Contents

Chapter 1

Creating the corpus

It was difficult.

Chapter 2

Gold standard

2.1 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).

2.2 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 Steingrmsson, 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* (Steingrmsson, Loftsson, and Way 2021).

2.3 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.

2.4 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)

2.5 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.

2.5.1 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

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 2.1: Translation examples of German compounds into Romansh

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 2.1 for examples.

Principle I 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.

2.5.2 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* 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.

Principle II 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

[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].

aligned using a 1-1 alignment to the Romansh participle, leaving the auxiliary unaligned and avoiding a 1-2 alignment.

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.

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.

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.

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.

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.

Steingrmsson, 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.

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