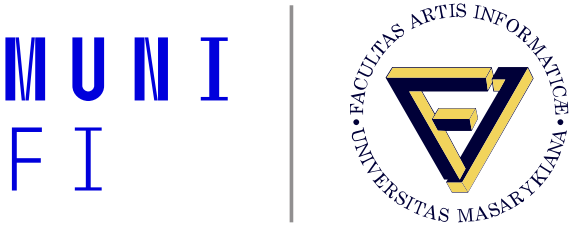

Evaluating the State-of-the-Art Sentence Alignment System on Literary Texts

Edoardo Signoroni
e.signoroni@mail.muni.cz

Faculty of Informatics, Masaryk University

December 11, 2021

Overview

Sentence alignment

- is:
  - the task of finding target sentences with the same meaning to that of the source in a bilingual corpus [5].
- useful for:
  - parallel corpora and Machine Translation (MT)
  - parallel concordances
  - translation equivalents
  - terminology extraction
  - many applications in Digital Humanities (DH, e.g. [2, 3])

Automatic Sentence Alignment

It has two parts:

- **score function** -> the likelihood that one or more target sentences are the translations of one or more source sentences
- **alignment algorithm** -> uses the scores to return a hypothesis alignment

Vecalign

Vecalign[1] [6] is the current **state-of-the-art** method.

It uses **bilingual sentence embedding similarity** (LASER, [1][2]) as its score function.

It is not being used for literary texts or in the wider area of DH.

[1] https://github.com/thompsonb/vecalign
[2] https://github.com/facebookresearch/LASER

Datasets

Two corpora:

- two versions of **"Alice's Adventures in Wonderland"**[3]
  823 sentences;
  EN, IT
- three editions of **"The Hobbit"**[9, 8, 7]
  2.200 lines avg.;
  EN, IT, CS [4]

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

[4] CS alignments were computed, but not evaluated

Datasets

	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.

Experiments

- Alice:
  1. one .txt for each language
  2. LASER and Vecaling were run with default parameters
- The Hobbit:
  1. converted to .txt
  2. split into sentences with Stanza[4] [5]
  3. LASER and Vecaling were run with default parameters

[5] LASER and Stanza gave different number of sentences.

Evaluation

Qualitative evaluation:

- 300 alignments, 100 from the beginning, the middle, and the end of each EN-IT corpus
- 1pt for a good alignment; 0pt for a bad alignment

Results

	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.

Manual analysis

Some interesting examples found during the analysis:[6]

[6] "$" was added after the Vecalign run to mark a new sentence in the dataframe.

Popular rimes

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 (

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

Popular rimes

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...

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

Dialogue

343	[374, 375]	[374, 375]	'Is that the way you manage?' Alice asked. $The Hatter shook his head mournfully. 'Not	— E tu fai così? — domandò Alice. $Il Cappellaio scosse mestamente la

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

Headings

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

Figure 4: A misaligned chapter heading.

Incipit

0	[0]	[0]	In this reprint several	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)

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

Named Entities

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

Figure 6: A split named entity: "Belladonna Took".

Many-to-1

In fact I will go so far as to send you on this adventure. $Very amusing for me, very good for you and profitable $"Sorry!	Anzi, farò di più: ti darò una bella parte in quest'avventura, mo
I don't want any adventures, thank you. $Not today. $Good morning! $But please come to tea - any time you like!	«Scusate! Io non voglio nessuna avventura, grazie! Non oggi!
Why not tomorrow? $Come tomorrow! $Good-bye!" $With that the hobbit turned and scuttled inside his roun	Detto questo lo Hobbit si girò, svi-gnandosela per la verde por

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

Missing blank

But you wouldn't get a safe path even then.	
$There are no safe paths in this part of the world.	Ma non troverete un sentiero sicuro nemmeno in questo caso.

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

Poems

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,

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

Conclusion

After this evaluation, we can say that:

- Vecaling works well for literary texts,
- but some issues remain in the handling of:
  - blanks;
  - short phrases;
  - sentence boundaries

Some of these problems may be caused by preprocessing (e.g. split NE and many-to-1 alignments)

Future Work

- Evaluate the performance on noisy or OCRed text;
- evaluate the impact of preprocessing;
- devise an automated evaluation.

Thank you!

References I

[1] Mikel Artetxe and Holger Schwenk. “Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond”. In: *CoRR* abs/1812.10464 (2018). arXiv: 1812.10464. URL: http://arxiv.org/abs/1812.10464.

[2] Christofer Meinecke, David Josef Wrisley, and Stefan Jänicke. “Automated Alignment of Medieval Text Version based on Word Embeddings”. In: (2020). DOI: https://doi.org/10.31219/osf.io/tah3y.

[3] Chiara Palladino, Maryam Foradi, and Tariq Yousef. “Translation Alignment for Historical Language Learning”. In: *Digital Humanities Quarterly* 15.3 (2021).

References II

[4] Peng Qi et al. “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. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[5] Steinþór Steingrımsson, Hrafn Loftsson, and Andy Way. “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*. Online: Association for Computational Linguistics, July 2020, pp. 182–190. DOI: 10.18653/v1/2020.acl-srw.25. URL: https://aclanthology.org/2020.acl-srw.25.

References III

[6] Brian Thompson and Philipp Koehn. “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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1342–1348. DOI: 10.18653/v1/D19-1136. URL: https://aclanthology.org/D19-1136.

[7] (transl. František Vrba) Tolkien John Ronald Reuel. *Hobit aneb Cesta tam a zase zpátky*.

[8] John Ronald Reuel Tolkien. *Lo Hobbit, o la Riconquista del Tesoro*. Bompiani, 2012.

References IV

[9] John Ronald Reuel Tolkien. *The Hobbit, or There and Back Again*. HarperCollins, 1995.

MUNI
FACULTY
OF INFORMATICS