vegn dà la preferenza a las persunas bosniacas che returnan da la Svizra .
A tal riguardo si dà la preferenza alla persone bosniache che ritornano dalla Svizzera .
En quest connex vegn dà la preferenza a las persunas bosniacas che returnan da la Svizra .
Hierbei wird bosnischen Personen , die aus der Schweiz zurückkehren , der Vorzug gegeben.

Figure C.2: Word alignment example Romansh–Italian and Romansh–German

for the same Romansh words, and it is exactly the same words where SimAlign fails: Romansh *en quest connex* ("in this context/matter") is not aligned correctly, neither in German nor in Italian. The same applies for Romansh *vegn* (literally "come", but here part of the passive construction), which is misaligned both times. This is also the case in Figure C.3. The same words are aligned correctly with German and with Italian, but in both cases Romansh *chantun* ("canton") remains unaligned.

In Figure C.4 word alignment with German is even better than with Italian. Here, every alignment is correct, whereas in the Italian example, Romansh *schilar* ("tackle") is not aligned to Italian *affronatare*, which should have been the case.

Finally, Figure C.5 is an example for many misalignments. In the German example, SimAlign succeeds in aligning Romansh *la derasaziuna da infecziuns* to German *die Durchseuchung*, but the rest of the alignments are wrong. The Italian example is completely misaligned.

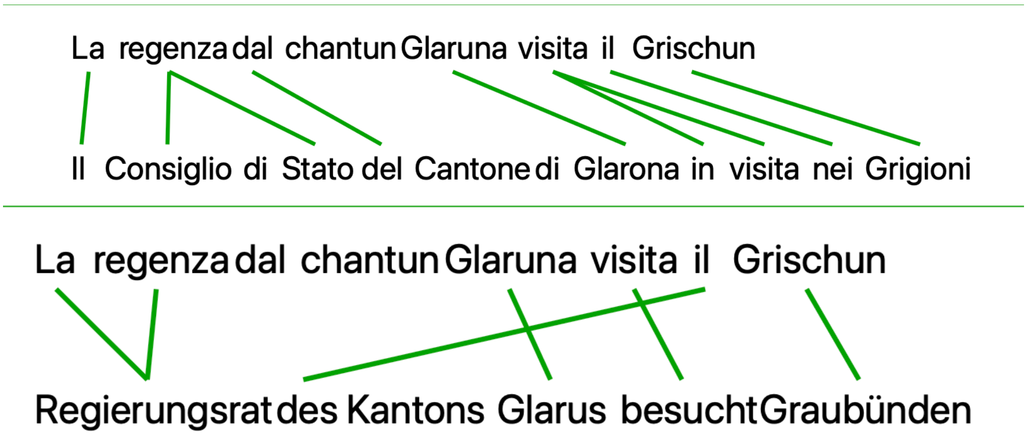
La regenza dal chantun Glaruna visita il Grischun
Il Consiglio di Stato del Cantone di Glarona in visita nei Grigioni
La regenza dal chantun Glaruna visita il Grischun
Regierungsrat des Kantons Glarus besucht Graubünden

Figure C.3: Word alignment example Romansh–Italian and Romansh–German

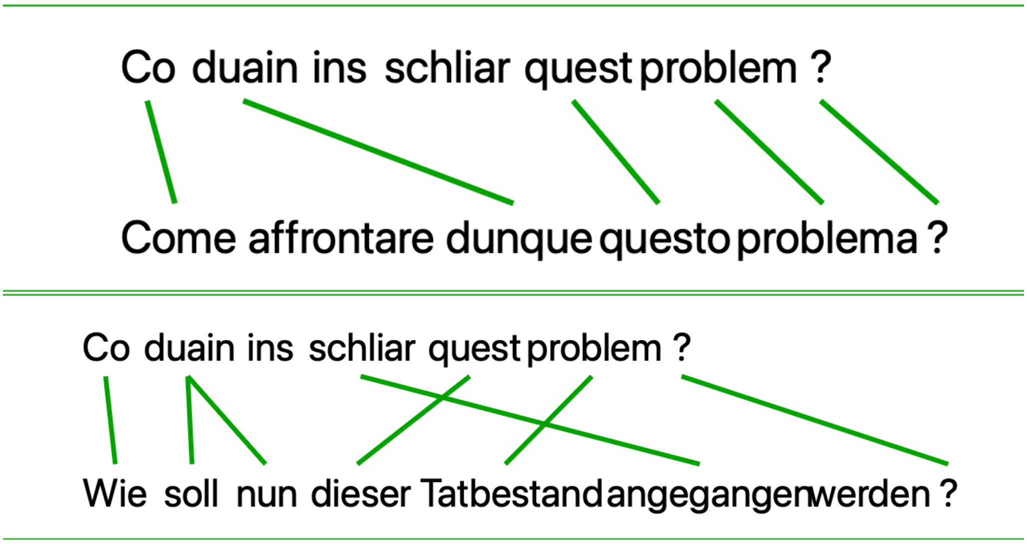
Co duain ins schliar quest problem ?
Come affrontare dunque questo problema ?
Co duain ins schliar quest problem ?
Wie soll nun dieser Tatbestand angegangen werden ?

Figure C.4: Word alignment example Romansh–Italian and Romansh–German

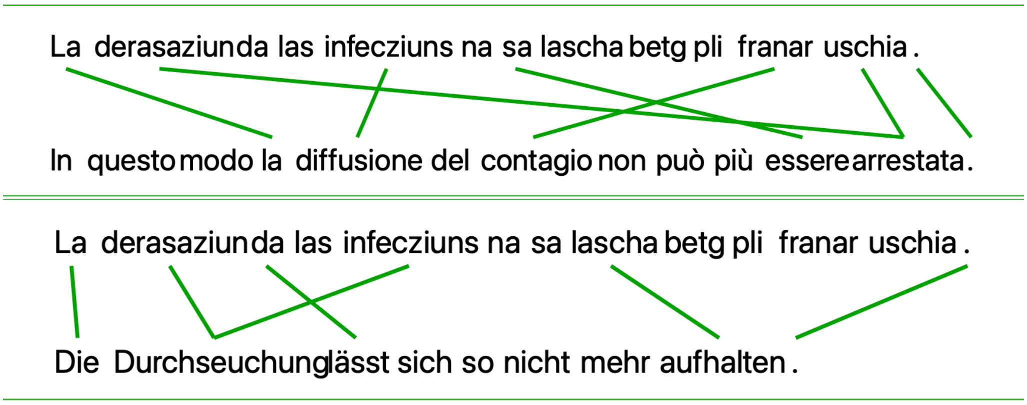
La derasaziun da las infecziuns na sa lascha betg pli franar uschia .
In questo modo la diffusione del contagio non può più essere arrestata .
La derasaziun da las infecziuns na sa lascha betg pli franar uschia .
Die Durchseuchung lässt sich so nicht mehr aufhalten .

Figure C.5: Word alignment example Romansh–Italian and Romansh–German

C.2 Summary

From observing these very few hand-picked cases, SimAlign doesn't seem to perform better when aligning Romansh to Italian. This is in spite of the higher similarity between Romansh and Italian, compared with German.

One possible explanation for this is that what mostly influences performance is the quality of the embeddings. If the Romansh word is similar enough to any of the words (or subwords) in the language model, alignment will work, regardless of the target language. Take for example Figure C.1. Here, all of the Romansh words are reminiscent of other seen languages and alignment works perfectly. However, in the case of Figure C.3, a suitable embedding for the Romansh word *chantun* cannot be looked-up for some reason, hence the word remains unaligned in both cases.