L-Università ta' Malta

university of groningen

Cross-lingual transfer learning with Persian

Sepideh Mollanorozy, Dr. Marc Tanti, Prof. Malvina Nissim

Sigtyp workshop @ EACL 2023

Dubrovnik, Croatia

06.05.2023

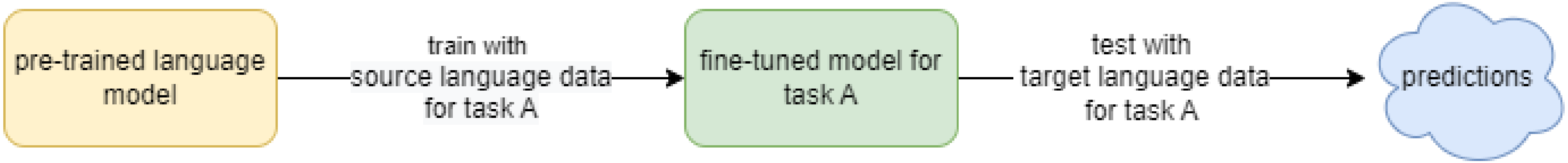
Table of contents

1. Introduction
2. Background
3. POS tagging
4. Sentiment analysis
5. Conclusion
6. References

1. Introduction

Transfer Learning

- Source and target language
- English as source, why?
- Language similarity, POS tagging
- Persian can be beneficial?


pre-trained language model
train with source language data for task A
fine-tuned model for task A
test with target language data for task A
predictions

Persian

- Country, Dialect:
  - Iran, Iranian Persian (Officially Persian)
  - Afghanistan, Dari
  - Tajikistan, Tajik
- Indo-European
- Persian alphabet (32 letters)
- SOV word order
- More than 85 million people

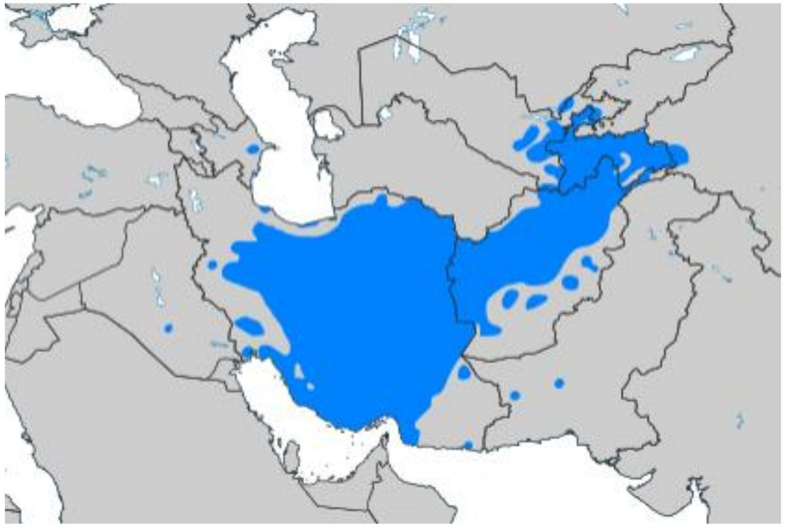


Fig2: regions where people's mother tongue is Persian (Commons, 2021b)

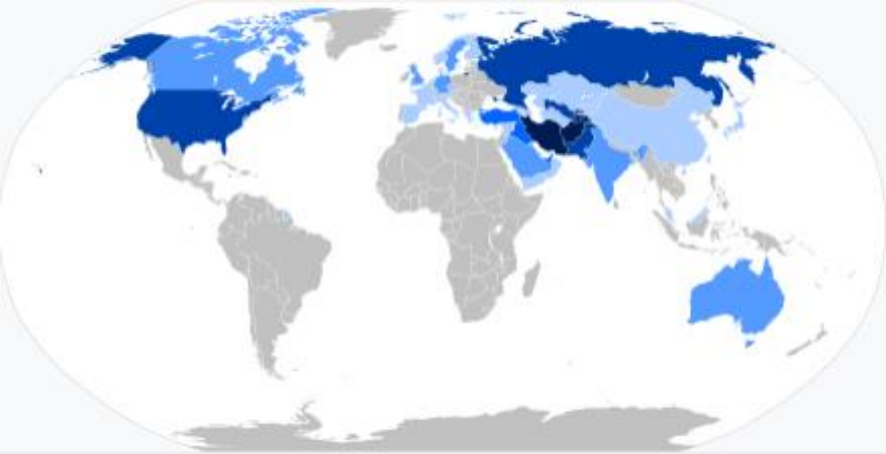


Fig3: Persian speakers around the world (Commons, 2021a)

