Evaluating the State-of-the-Art Sentence Alignment System on Literary Texts

Edoardo Signoroni

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
e.signoroni@mail.muni.cz

Abstract. Sentence alignment is a useful task with many applications in Natural Language Processing and Digital Humanities. This paper presents an evaluation of Vecalign, the state-of-the-art method for automatic sentence alignment, on two bilingual corpora built from literary texts. This preliminary study shows that Vecalign performs well for literary texts and gives insights on its remaining issues through a qualitative evaluation of the output alignments.

Keywords: Parallel corpora · Automatic alignment · Literary text

Introduction

Sentence alignment is the Natural Language Processing (NLP) task of taking parallel documents split into sentences and finding a bipartite graph which matches minimal groups of sentences that are translation of each other [20]. In other words, to find target sentences with the same meaning to that of the source segments in multilingual texts [19].

This task is important to build bilingual corpora on which statistical Machine Translation (MT) systems could be trained. While neural MT approaches seem to be performing much better with sizable amounts of data, Kim et al. (2020) [6] shows that supervised and semi-supervised baselines outperform the best unsupervised systems.

Good alignment is also crucial for lexicography, as it can be leveraged to display parallel concordances and to find translation equivalents, and for terminology extraction.

Parallel corpora alignment is also being used in Digital Humanities (DH) with various purposes, such as historical language learning [10] or version alignment for medieval texts [8].

After a brief overview of the related work (Section 1), and a description of the methodology employed for this work (Section 2), the paper evaluates the performance of Vecalign [20] through a qualitative manual analysis (Section 3) of its automatic alignment of two corpora built from literary texts.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, pp. 115–124, 2021. © Tribun EU 2021

1 Related Work

This section will present some related work relevant to this study, firstly describing currently employed sentence alignment methods, and then briefly covering their application on literary texts and in DH.

1.1 Sentence Alignment Methods

The first automatic alignment methods were simple: they align sentences according to their length in words [4] or characters [5]. These algorithms do not work for text with sentences that have the same length, such as list of names or dates. Other systems worked with correspondence rules [17].

Newer approaches employ either external dictionaries or by training a translation model on the parallel text itself [22,9]. They also add some heuristics, such as limiting the search space to be near the diagonal. These systems, however, do not work with small texts because the occurrences of a given word are few. More recent methods introduced MT-based scoring [15,16], such as BLEU [11].

Steingrimsson et al. (2020) [19] review the current literature on the topic of sentence alignment and parallel corpora filtering. They then devise a new pipeline for aligning and filtering parallel corpora in sparse data conditions building on existing methods, such as those in Sennrich et. al. (2011) [16] and Artetxe et al. (2018a) [1]. Their proposed method is language pair independent and assumes unaligned bitexts and monolingual corpora.

The state-of-the-art systems use bilingual sentence embeddings, with their similarity used as the scoring function for alignment [20]. This is the method that it is employed for this paper, and it will be further described in Section 2.1.

The latest work on sentence alignment was presented at the Fifth Conference on Machine Translation (WMT2020), which featured a shared task on "Parallel Corpus Filtering and Alignment for Low-resource conditions" [7].

1.2 Work in Digital Humanities

Steinbach and Rehbein (2019) [18] demonstrate a pipeline for the parallelization and the annotation transfer for literary texts. For the sentence alignment they use Bleualign [16].

Meinecke, Wrisley, and Jänicke (2019) [8] employ the gensim implementation [14] of fastText [3] word embeddings and sentence embeddings similarity to compare and align different versions of the same medieval text.

The use of automatic alignment in DH is varied and broad. Some examples include Pataridze and Kindt (2018) [12], the Rosetta Stone project[1], or Zhekova et al. (2015) [24]. It seems common for these works to present their own domain-specific tools, such as UGARIT [2]. It is out of the scope of this paper to survey

[1] https://rosetta-stone.dh.uni-leipzig.de/rs/home/

[2] http://ugarit.ialigner.com/index.php

all the different application of automatic alignment systems in DH, nonetheless the examples above should give and idea of the variety of uses it has.

