Multilingual NLP
Lab 4 – Comparing the grammatical ability of monolingual and multilingual language models

AN Ji

December 25, 2024

🔗 Code on Colab

1 Use two multilingual models (e.g. mBERT and mT5) and monolingual models, compute the probability of generating each sentence.

For each of the 35 combinations made by 5 languages and 7 syntactic relationships, We sample the first 100 sentence pairs as our evaluation set. The total number of sentences is 7,000.

We use the following models for our experiment.

Multilingual model for all languages: `mBERT`

Monolingual models:

- English: `FacebookAI/roberta-base`
- French: `almanach/camembert-base`
- German: `dbmdz/bert-base-german-cased`
- Hebrew: `dicta-il/dictabert`
- Russian: `DeepPavlov/rubert-base-cased`

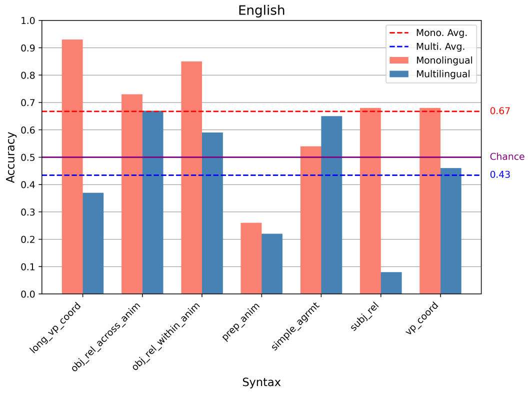
2 Assess whether the grammatical sentence gets a higher probability than the agrammatical sentence.

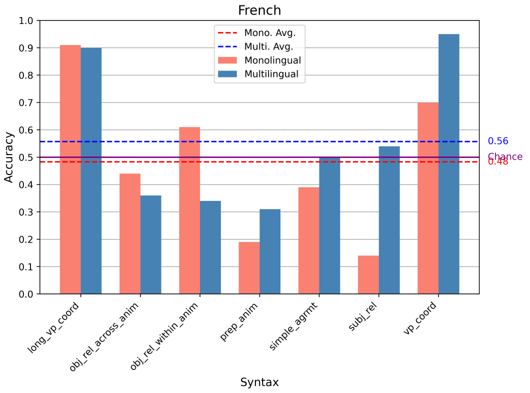
For each combination, we compute the accuracies of both models by counting the number of pairs where grammatical sentence is assigned a higher probability than its agrammatical counterpart. The raw results are presented in Table 3.

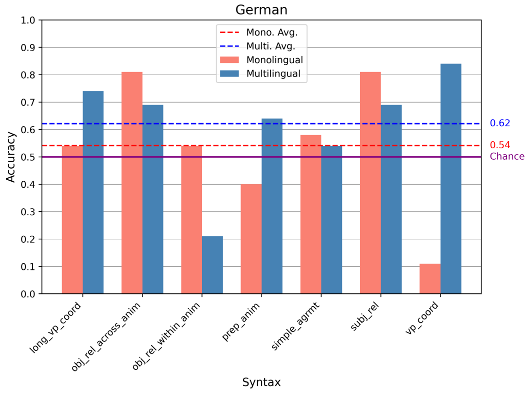
3 Conclude

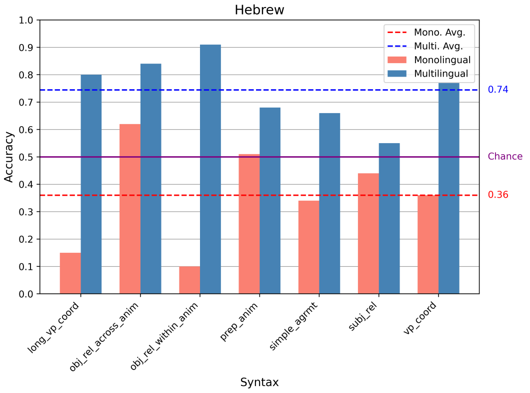
3.1 For each language, plot the results taking as an example the plot in Figure 1. What do these plots show?

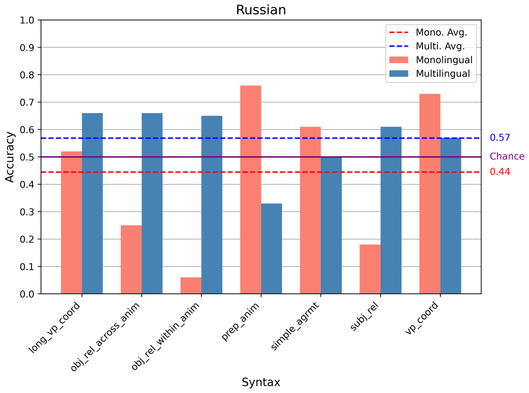
Before all, we should point out that the inferred probabilities may fluctuate from one inference to another, and might be affected by a couple of factors, such as the order of sentences passed in models.











- **English**: `RoBERTa` surpasses 0.5 and outperforms `mBERT` in every syntactic aspect except `simple agreement` (e.g. *the author laughs* vs. **the author laugh*). `mBERT` does not perform well with only 3 groups higher than 0.5. It's actually the only language where monolingual model defeats the multilingual counterpart on grammatical sensibility.

