Ollscoil na Gaillimhe
University of Galway

Insight

Findings of the SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages

Oksana Dereza, Adrian Doyle, Priya Rani,
Atul Kr. Ojha, Pádraic Moran, John P. McCrae

21-22 March 2024, St. Julians, Malta
The 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP

Overview

Ollscoil na Gaillimhe
University of Galway

- 3 problems & 13 languages in the constrained track
- 5 problems & 16 languages in the unconstrained track
- 14 registrations, 3 submissions to each track
- 2 successful submissions to each track → 4 system description papers

- GitHub: https://github.com/sigtyp/ST2024
- List of text sources: ./ST2024/blob/main/list_of_text_sources.md
- Constrained track: https://codalab.lisn.upsaclay.fr/competitions/16822
- Unconstrained track: https://codalab.lisn.upsaclay.fr/competitions/16818
- Updated dataset (with test labels): https://doi.org/10.5281/zenodo.10655061

Ollscoil na Gaillimhe
University of Galway

Subtasks & Data

University *of* Galway.ie

Subtasks

Ollscoil na Gaillimhe
University of Galway

A. Constrained

1. POS-tagging
2. Morphological feature prediction
3. Lemmatisation

B. Unconstrained

1. POS-tagging
2. Morphological feature prediction
3. Lemmatisation
4. Filling the gaps (mask filling)
   - Word-level
   - Character-level

Data Sources

Ollscoil na Gaillimhe
University of Galway

St Gall Priscian Glosses, Würzburg Glosses

- Old Irish

CELT

- Old Irish
- Middle Irish
- Early Modern Irish

Historical Irish Corpus

- Early Modern Irish

MGTSZ

- Old Hungarian

Universal Dependencies v. 2.12

- Ancient Greek
- Ancient Hebrew
- Classical Chinese
- Coptic Gothic
- Historical Icelandic
- Latin
- Old Church Slavonic
- Old East Slavic
- Old French
- Vedic Sanskrit

Language	Code	Script	Dating	Train-T	Valid-T	Test-T	Train-S	Valid-S	Test-S
Ancient Greek ♣	grc	Greek	800 BCE – 110 CE	334,043	41,905	41,046	24,800	3,100	3,101
Ancient Hebrew ♢	hbo	Hebrew	900 – 999 CE	40,244	4,862	4,801	1,263	158	158
Classical Chinese ♠	lzh	Hanzi	47 – 220 CE	346,778	43,067	43,323	68,991	8,624	8,624
Coptic ♢	cop	Coptic	0 – 199 CE	57,493	7,282	7,558	1,730	216	217
Gothic ♣	got	Latin	400 – 799 CE	44,044	5,724	5,568	4,320	540	541
Medieval Icelandic ♣	isl	Latin	1150 – 1680 CE	473,478	59,002	58,242	21,820	2,728	2,728
Classical & Late Latin ♣	lat	Latin	100 BCE – 399 CE	188,149	23,279	23,344	16,769	2,096	2,097
Medieval Latin ♣	latm	Latin	774 – early 1300s CE	599,255	75,079	74,351	30,176	3,772	3,773
Old Church Slavonic ♣	chu	Cyrillic	900 – 1099 CE	159,368	19,779	19,696	18,102	2,263	2,263
Old East Slavic ♣	orv	Cyrillic	1025 – 1700 CE	250,833	31,078	32,318	24,788	3,098	3,099
Old French ♣	fro	Latin	1180 CE	38,460	4,764	4,870	3,113	389	390
Vedic Sanskrit ♣	san	Latin (transcr.)	1500 – 600 BCE	21,786	2,729	2,602	3,197	400	400
Old Hungarian ♡	ohu	Latin	1440 – 1521 CE	129,454	16,138	16,116	21,346	2,668	2,669
Old Irish ♣	sga	Latin	600 – 900 CE	88,774	11,093	11,048	8,748	1,093	1,094
Middle Irish ♣	mga	Latin	900 – 1200 CE	251,684	31,748	31,292	14,308	1,789	1,789
Early Modern Irish ♣	ghc	Latin	1200 – 1700 CE	673,449	115,163	79,600	24,440	3,055	3,056

Ollscoil na Gaillimhe
University of Galway

Evaluation & Results

University *of* Galway.ie

Evaluation

Ollscoil na Gaillimhe
University of Galway

Task	Metrics
POS-tagging	Accuracy @1, F1
Morphological feature prediction	Macro-average of Accuracy @1 per tag
Lemmatisation	Accuracy @1, Accuracy @3
Filling the gaps (word-level)	Accuracy @1, Accuracy @3
Filling the gaps (character-level)	Accuracy @1, Accuracy @3

- Final score = macro-average of per-task scores
- Each task is weighed equally

- Multiple metrics for every task, except morphological feature prediction
- Unweighted average of the metrics as the score for the task