None of the work on literary texts or in DH seems to take advantage of Vecalign [20] as the state-of-the-art alignment system.

2 Methods

This section will discuss the methodology of this work, presenting the tools and the corpus on which they were tested.

2.1 Vecalign and LASER

Vecaling[3] was chosen as the automatic alignment system for two main reasons: i. it is the current state-of-the-art system; ii. it seems to be still untested on literary texts.

Vecaling propose a new scoring function based on the similarity of bilingual sentence embeddings. The method computes sentence embedding similarity scores with cosine similarity normalized with randomly selected embeddings. It then averages adjacent pairs of sentence embeddings in both documents and align these approximate embeddings, iteratively refining this alignment using the original embeddings and a small window around them.

Following the Vecalign paper, LASER[4] was used to compute the sentence embeddings. This tool is based on the architecture for language agnostic sentence embeddings presented in Artexte and Schwenk (2019) [2].

2.2 Corpora

Two corpora were used for the experiments: i. a manually aligned version of Lewis Carrol's "Alice's Adventures in Wonderland"[5]; ii. three versions of J.R.R. Tolkien's "The Hobbit".

The first corpus consists of 823 sentences from "Alice's Adventures in Worderland" manually aligned and reviewed by András Farkas in nine languages. Only English and Italian were considered. This corpus was considered as a possible gold-standard to automatically evaluate the performance of Vecalign, however this was proven to be problematic for several reasons, which will be mentioned in the following section.

The second corpus is from J.R.R. Tolkien's book "The Hobbit" [21]. Three unaligned editions in three different languages (English, Czech, and Italian) where collected. The full .txt files averaged around 2.200 lines.

Table 1 summarizes the size of the two corpora.

[3] `https://github.com/thompsonb/vecalign`

[4] `https://github.com/facebookresearch/LASER`

[5] Retrieved from `https://farkastranslations.com/books/Carroll_Lewis-Alice_in_wonderland-en-hu-es-it-pt-fr-de-eo-fi.html`

	number of lines	number of sentences
alice_en	824	824
alice_it	824	824
hobbit_en	1989	5770
hobbit_it	2372	5342

Table 1: Number of lines in the .txt and number of sentences after preprocessing.

3 Experiments and Evaluation

Since we are not dealing with text scraped from the web, or processed with Optical Character Recognition (OCR) algorithms, or otherwise overly noisy data, not much preprocessing was needed.

Alice's corpus did not need specific preprocessing: it was easily downloaded in .csv form and the sentences for English and Italian were stored in separate .txt files. LASER sentence embeddings were trained with standard parameters and Vecaling was run with default settings. The output alignment was stored as a .csv file. Since Vecaling gives its output alignments as pairs of lists of sentences IDs, these were leveraged to add the text of the sentences to the .csv to qualitatively evaluate the resulting alignment. In case of alignments between multiple sentences, these were split by the special character *$* in order for them to be distinguishable in the .csv.

The Hobbit's corpus underwent some preprocessing stages. The text was first obtained in .doc format, it was then converted into .txt to be processed. By doing so, some features of the book, such as illustrations, images, and page numbers were lost. The text was split into sentences with [13], even if LASER is capable of handling training of sentence embeddings from raw text. Future work may address if this step actually has any impact on the output, since a preliminary observation has shown that the text was divided in a different number of sentences by LASER and Stanza. Sentences were stored in a separate .txt file for each language. LASER and Vecalign were again used with their default configuration. The resulting alignments were stored in three .csv files.