Research questions

- Language similarity and Persian for POS tagging
- Linguistic features of the matching languages
- Performance of ParsBERT and XLM-RoBERTa
- Matching languages with Persian for Sentiment analysis

2. Background

Language Similarity

- LDND distance measure (Wichmann et al. , 2010)
- Levenshtein distance (LD)
  - minimum number of times needed to add, delete, or substitute a character
- normalized LD (LDN)
  - dividing LD by the maximum length
  - Omit the influence of long words transformed into short words (high LD)
- LDND
  - Dividing the LDN value by the mean of all LDN values between each two words

Language similarity and transfer learning

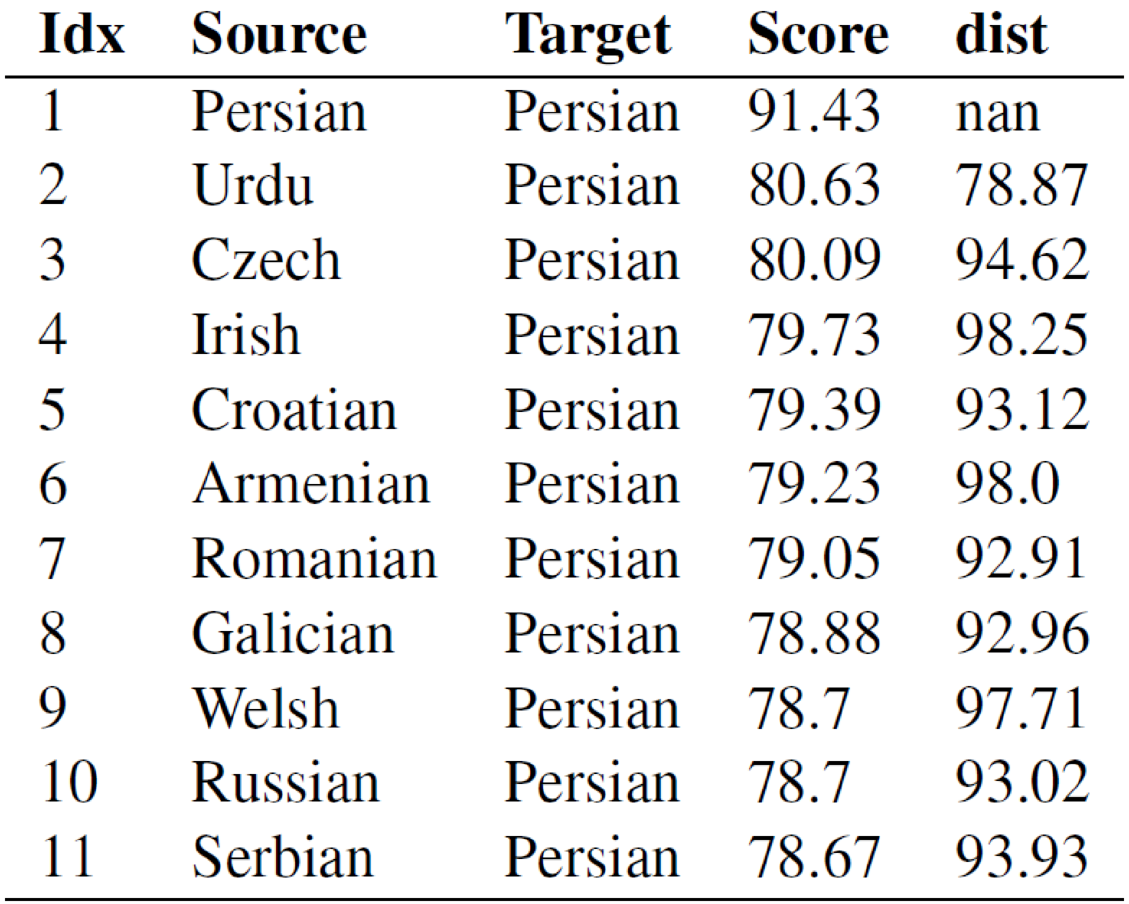
- de Vries et al. (2022)
  - POS tagging task
  - Search for good pairs and success factors
  - pre-trained multilingual language model XLM-RoBERTa (Conneau et al., 2019)
  - No global source languages
  - Success factors:
    - target in pre-training
    - LDND distance