- Morphological annotation: punishment for predicting non-existent features

Baselines

Ollscoil na Gaillimhe
University of Galway

- Provided for problems 1-3
- Simple multilayer perceptron architecture with no pretraining
- Individual models for each language for POS-tagging and lemmatisation
- Language-agnostic models for each of the 44 morphological features
- 16 features extracted for each token + language codes
- Hyperparameters common for all models:
    - Hidden layers: 2
    - Activation: ReLU
    - Dropout: 20%
    - Optimiser: Adam
    - Early stopping

Submissions: constrained

Ollscoil na Gaillimhe
University of Galway

Heidelberg-Boston

- Constrained track winner
- Results on par with the winner of the unconstrained track
- POS-tagging & morphological feature prediction: hierarchical tokenisation + DeBERTa-V3
- Lemmatisation: character-level nanoT5
- Models pre-trained from scratch for each language in the dataset individually
- Avg. 95.25%, 93.67% and 96.18% for POS-tagging, lemmatisation & morph. features

Team 21a

- Multilingual model pretrained on the whole dataset, LiBERTus
- Follows RoBERTa architecture
- Loses to Heidelberg-Boston despite using cross-lingual transfer
- Struggles with multiword expressions
- Avg. 82.47%, 81.98% and 90.70% for POS-tagging, lemmatisation & morphological feature prediction

Submissions: unconstrained

Ollscoil na Gaillimhe
University of Galway

UDParse

- Problems 1-3: UDParse paser on top of different pre-trained transformer models (mBERT, XLM-RoBERTa, GPT2, heBERT, slavicBERT)
- Word-level mask filling: distilBERT
- Character-level mask filling: embeddingless n-gram model
- Avg. 96.09%, 86.47% and 96.68% for POS-tagging, lemmatisation & morph. features
- Avg. 3.77% and 55.62% for word-level & character-level mask filling

TartuNLP

- Fine-tuned stacked language- and task-specific adapters for XLM-RoBERTa
- Applied the same approach uniformly to all 5 tasks and 16 languages
- Avg. 85.67% and 88.14% for POS-tagging & morph. feature prediction
- Avg. 94.88% for lemmatisation
- Avg. 5.95% and 48.38% for word-level & character-level mask filling

Problems 1-3

Ollscoil na Gaillimhe
University of Galway

		AVG	chu	cop	fro	got	grc	hbo	isl	lat	latm	lzh	ohu	orv	san
POS-tagging															
Baseline		92.76	93.36	94.98	91.57	93.73	90.33	94.07	94.00	92.39	97.22	90.91	93.59	90.33	89.37
Constrained	HDB-BOS	**95.25**	**96.57**	**96.92**	**93.10**	**95.41**	**96.39**	**96.68**	**96.08**	**95.54**	**98.43**	**92.92**	**95.98**	**94.46**	**89.71**
	Team 21a	82.47	94.62	42.65	85.14	93.48	93.49	27.26	93.85	92.43	94.41	81.79	94.42	91.23	87.32
Unconstrained	UDParse	**96.09**	**97.00**	**97.33**	**96.01**	**96.47**	**96.49**	**97.84**	**96.88**	**96.83**	**98.79**	**93.76**	**96.71**	**94.99**	**90.02**
	TartuNLP	85.67	66.35	60.99	94.51	92.72	95.72	94.15	96.67	95.86	**98.79**	83.28	75.14	75.67	83.83
Lemmatisation															
Baseline		91.95	89.60	95.74	91.93	91.95	91.06	95.28	93.78	92.08	97.03	98.81	89.43	84.44	84.24
Constrained	HDB-BOS	**93.67**	**94.49**	**95.07**	**92.63**	**93.31**	**94.08**	**97.29**	**96.63**	**96.00**	**98.46**	99.18	**85.92**	**90.09**	**84.59**
	Team 21a	81.98	79.59	46.32	83.32	90.79	88.30	61.75	94.58	92.35	97.22	**99.84**	69.97	78.44	83.21
Unconstrained	UDParse	86.47	59.56	74.78	92.47	92.81	**94.02**	96.85	**97.96**	96.74	**98.91**	**99.96**	63.43	68.55	88.10
	TartuNLP	**94.88**	**92.70**	**98.28**	**95.11**	**95.41**	93.39	**98.15**	97.23	**96.99**	98.69	99.91	**86.91**	**89.23**	**91.48**
Morphological feature prediction															
Baseline		33.32	85.07	47.41	28.27	18.95	25.10	42.78	35.83	18.17	30.94	43.58	23.20	25.55	08.34
Constrained	HDB-BOS	**96.18**	**96.04**	**98.60**	**97.87**	**95.32**	**97.46**	**97.46**	**95.29**	**95.17**	**98.68**	**95.52**	**96.30**	**95.00**	**91.58**
	Team 21a	90.70	94.06	80.47	94.08	93.96	96.50	71.20	94.79	93.31	97.98	85.98	94.64	92.16	90.00
Unconstrained	UDParse	**96.68**	**96.49**	**98.88**	**98.33**	**96.23**	**97.78**	**97.05**	**95.92**	**96.66**	**98.83**	**96.24**	**96.62**	**95.16**	**92.60**
	TartuNLP	88.14	67.14	74.86	98.01	92.40	97.33	95.14	95.53	95.91	**98.83**	88.75	75.62	80.00	86.33