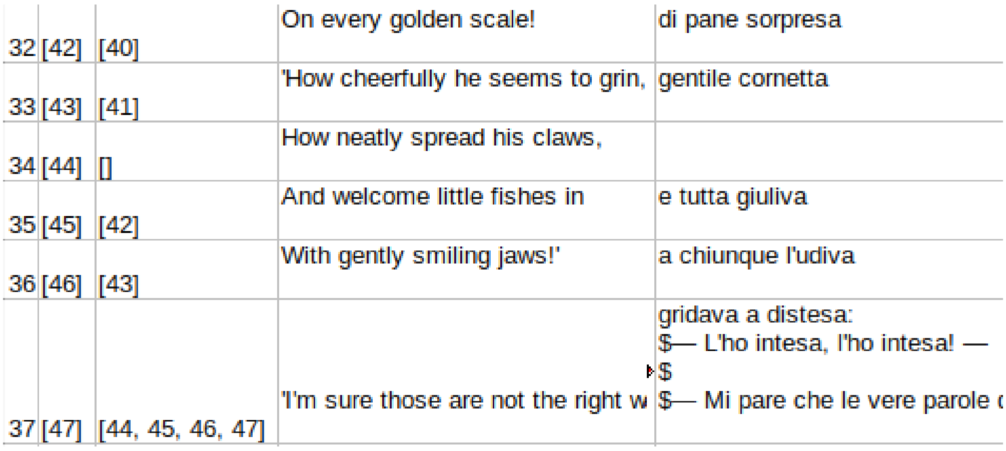
	start	mid	end	average
alice_en_it	85	98	100	**94,33**
hobbit_en_it	83	96	99	**92,67**

Table 2: Scores for the manual evaluation batches: the first (start), central (mid), and last (end) one hundred EN to IT alignments and the overall average score for each corpus.

Evaluation proved to be more complex than anticipated. Several automated methods were considered to evaluate the alignment quality. The Alice corpus was considered as reference for the design phase of the evaluation, since it has a gold standard. After taking an overview of the resulting outputs, an automated method of evaluation was tentatively devised. However, all of the proposed methodologies proved to be flawed. For example, a simple automated comparison between proposed alignment and gold standard alignment was revealed to be ineffective since it did not consider 1-to-many and many-to-one alignments. A MT-based method based on word lists comparison and BLEU score was considered, but proved to be unwieldy. Devising an automated evaluation method for The Hobbit corpus was even more challenging, since there was non gold standard available.

It was then decided to provide a qualitative evaluation of the results by manually assigning a score (0 for a bad alignment and 1 for a good alignment) to three batches of 100 alignments, one from the beginning, one from the main body, and one from the end. The scores were then averaged. Albeit simple, this method still provided some useful insights on the performance and the issues of Vecalign. The scores are given in Table 2

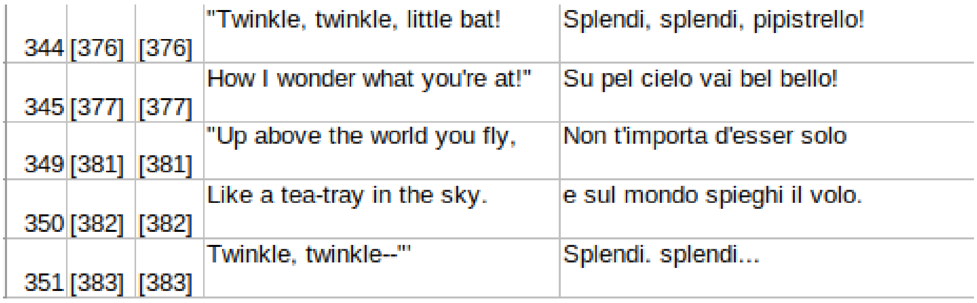
On the Alice corpus, 94.33% out of the 300 evaluated alignments where judged to be good. The first batch was the worst one, with 85/100, while the other two had respectively 98/100 and 100/100. Some interesting facts were uncovered by the analysis.



32	[42]	[40]	On every golden scale!	di pane sorpresa
33	[43]	[41]	'How cheerfully he seems to grin,	gentile cornetta
34	[44]	[]	How neatly spread his claws,	
35	[45]	[42]	And welcome little fishes in	e tutta giuliva
36	[46]	[43]	With gently smiling jaws!'	a chiunque l'udiva
37	[47]	[44, 45, 46, 47]	'I'm sure those are not the right w	gridava a distesa: $— L'ho intesa, l'ho intesa! — $ $— Mi pare che le vere parole (

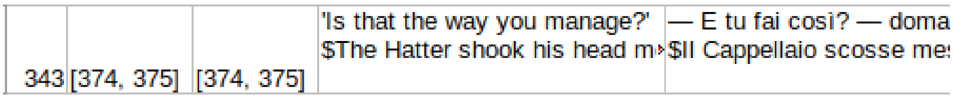
Fig. 1: The adaptation of a popular rime that confounds the alignment. The Italian version is not the translation of the English text.