3. POS tagging

POS tagging analysis

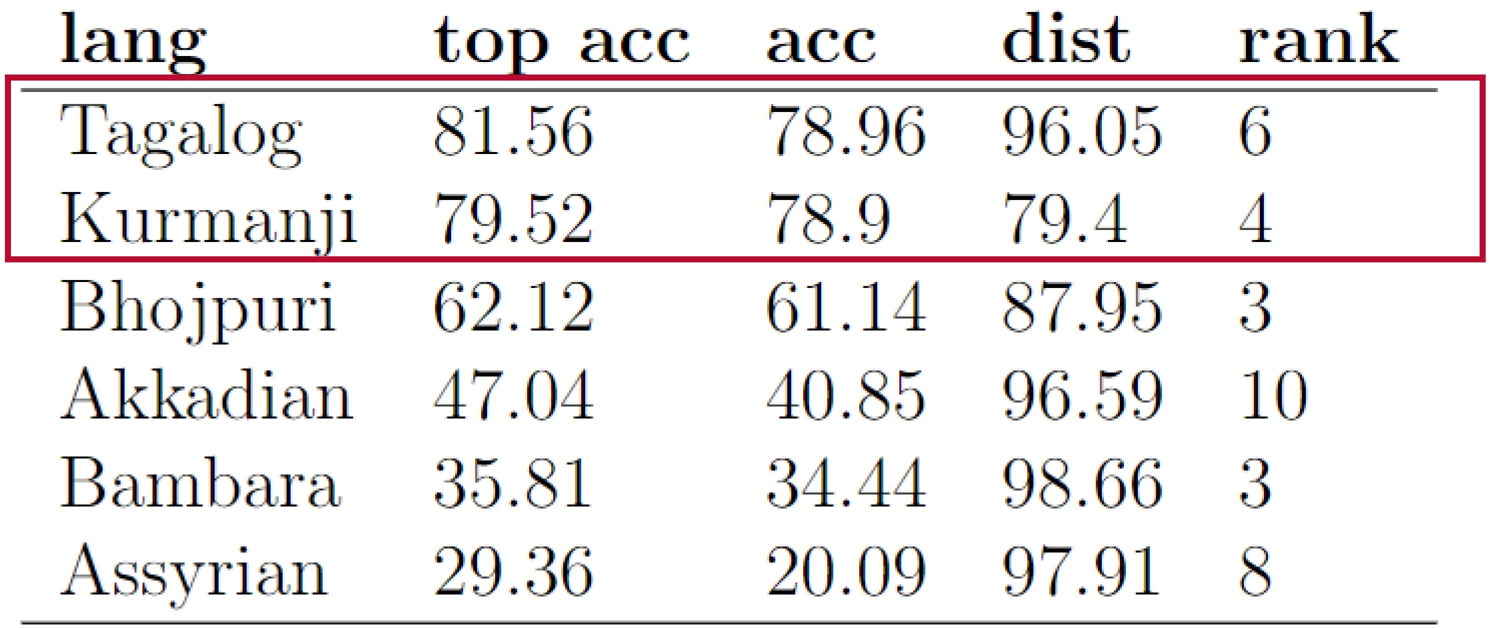
- UD dataset, 17 tags
- 65 source and 105 target
- Pre-trained (CommonCrawl data) XLM-RoBERTa language model
- Fine-tune and test with source-target combinations
- Accuracy score
- LDND distance

Persian as target



Idx	Source	Target	Score	dist
1	Persian	Persian	91.43	nan
2	Urdu	Persian	80.63	78.87
3	Czech	Persian	80.09	94.62
4	Irish	Persian	79.73	98.25
5	Croatian	Persian	79.39	93.12
6	Armenian	Persian	79.23	98.0
7	Romanian	Persian	79.05	92.91
8	Galician	Persian	78.88	92.96
9	Welsh	Persian	78.7	97.71
10	Russian	Persian	78.7	93.02
11	Serbian	Persian	78.67	93.93

