An Automatic Model and Gold Standard for Translation Alignment of Ancient Greek

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis

The paper illustrates a workflow for developing and evaluating automatic translation alignment models for Ancient Greek. We designed an annotation Style Guide and a gold standard for the alignment of Ancient Greek-English and Ancient Greek-Portuguese, measured inter-annotator agreement and used the resulting dataset to evaluate the performance of various alignment models.

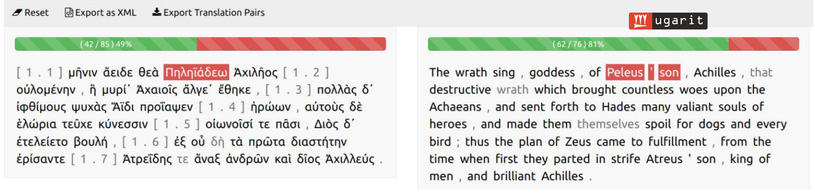
ALIGNMENT GUIDELINES & GOLD STANDARD

The datasets for gold standard include:

- The Iliad
- Plato's Crito
- Xenophon's Cyropaedia

The chosen datasets provide sufficient diversity of:

- Language (Homeric to Koine Greek)
- Text genre (poetry, prose, and dialogue)
- Style of translation



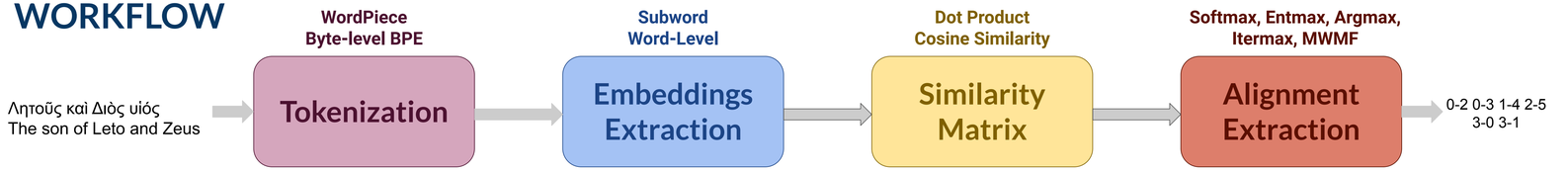
Both the Gold Standard and the Guidelines for Ancient Greek-English were created by two experts who had previously worked with UGARIT.

The Guidelines for Portuguese were created in a similar fashion: they designed by two domain experts, who also manually aligned the corpus

	GRC-ENG	GRC-POR
Sentences	275	183
GRC Tokens	5.359	3.216
GRC Types	2.347	1.587
ENG/POR Tokens	7.515	3.71
ENG/POR Types	1.634	1.355
Sure Alignments	6.24	3.028
Possible Alignments	1.423	864
IAA	86.17%	83.31%

An overview of the gold standards datasets and their Inter-Annotator Agreement

ALIGNMENT WORKFLOW



Experiment	Input Model	Training Objectives	Languages	Data Size	Source
EX1	mBERT, XLM-R	MLM, SO, TLM, PSI	GRC-ENG	32.5k par. sentences	Perseus
EX2	EX1 fine-tuned		GRC-LAT	8k par. sentences	DFHG
EX3	EX2 fine-tuned	MLM	GRC Monolingual	12 Millions Tokens	Perseus, First1kGreek TreeBanking
EX4	EX2 fine-tuned	MLM, SO, TLM, PSI	GRC-ENG GRC-LAT GRC-KAT	45k par. sentences	Perseus, DFHG, UGARIT
EX5	EX3 fine-tuned				
EX6	EX5 fine-tuned	SO	Mixed Dataset	2.2k par. sentences	UGARIT

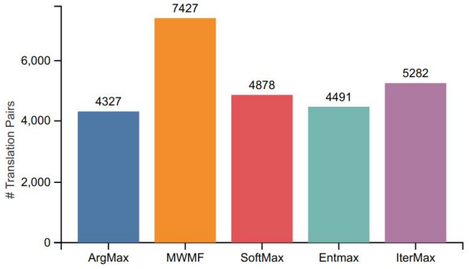
An overview of the conducted experiments

		Precision	Recall	F1	AER
Baseline	Giza++	25.59%	24.60%	25.09%	74.88%
	fast align	25.62%	30.14%	27.70%	72.47%
	Eflomal	34.84%	35.59%	35.21%	64.81%

		mBERT				XLM-R			
		Precision	Recall	F1	AER	Precision	Recall	F1	AER
Zero-Shot	Softmax	30.08%	26.66%	28.27%	71.66%	21.68%	13.53%	16.66%	83.16%
Ex6	Softmax	63.84%	61.27%	**62.53%**	37.40%	76.11%	75.61%	75.86%	24.13%
	Entmax	65.49%	57.41%	61.18%	38.61%	**77.45%**	72.69%	74.99%	24.89%
	Match	50.00%	**72.61%**	59.22%	41.50%	58.79%	**86.17%**	69.89%	31.01%
	Argmax	**66.01%**	54.92%	59.96%	39.76%	77.25%	71.10%	74.05%	25.81%
	Itermax	59.67%	64.06%	61.79%	**38.35%**	72.22%	81.02%	**76.37%**	**23.91%**

Evaluation results on Ancient Greek-Portuguese gold standard.

A comparison among the alignment extraction approaches regarding the number of translation pairs they produce (XLM-R, EX6)



		Precision	Recall	F1	AER
Baseline	Giza++	37.25%	29.26%	32.78%	67.01%
	FastAlign	37.37%	35.64%	36.48%	63.47%
	Eflomal	47.17%	42.93%	44.95%	54.95%

		mBERT				XLM-R			
		Precision	Recall	F1	AER	Precision	Recall	F1	AER
Zero-Shot	Softmax	37.14%	21.09%	26.90%	72.70%	37.59%	11.84%	18.01%	81.80%
Ex1	Softmax	52.98%	38.21%	44.40%	55.28%	54.61%	28.21%	37.20%	62.46%
Ex2	Softmax	55.89%	40.24%	46.79%	52.86%	55.62%	28.97%	38.10%	61.56%
Ex3	Softmax	54.03%	39.68%	45.76%	53.94%	53.58%	24.21%	33.35%	66.33%
Ex4	Softmax	65.06%	48.08%	55.30%	44.33%	65.22%	36.39%	46.72%	52.88%
Ex5	Match	58.09%	65.03%	61.36%	38.81%	62.46%	72.70%	67.19%	33.05%
	Itermax	69.64%	58.35%	63.50%	36.25%	77.07%	58.28%	66.37%	33.32%
Ex6	Softmax	80.80%	56.91%	66.78%	32.72%	90.73%	67.91%	77.68%	21.89%
	Entmax15	83.86%	53.76%	65.52%	33.93%	92.61%	64.18%	75.82%	23.69%
	Match	65.42%	**72.76%**	68.90%	31.31%	77.85%	**85.50%**	**81.50%**	**18.72%**
	Argmax	**84.95%**	52.47%	64.87%	34.57%	**93.44%**	62.57%	74.95%	24.54%
	Itermax	78.43%	64.08%	**70.53%**	**29.14%**	89.66%	72.05%	79.90%	19.73%

Evaluation results on Ancient Greek-English gold standard.

LREC 2022 Marseille

GitHub