First, while many of the alignments (a.) were correct, often they were not exact translations of the source sentences. This seems to hold true for the whole text, but some peculiar cases are rimes such as in a. 31 through 37 (Fig. 1) where not translated at all, but adapted to reflect the target culture. This also holds true for other translation choices as well, such as in a. 43 where the original reference to William the Conqueror is changed to Napoleon. The different adaptation seems to be irrelevant with regards to the performance if it is limited to a single



344	[376]	[376]	"Twinkle, twinkle, little bat!	Splendi, splendi, pipistrello!
345	[377]	[377]	How I wonder what you're at!"	Su pel cielo vai bel bello!
349	[381]	[381]	"Up above the world you fly,	Non t'importa d'esser solo
350	[382]	[382]	Like a tea-tray in the sky.	e sul mondo spieghi il volo.
351	[383]	[383]	Twinkle, twinkle--'"	Splendi. splendi...

Fig. 2: Another localized popular rime. In this case, however, the alignment is maintained.

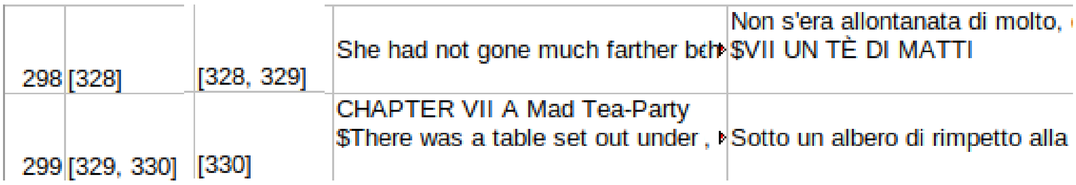


343	[374, 375]	[374, 375]	'Is that the way you manage?' $The Hatter shook his head m▸	— E tu fai cosi? — doma $Il Cappellaio scosse me

Fig. 3: A 2-to-2 alignment due to direct discourse markers and punctuation.

word, but in case of longer segments, it can lead to misalignment, such as in the aforementioned a. 31-37. The algorithm reports higher alignment cost for sections such as these.

Second, there is a general tendency to generate a 2-to-2 alignment between a short phrase with direct dialogue and a longer following sentence. This is most probably due to the presence of punctuation. However, this does not impact the alignment quality since the sentences are correctly paired (Fig. 3)



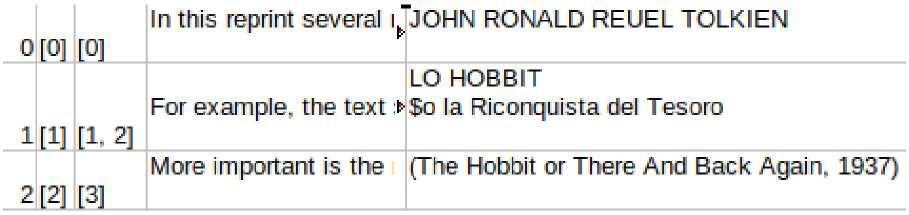
298	[328]	[328, 329]	She had not gone much farther bef▸	Non s'era allontanata di molto, $VII UN TÈ DI MATTI
299	[329, 330]	[330]	CHAPTER VII A Mad Tea-Party $There was a table set out under ,▸	Sotto un albero di rimpetto alla

Fig. 4: A misaligned chapter heading.

Third, often the chapter header is misaligned in a 1-to-2 or a 2-to-1 alignment together with the preceding or the following sentence (Fig. 4). Different choices in the typesetting of, for example, the direct discourse marker, did not impact the performance of the algorithm.