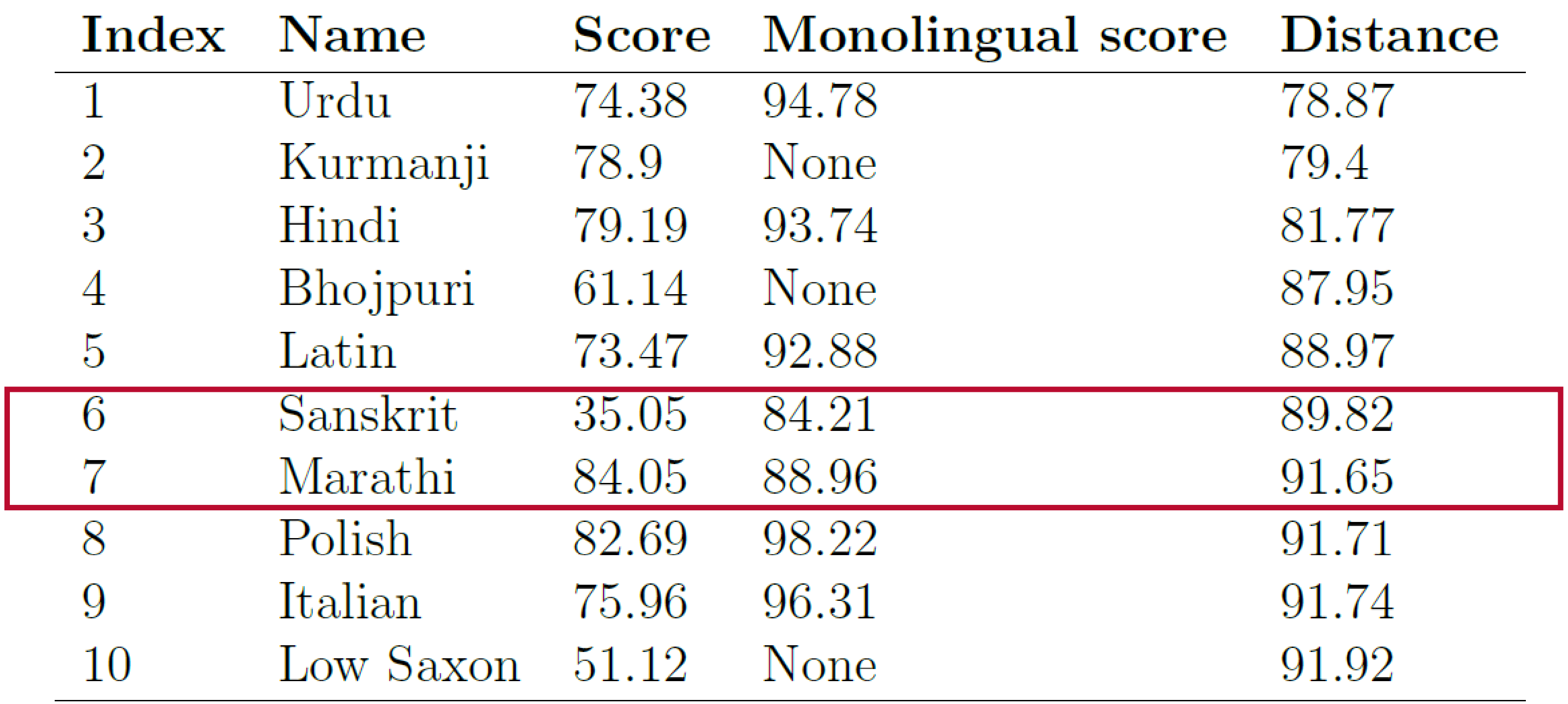
Persian source, low-resource target



lang	top acc	acc	dist	rank
Tagalog	81.56	78.96	96.05	6
Kurmanji	79.52	78.9	79.4	4
Bhojpuri	62.12	61.14	87.95	3
Akkadian	47.04	40.85	96.59	10
Bambara	35.81	34.44	98.66	3
Assyrian	29.36	20.09	97.91	8

