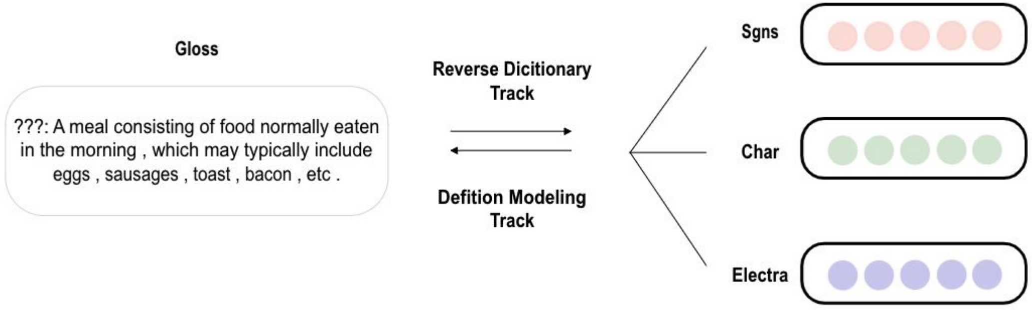
1Cademy at Semeval-2022 Task 1: Investigating the Effectiveness of Multilingual, Multitask, and Language-Agnostic Tricks for the Reverse Dictionary Task

Zhiyong Wang*, Ge Zhang*, Nineli Lashkarashvili

Task Description

- Three embeddings: skip-gram embeddings, character-based embeddings, and transformer-based contextualized embeddings.
- Five languages: English, Spanish, French, Italian, and Russian
- Heads of glosses are hidden



Results and Findings

1. Do all architectures yield comparable results?

2. What are the effects of Re-Tokenization?

3. What are the effects of combining different languages as inputs?

4. What are the effects of handling all tracks together?

Different Architectures

- Transformer-based model is hard to tune with small amounts of data

- ELMo-based models achieve best results

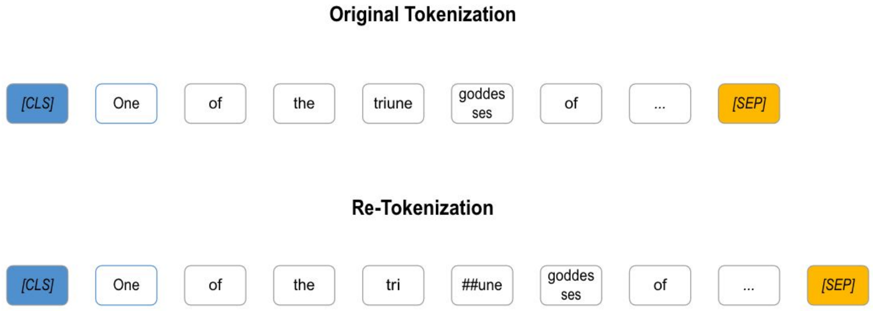
Word Representations	SGNS			Char			Electra		
Monolingual Models	MSE	COS	RANK	MSE	COS	RANK	MSE	COS	RANK
RNN+WordPiece	1.000	0.249	0.310	0.158	0.778	0.442	**1.454**	**0.832**	0.433
LSTM+WordPiece	0.990	0.228	0.375	0.148	0.791	0.458	1.491	0.831	0.449
Transformer+WordPiece	1.042	0.214	0.367	0.194	0.780	0.453	1.796	0.827	0.486
BiRNN+WordPiece	0.989	0.221	0.395	0.150	0.791	0.454	1.483	0.832	0.449
Elmo+WordPiece	1.041	0.252	0.282	0.161	0.772	0.430	1.512	0.829	0.434
Elmo+BPE	1.037	0.250	**0.250**	0.162	0.774	0.443	1.537	0.822	0.436
Elmo+ULM	1.022	**0.265**	0.259	0.157	0.781	**0.430**	1.525	0.829	0.432
Elmo+WordPiece+DWA	**0.985**	0.246	0.298	**0.142**	**0.799**	0.447	1.514	0.827	**0.428**

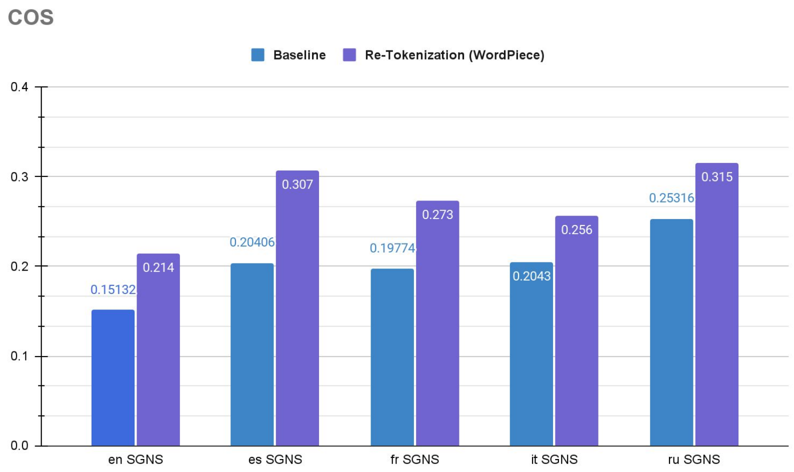
Table 1: Experiment results on English resource test data using the monolingual models. Check section 2 for word algorithm representations' abbreviation. Check section 3 for details of monolingual models.