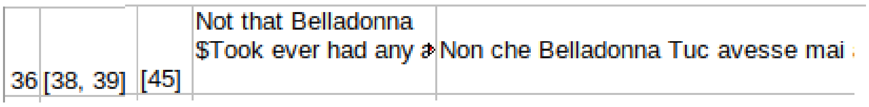
On the Hobbit corpus, 92.67% out of the 300 evaluated alignments were judged as correct. Again the first batch was the worst one, with 83/100, while the others scored 96/100 and 99/100. This corpus was slightly noisier than the Alice one, since the two Hobbit books differed in some editorial choices.

The first 10 alignments are all incorrect: the beginning of the book is completely different in the two editions, nonetheless Vecalign paired sentences in a miscellaneous assortment of 1-to-1, 1-to-2, and 1-to-many alignments



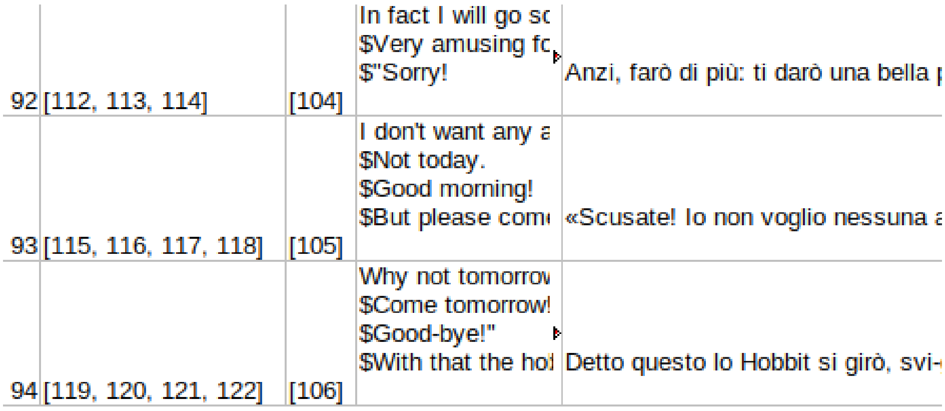
0	[0]	[0]	In this reprint several ı	JOHN RONALD REUEL TOLKIEN
1	[1]	[1, 2]	For example, the text :	LO HOBBIT $o la Riconquista del Tesoro
2	[2]	[3]	More important is the	(The Hobbit or There And Back Again, 1937)

Fig. 5: A section of the misaligned beginning of the Hobbit corpus.



36	[38, 39]	[45]	Not that Belladonna $Took ever had any a	Non che Belladonna Tuc avesse mai

Fig. 6: A split named entity: ”Belladonna Took”.



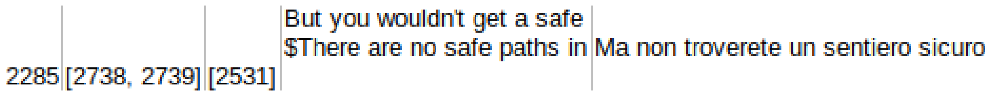
92	[112, 113, 114]	[104]	In fact I will go sc $Very amusing fc $"Sorry!	Anzi, farò di più: ti darò una bella
93	[115, 116, 117, 118]	[105]	I don't want any a $Not today. $Good morning! $But please com	«Scusate! Io non voglio nessuna a
94	[119, 120, 121, 122]	[106]	Why not tomorrow $Come tomorrow! $Good-bye!" $With that the ho	Detto questo lo Hobbit si girò, svi-

Fig. 7: An erroneous many-to-1 alignment. Only the last one is correctly aligned.

(Fig. 5) The ideal output should have been a series of blanks on both sides, alternatively.