Lowest dist among others

LDND distance with Persian

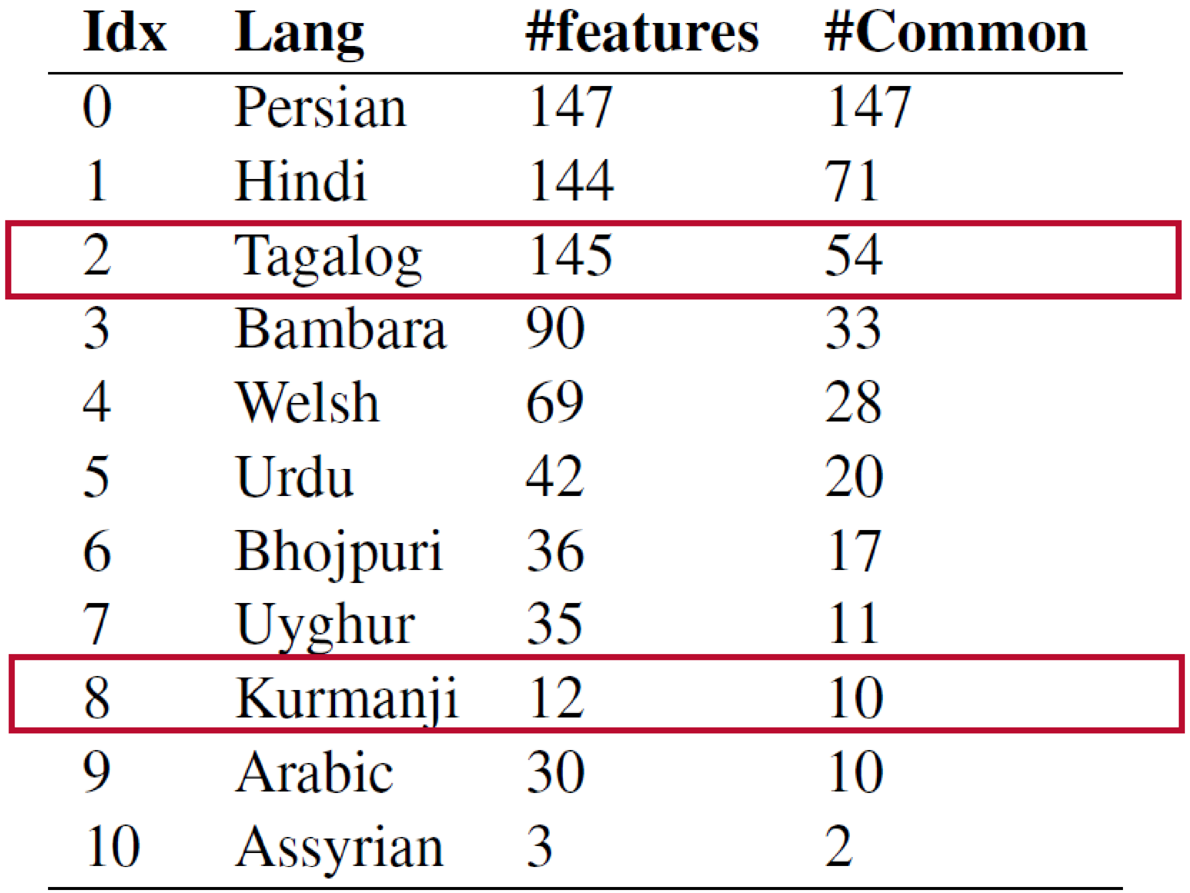


Index	Name	Score	Monolingual score	Distance
1	Urdu	74.38	94.78	78.87
2	Kurmanji	78.9	None	79.4
3	Hindi	79.19	93.74	81.77
4	Bhojpuri	61.14	None	87.95
5	Latin	73.47	92.88	88.97
6	Sanskrit	35.05	84.21	89.82
7	Marathi	84.05	88.96	91.65
8	Polish	82.69	98.22	91.71
9	Italian	75.96	96.31	91.74
10	Low Saxon	51.12	None	91.92

LDND not a good measure

WALS linguistic features

- (Dryer & Haspelmath, 2013)
- Help to explain neural network performance
- Language similarity measure based on number of common features
- Potential ground for Tagalog high score