Problem 4

Ollscoil na Gaillimhe
University of Galway

	AVG	chu	cop	fro	got	grc	hbo	isl	lat	latm	lzh	ohu	orv	san	sga	mga	ghc
Mask filling: word-level																	
UDParse	3.77	**2.80**	0.00	3.28	2.67	**3.07**	**5.39**	3.42	3.51	4.73	6.10	**6.31**	5.03	3.86	2.79	**4.03**	3.29
TartuNLP	**5.95**	2.42	**1.87**	**7.22**	**3.40**	3.01	0.00	**16.90**	**11.45**	**14.39**	**10.46**	0.06	**6.05**	**4.79**	**3.21**	3.99	**6.00**
Mask filling: character-level																	
UDParse	**55.62**	**66.77**	0.00	**62.77**	**74.59**	**68.46**	**36.85**	**66.45**	**67.91**	**72.93**	0.00	**66.52**	**66.77**	**70.10**	**58.38**	**53.38**	**58.09**
TartuNLP	48.38	53.79	**45.10**	52.46	67.34	61.15	18.56	57.32	65.79	69.84	**0.25**	45.65	48.04	64.52	34.86	39.49	49.88

Overall results

		AVG	chu	cop	fro	got	grc	hbo	isl	lat	latm	lzh	ohu	orv	san	sga	mga	ghc
	Baseline	72.68	89.35	79.38	70.59	68.21	68.83	77.38	74.54	67.55	75.07	77.77	68.74	66.77	60.65	–	–	–
Constrained	HDB-BOS	**95.02**	**95.70**	**96.65**	**94.54**	**94.68**	**95.98**	**97.14**	**96.00**	**95.57**	**98.53**	**95.88**	**92.73**	**93.18**	**88.62**	–	–	–
	Team 21a	85.05	89.42	56.48	87.51	92.74	92.76	53.41	94.41	92.69	96.54	89.21	86.34	87.28	86.84	–	–	–
Unconstrained	UDParse	**61.93**	**71.15**	**58.90**	**71.10**	**73.07**	**71.84**	**67.05**	71.98	72.38	74.79	**59.20**	**70.61**	**69.15**	**69.61**	**30.59**	**28.71**	**30.69**
	TartuNLP	55.74	49.85	51.52	68.93	69.74	70.25	60.94	**72.88**	**73.15**	**76.15**	56.54	51.98	55.66	65.51	19.03	21.74	27.94

Discussion

Ollscoil na Gaillimhe
University of Galway

- All participants used transformer architectures, with RoBERTa and its modifications being the most popular one
- All participants outperformed the baselines for morphological feature prediction with the best average result about 96% across 13 languages
- Only the winning teams beat the baselines for POS-tagging and lemmatisation, achieving average results of 95.25% and 93.67% respectively in the constrained setting, and 96.09% and 94.88% in the unconstrained setting
- In word-level mask filling, the best average result over 16 languages was 5.95% and the best result for an individual language was 16.9% for Medieval Icelandic
- At the character level, the best average result over 16 languages was 55.62% and the best result for an individual language was 74.59% for Gothic
- The difference between the smallest (Vedic Sanskrit, 21K) and largest (Medieval Latin, 599K) corpora is not that dramatic: e.g. avg. 9.5% for POS-tagging and avg. 11.3% for morph. feature prediction

Discussion

Ollscoil na Gaillimhe
University of Galway

- Cross-lingual and cross-temporal transfer could have played an important role in the systems that used XLM-RoBERTa
- Similar results can be achieved with pre-training on modestly sized monolingual data without any transfer
- A simple character n-gram model can be more effective for mask filling in a low-resource setting than transformers
- Generally, mask filling tasks are much harder than others, especially on word level
    - High lexical variety
    - Orthographic variation
    - Relatively short sentences
    - Code-switching
    - Data scarcity
    - Composite characters and vowel markings in Coptic & Ancient Hebrew
    - Non-trivial character decomposition in Classical Chinese

Ollscoil na Gaillimhe
University of Galway

Thank you for your attention!

oksana.dereza@insight-centre.org

University
*of*Galway.ie