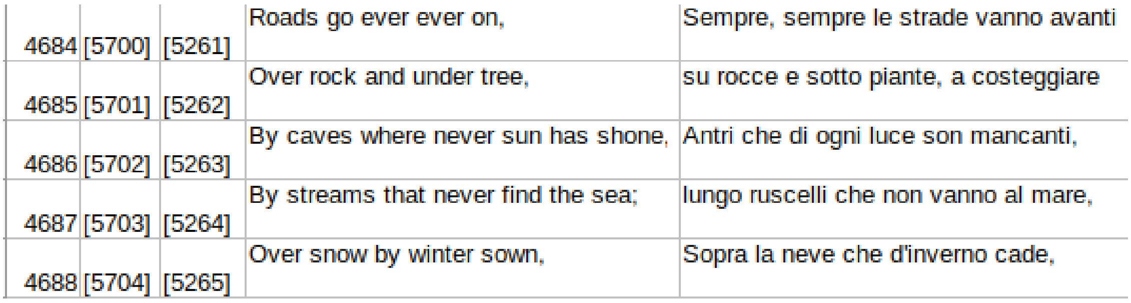
In some cases, e.g. a. 32, 36, and 37, preprocessing tricked the algorithm into creating a 2-to-1 alignment. For example, an unrecognized named entity could be split in the middle, generating a new sentence (Fig. 6). These preprocessing problems are likewise found in other sections of the text, e.g. a. 79-80, giving rise to unwanted many-to-many alignments(Fig. 7).

These problems, however, seem to be more due to differences in the tokenization model between the two languages, than due to Vecalign. Nonetheless, they are somewhat useful to this analysis, since they show that Vecalign is not totally impervious to errors when dealing with short sentences, such as in a. 92-94. In other cases, e.g. a. 4730 and 4724, the system coped well with differences in punctuation and sentence structure that influenced tokenization and sentence splitting. Moreover, the Italian version of the text contained some line break markings (”-”) inside words, but this seems not to have influenced the quality of the alignment.



2285	[2738, 2739]	[2531]	But you wouldn't get a safe $There are no safe paths in	Ma non troverete un sentiero sicuro

Fig. 8: A missing blank in the target alignment. The second sentence is not in the Italian version.



4684	[5700]	[5261]	Roads go ever ever on,	Sempre, sempre le strade vanno avanti
4685	[5701]	[5262]	Over rock and under tree,	su rocce e sotto piante, a costeggiare
4686	[5702]	[5263]	By caves where never sun has shone,	Antri che di ogni luce son mancanti,
4687	[5703]	[5264]	By streams that never find the sea;	lungo ruscelli che non vanno al mare,
4688	[5704]	[5265]	Over snow by winter sown,	Sopra la neve che d'inverno cade,

Fig. 9: A poem-like section. Most of it is correctly aligned.

Sometimes, a blank was expected, but Vecalign choose to merge the unaligned sentence with the following one. This is the case with a. 2285 (Fig. 8).

Lastly, in the Hobbit as well are found some songs that could be considered rimes or poems, both in structure and content. The a. 4684-4626 are a good example of this case: apart from the last two lines that confound the algorithm, the other are correctly aligned, unlike the first Alice rime. This could be due to the fact that in the Hobbit the poem is translated, and not adapted (Fig. 9).

4 Conclusion and Future Work

This paper described two experiments that tested and evaluated Vecaling, the state-of-the-art method for automatic sentence alignment, on two corpora of literary texts. The system was shown to perform well, even if some issues, such as not optimal handling of blank-aligned sentences and the management of short phrases and sentence boundaries, remain to be resolved.

Future work may address issues in automatic sentence alignment such as dealing with noisy or OCRed text and evaluate the impact of preprocessing, such as sentence splitting and text cleaning, on the final alignment task. Moreover, a good automatic quantitative evaluation framework should be devised to complement qualitative manual evaluation.

English-Czech and Czech-Italian alignments of the Hobbit corpus were computed, but not evaluated, and are available for future research.

Acknowledgements. The author's work is funded by Masaryk University and by Lexical Computing. The author would like to thank his supervisor doc. Mgr. Pavel Rychlý, Ph.D. for the feedback and his colleagues in the NLP lab and at Lexical Computing for the help with practical issues.