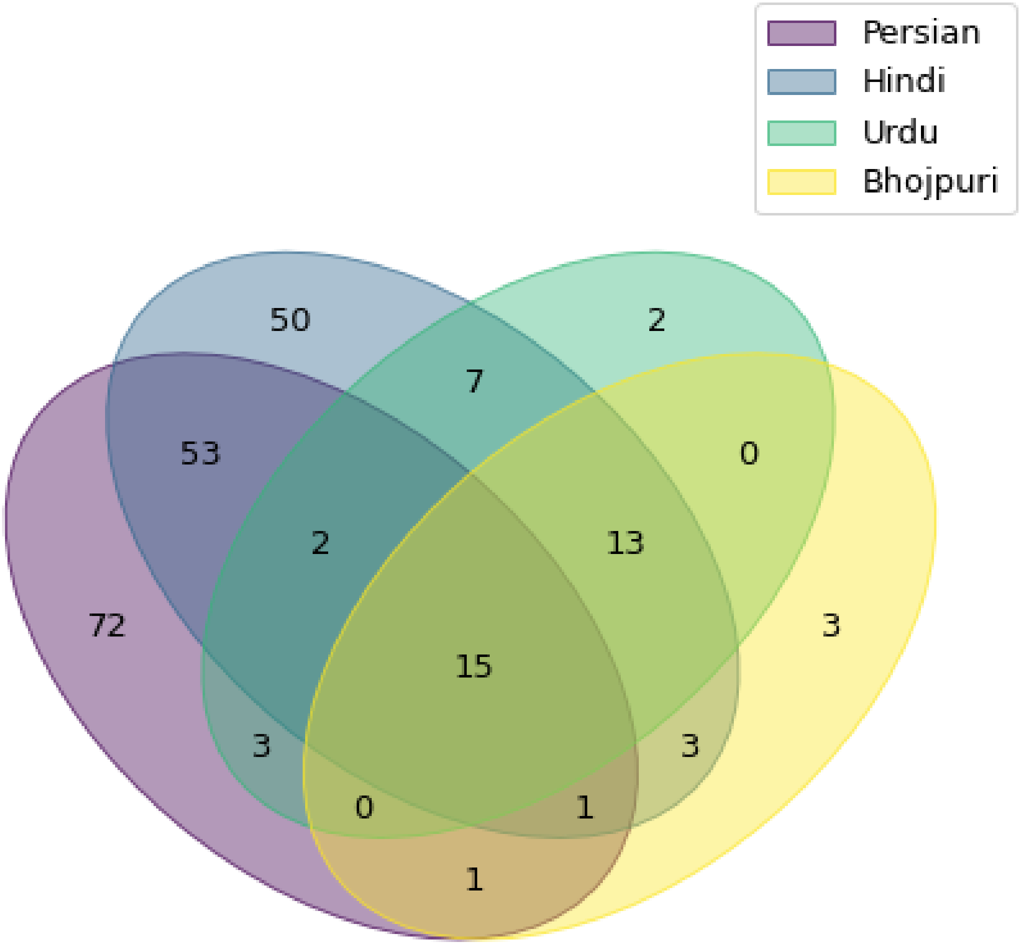


Idx	Lang	#features	#Common
0	Persian	147	147
1	Hindi	144	71
2	Tagalog	145	54
3	Bambara	90	33
4	Welsh	69	28
5	Urdu	42	20
6	Bhojpuri	36	17
7	Uyghur	35	11
8	Kurmanji	12	10
9	Arabic	30	10
10	Assyrian	3	2

WALS linguistic features

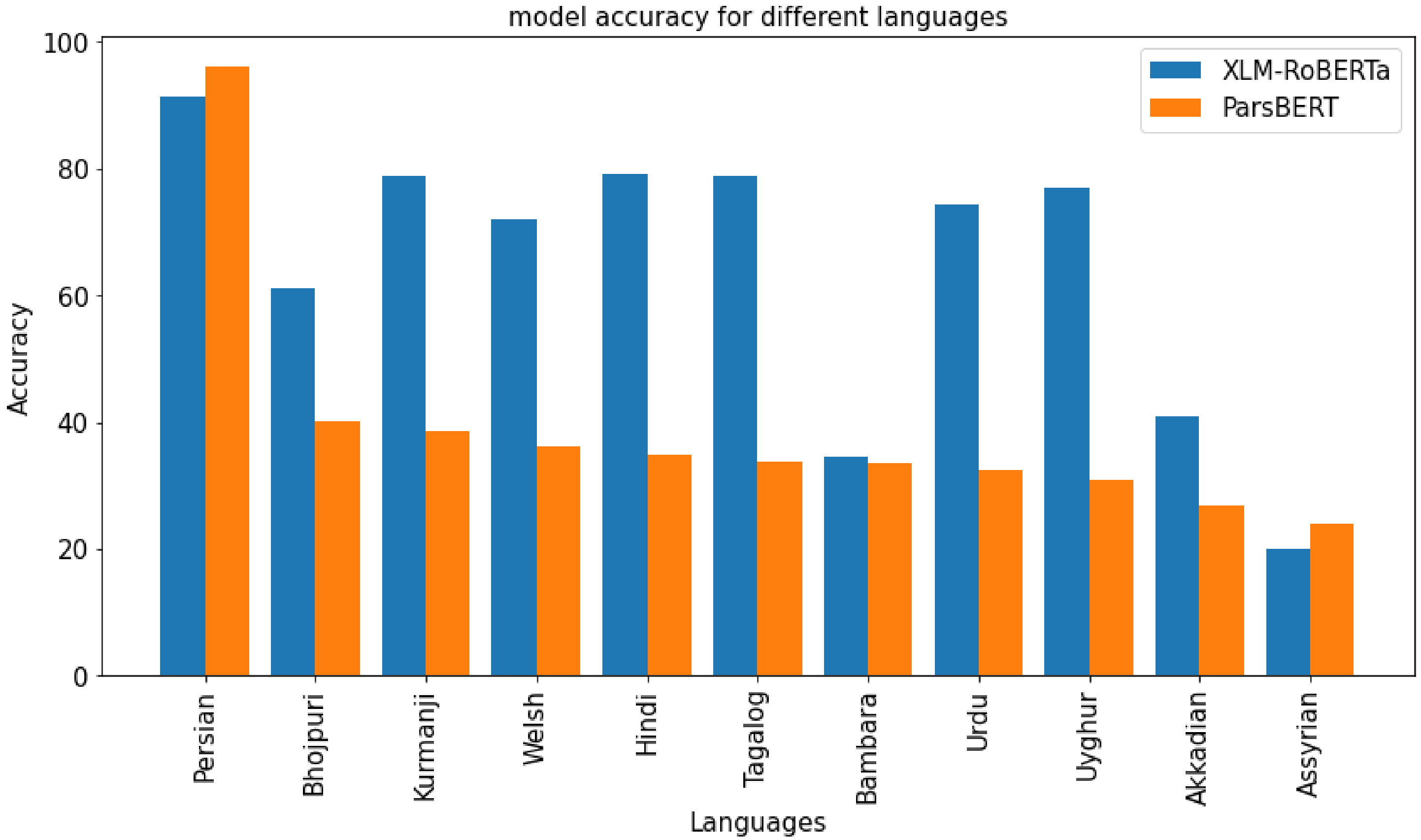
Mostly syntactic features:

- SOV
- Demonstrative-Noun
- Numeral-Noun
- initial position of Polar Question Particles



ParsBERT

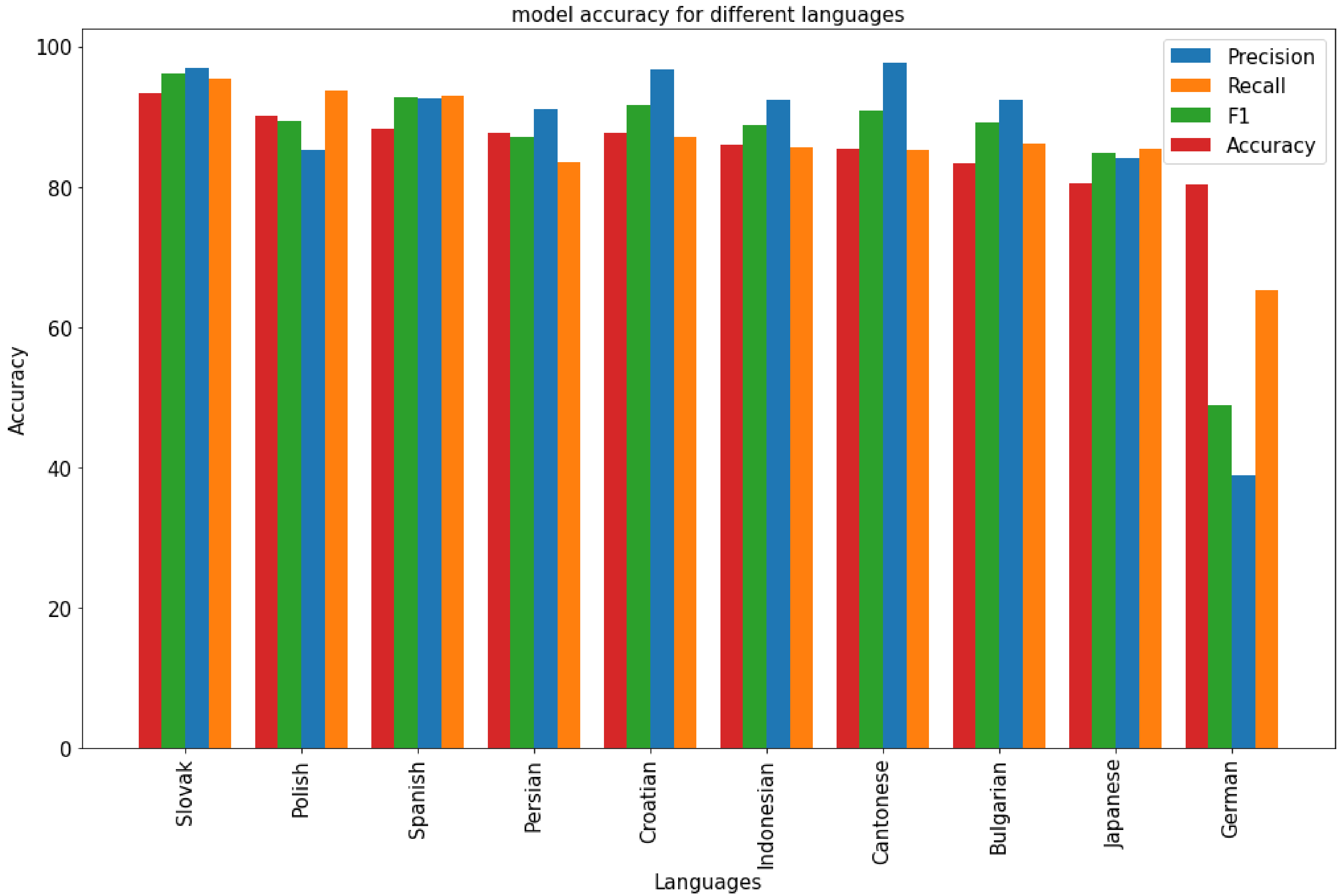
- (Farahani et al, 2021)
- Pre-trained monolingual Persian language model:
    - MLM
    - next sentence prediction
- Fine-tune with Persian
- Inference with others