Re-Tokenization

- Three widely-used tokenization algorithm: BPE, WordPiece, and Unigram:

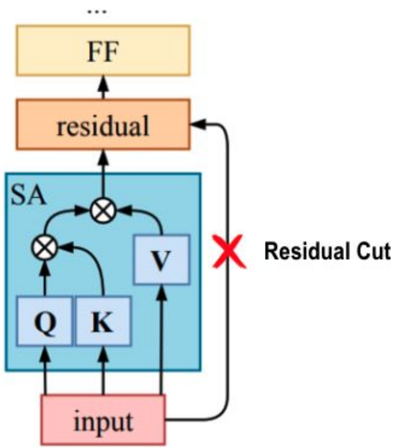
- The re-tokenization can improve model performance on this task significantly.





Multilingual Models

1. Adding Language Token
2. Residual Cut (Liu et al.,2020)



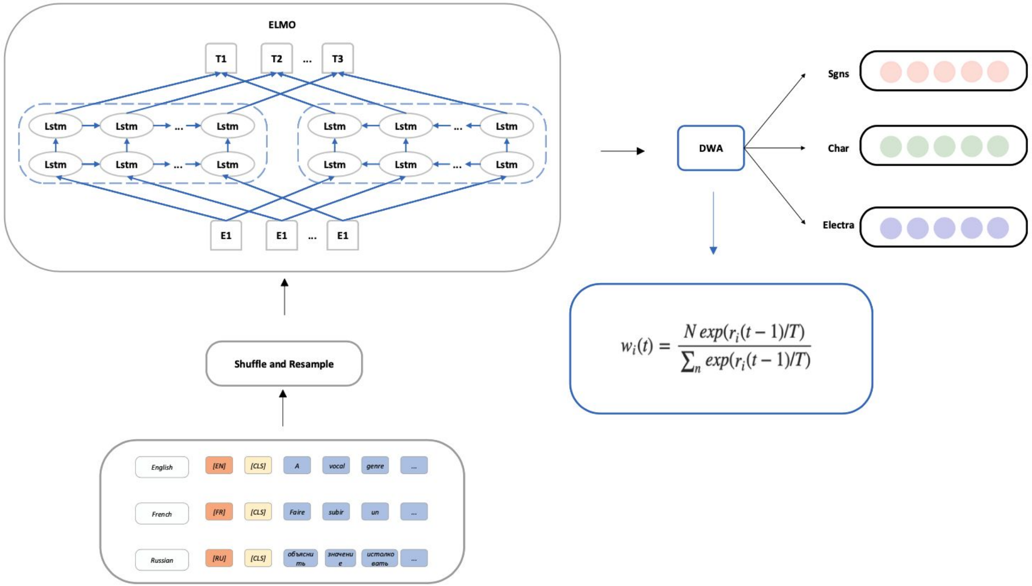
Liu, D., Niehues, J., Cross, J., Guzmán, F., & Li, X. (2020). Improving zero-shot translation by disentangling positional information. arXiv preprint arXiv:2012.15127.

Multilingual Models

Languages	EN			ES			FR			IT			RU		
Multilingual Models	MSE	COS	RANK	MSE	COS	RANK	MSE	COS	RANK	MSE	COS	RANK	MSE	COS	RANK
Transformer	1.023	0.201	0.400	0.977	0.300	**0.310**	1.051	0.278	**0.338**	1.143	**0.280**	**0.340**	0.564	0.318	0.363
Transformer+RC	1.029	0.199	0.417	1.005	0.298	0.329	1.069	0.253	0.374	1.189	0.267	0.364	0.601	0.279	0.409
Transformer+ALT	1.043	**0.215**	**0.397**	1.014	**0.308**	0.310	1.103	**0.280**	0.350	1.158	0.276	0.341	0.603	**0.326**	**0.337**
Transformer+RC+ALT	**1.011**	0.159	0.500	**0.955**	0.266	0.422	**1.044**	0.271	0.360	**1.129**	0.264	0.376	**0.561**	0.308	0.371

Table 2: Experiment results on **SGNS** word representation using the multilingual Transformer-based models. Check section 3 for details of multilingual models. **RC** represents the Residual Cutting trick. **ALT** represents the Adding Language Token trick.

Multitask Model



Word Representations	SGNS			Char			Electra		
Multilingual Models	MSE	COS	RANK	MSE	COS	RANK	MSE	COS	RANK
Elmo_EN	1.023	0.238	0.317	0.177	0.759	**0.447**	1.555	0.818	**0.440**
Elmo+ALT_EN	**1.014**	**0.246**	**0.300**	**0.164**	**0.762**	0.449	**1.540**	**0.825**	0.441
Elmo_ES	**0.953**	0.342	**0.234**	0.532	0.810	0.405	NA	NA	NA
Elmo+ALT_ES	0.960	**0.351**	0.235	**0.511**	**0.822**	**0.393**	NA	NA	NA
Elmo_IT	**1.094**	0.343	**0.218**	0.355	0.720	0.403	NA	NA	NA
Elmo+ALT_IT	1.106	**0.343**	0.214	**0.354**	**0.735**	**0.387**	NA	NA	NA
Elmo_FR	**1.001**	0.313	0.255	0.388	0.752	**0.411**	1.298	0.845	0.445
Elmo+ALT_FR	1.004	**0.321**	**0.246**	**0.387**	**0.757**	0.411	**1.228**	**0.859**	**0.439**
Elmo_RU	**0.547**	0.357	0.247	0.145	0.816	**0.398**	0.891	**0.729**	0.386
ELmo+ALT_RU	0.563	**0.368**	**0.232**	**0.137**	**0.828**	0.400	**0.887**	0.728	**0.384**

Table 3: Experiment results of the multilingual ELmo-based models. **ALT** represents the Adding Language Token trick.

- The Dynamic Weight Averaging (DWA) can keep different tasks converging at the same speed.
- The ELMo-based multitask model with DWA is as competitive as ELMo-based single task model across different languages.

Liu, S., Johns, E., & Davison, A. J. (2019). End-to-end multi-task learning with attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1871-1880).