References

1. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 789–798. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1073, https://aclanthology.org/P18-1073
2. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. CoRR **abs/1812.10464** (2018), http://arxiv.org/abs/1812.10464
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
4. Brown, P.F., Lai, J.C., Mercer, R.L.: Aligning sentences in parallel corpora. In: Proceedings of the 29th Annual Meeting on Association for Computational Linguistics. p. 169–176. ACL '91, Association for Computational Linguistics, USA (1991). https://doi.org/10.3115/981344.981366, https://doi.org/10.3115/981344.981366
5. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. Computational Linguistics **19**(1), 75–102 (1993), https://aclanthology.org/J93-1004
6. Kim, Y., Graça, M., Ney, H.: When and why is unsupervised neural machine translation useless? (2020)
7. Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.J., GuzmÃ¡n, F.: Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. In: Proceedings of the Fifth Conference on Machine Translation. pp. 726–742. Association for Computational Linguistics, Online (November 2020), https://www.aclweb.org/anthology/2020.wmt-1.78
8. Meinecke, C., Wrisley, D.J., Jänicke, S.: Automated alignment of medieval text version based on word embeddings (2020). https://doi.org/https://doi.org/10.31219/osf.io/tah3y
9. Moore, R.C.: Fast and accurate sentence alignment of bilingual corpora. In: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers. pp. 135–144. Springer, Tiburon, USA (Oct 8-12 2002), https://link.springer.com/chapter/10.1007/3-540-45820-4_14
10. Palladino, C., Foradi, M., Yousef, T.: Translation alignment for historical language learning. Digital Humanities Quarterly **15**(3) (2021)
11. Papineni, K., Roukos, S., Ward, T., jing Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 311–318 (2002)
12. Pataridze, T., Kindt, B.: Text Alignment in Ancient Greek and Georgian: A Case-Study on the First Homily of Gregory of Nazianzus. Journal of Data Mining and Digital Humanities **Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages** (Jan 2018), https://hal.archives-ouvertes.fr/hal-01294591
13. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020), https://nlp.stanford.edu/pubs/qi2020stanza.pdf

14. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
15. Sennrich, R., Volk, M.: Mt-based sentence alignment for ocr-generated parallel texts. In: The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010) (11 2010). https://doi.org/10.5167/uzh-38464
16. Sennrich, R., Volk, M.: Iterative, mt-based sentence alignment of parallel texts. In: NODALIDA (2011)
17. Simard, M., Foster, G.F., Isabelle, P.: Using cognates to align sentences in bilingual corpora. In: Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. Montréal, Canada (Jun 25-27 1992), https://aclanthology.org/1992.tmi-1.7
18. Steinbach, U., Rehbein, I.: Automatic alignment and annotation projection for literary texts. In: Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. pp. 35–45. Association for Computational Linguistics, Minneapolis, USA (Jun 2019). https://doi.org/10.18653/v1/W19-2505, https://aclanthology.org/W19-2505
19. Steingrímsson, S., Loftsson, H., Way, A.: Effectively aligning and filtering parallel corpora under sparse data conditions. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 182–190. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-srw.25, https://aclanthology.org/2020.acl-srw.25
20. Thompson, B., Koehn, P.: Vecalign: Improved sentence alignment in linear time and space. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 1342–1348. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1136, https://www.aclweb.org/anthology/D19-1136
21. Tolkien, J.R.R.: The Hobbit, or There and Back Again. George Allen & Unwin (1937)
22. Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., Trón, V.: Parallel corpora for medium density languages. Recent Advances in Natural Language Processing IV pp. 247–258 (01 2007). https://doi.org/10.1075/cilt.292.32var
23. Xu, Y., Max, A., Yvon, F.: Sentence alignment for literary texts: The state-of-the-art and beyond. In: Linguistic Issues in Language Technology, Volume 12, 2015 - Literature Lifts up Computational Linguistics. CSLI Publications (Oct 2015), https://aclanthology.org/2015.lilt-12.6
24. Zhekova, D., Zangenfeind, R., Mikhaylova, A., Nikolaienko, T.: Sentence-alignment and application of russian-german multi-target parallel corpora for linguistic analysis and literary studies. MATLIT: Materialities of Literature **4**(1), 45–61 (Feb 2015). https://doi.org/10.14195/2182-8830_4-1_3, https://impactum-journals.uc.pt/matlit/article/view/2182-8830_4-1_3