4. Sentiment Analysis

Data Collection

- Binary dataset
- 23 languages from Martinez-Garcia et al. (2021):
  - Algerian, Arabic, Basque, Bulgarian, Cantonese, Chinese, Croatian, English, Finnish, German, Greek, Hebrew, Indonesian, Japanese, Korean, Maltese, Norwegian, Russian, Slovak, Spanish, Thai, Turkish, and Vietnamese
- 8 languages from various sources gathered
- Identical structure
- Public access at https://huggingface.co/sepidmnorozy


model accuracy for different languages
Accuracy
100
80
60
40
20
0
Precision
Recall
F1
Accuracy
Slovak
Polish
Spanish
Persian
Croatian
Indonesian
Cantonese
Bulgarian
Japanese
German
Languages

5. Conclusion

Conclusion

- Monolingual Persian 91.43% POS tagging! Persian best case for itself!
- Persian a potential good source for Kurmanji and Tagalog for other tasks
- ParsBERT outperforms XLM-RoBETa only for monolingual Persian 96%
- Monolingual Persian is not the best for sentiment analysis
- Task-dependent

References

Commons, W. (2021a). File:map of persian speakers.svg — Wikimedia commons, the free media repository. Retrieved from https://commons.wikimedia.org/w/index.php?title=File:Map of Persian speakers. svgoldid=527368091 ([Online; accessed 13-February-2022])

Commons, W. (2021b). File:persian language location map.svg — wikimedia commons, the free media repository. Retrieved from https://commons.wikimedia.org/w/index.php?title=File:Persian Language Location Map.svgoldid=606196262 ([Online; accessed 13-February-2022])

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzm´an, F., Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. Retrieved from https://arxiv.org/abs/1911.02116 doi: 10.48550/ARXIV.1911.02116

de Vries, W., Bartelds, M., Nissim, M., & Wieling, M. (2021). Adapting monolingual models: Data can be scarce when language similarity is high. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 . Retrieved from http://dx.doi.org/10.18653/v1/2021.findings-acl.433 doi: 10.18653/v1/2021.findings-acl.433

References

de Vries, W., Wieling, M., & Nissim, M. (2022, 05). Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers) (p. 7676-7685). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.529 doi: 10.18653/v1/2022.acl-long.529

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). Wals online. Retrieved from https://wals.info/

Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2020, 05). Parsbert: Transformer-based model for persian language understanding.

Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021, 10). ParsBERT: Transformer-based model for persian language understanding. Neural Processing Letters, 53 (6), 3831–3847. Retrieved from https://doi.org/10.1007%2Fs11063-021-10528-4 doi: 10.1007/s11063-021-10528-4

References

Martinez-Garcia, A., Badia, T., & Barnes, J. (2021, 08). Evaluating morphological typology in zero-shot cross-lingual transfer. Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.244 doi: 10.18653/v1/2021.acl-long.244

Wichmann, S., Holman, E. W., Bakker, D., & Brown, C. H. (2010). Evaluating linguistic distance measures. Physica A: Statistical Mechanics and its Applications, 389 (17), 3632-3639. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378437110003997 doi:https://doi.org/10.1016/j.physa.2010.05.011

Thanks for your attention!

Any questions?

You can reach me at:

sepid.mnorozy@gmail.com