- **French**: `mBERT` reaches 0.5 on 4 groups (`long_vp_coord`, `vp_coord`, `simple_agrmt` and `subj_rel`), while `camemBERT` is not so competitive on the right-hand 4 groups. The overall performance difference is one of the smallest (0.08, the other one is German).

- **German**: Similar to French, the two models are better at complementary groups and it's relatively harder to tell which is better, and the distribution is relatively dispersed and unstable across all groups.

- **Hebrew**: The performance of `mBERT` is obviously more evenly distributed across all syntactic groups in Hebrew. It also yields the best overall performance (0.74) among all languages, while `DictaBERT` struggles a lot in judging sentence grammaticality.

- **Russian**: `mBERT` slightly outperform `RuBERT`, the monolingual counterpart, after German and French. It also performs more evenly, while `RuBERT` seems to be much weaker in the same task, similar to Hebrew model `DictaBERT`.

3.2 For each language, compute the morphological complexity using the C_{WALS} metric introduced in Bentz et al. [1]. How does this metric correlate with the grammatical scores? You can, for instance, compute the Spearman correlation coefficient.

Given Spearman correlation coefficients computed based on our model performances, we notice a strong positive correlation (+0.700) between C_{WALS} and multilingual model's overall grammatical score, as well as an even much stronger negative correlation (-1.000) between C_{WALS} and monolingual model's overall performance. The overall grammatical scores for each language and the C_{WALS} are shown in Table 1.

3.3 Another way of calculating the morphological complexity of each language is to look at the number of tokens in each corpus. Do you observe a link between the number of tokens and performance?

We use `mBERT` tokenizer to tokenize the sentences of our sample for crosslingual consistency, although this choice might bring some kind of tiny bias since we also use `mBERT` as one of our models. The token counts are presented in Table 2 together with grammatical scores of both kinds of models.

We also compute the Spearman score to quantify their correlation. The results indicate that the total number of tokens, as another metric representing morphological complexity of languages, has a very strong positive correlation (+1.000) with multilingual model's overall performance, as well as a strong negative correlation (-0.700) with monolingual model's overall performance.

Note that the results converges with those of C_{WALS} and this could, to some extent, justify the conclusion of Bentz et al. [1].

language	multi_accuracy	mono_accuracy	c_wals
en	0.434286	0.667143	0.329252
de	0.621429	0.541429	0.397002
fr	0.557143	0.482857	0.434112
ru	0.568571	0.444286	0.453401
he	0.744286	0.360000	0.529145

Table 1: Overall performances of multilingual and monolingual models for all languages alongside their C_{WALS}. Languages are reorganized in ascending order according to morphological complexity represented by their C_{WALS}.

language	multi_accuracy	mono_accuracy	num_tokens
en	0.434286	0.667143	12002
de	0.621429	0.541429	16032
fr	0.557143	0.482857	14729
ru	0.568571	0.444286	15158
he	0.744286	0.360000	17026

Table 2: Overall performances of multilingual and monolingual models for all languages alongside their number of tokens in our corpus. Tokens are based on `mBERT` tokenizer

.

Appendix

	language	syntax	multi_accuracy	mono_accuracy
0	de	long_vp_coord	0.74	0.54
1	de	obj_rel_across_anim	0.69	0.81
2	de	obj_rel_within_anim	0.21	0.54
3	de	prep_anim	0.64	0.40
4	de	simple_agrmt	0.54	0.58
5	de	subj_rel	0.69	0.81
6	de	vp_coord	0.84	0.11
7	en	long_vp_coord	0.37	0.93
8	en	obj_rel_across_anim	0.67	0.73
9	en	obj_rel_within_anim	0.59	0.85
10	en	prep_anim	0.22	0.26
11	en	simple_agrmt	0.65	0.54
12	en	subj_rel	0.08	0.68
13	en	vp_coord	0.46	0.68
14	fr	long_vp_coord	0.90	0.91
15	fr	obj_rel_across_anim	0.36	0.44
16	fr	obj_rel_within_anim	0.34	0.61
17	fr	prep_anim	0.31	0.19
18	fr	simple_agrmt	0.50	0.39
19	fr	subj_rel	0.54	0.14
20	fr	vp_coord	0.95	0.70
21	he	long_vp_coord	0.80	0.15
22	he	obj_rel_across_anim	0.84	0.62
23	he	obj_rel_within_anim	0.91	0.10
24	he	prep_anim	0.68	0.51
25	he	simple_agrmt	0.66	0.34
26	he	subj_rel	0.55	0.44
27	he	vp_coord	0.77	0.36
28	ru	long_vp_coord	0.66	0.52
29	ru	obj_rel_across_anim	0.66	0.25
30	ru	obj_rel_within_anim	0.65	0.06
31	ru	prep_anim	0.33	0.76
32	ru	simple_agrmt	0.50	0.61
33	ru	subj_rel	0.61	0.18
34	ru	vp_coord	0.57	0.73

Table 3: Accuracies of multilingual model (mBERT) and monolingual models on each combination of language and syntactic relationship

References

[1] Bentz, C. et al. (Dec. 2016). "A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora". In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. Ed. by Brunato, D. et al. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 142–153. URL: https://aclanthology.org/W16-4117.