Note this is for the 2019 unconference. We will have 2020 ones soon.

NLP UnConference - Language Identification

Top level points discussed:

1) Rule based methods vs ML for under-resourced languages

Rule-based methods may achieve high performance but the problem is that linguistic experts are needed for each language. Hence, more time is required and language do change as the years pass.
Machine learning methods perform well but for text, our under-resourced languages are not well represented in terms of writing. Some characters are not available in the keyboard hence people use other characters as placeholders.

2) The struggle to deal with:
    a) multiple accents or dialects for spoken LID (words are pronounced differently)
    b) removing background noise + transcribing the speech are large sources of error.
3) Do we really need neural methods for LID ?
4) LID for text in non english alphabets.

GitHub repo for subword inspiration with paper.

Possible Further Discussion Points

Naive Bayesian model vs. pre-trained language model

Related to point 3 above - From the literature it seems as if naive Bayesian character n-gram models do very well. Neural methods should be able to take more semantic context into account and should be able to do better. Maybe more so when code switching is present.
Has anybody done any work comparing a naive Bayesian model with a model based on a pre-trained language model?

Language Identification Shared Task

It would possibly be useful to start a shared language ident corpus and challenge. The code and corpus at https://github.com/praekelt/feersum-lid-shared-task could be a starting point, but more work needs to be done.

Text Language Identification References

- *The Workshop on NLP for Similar Languages, Varieties and Dialects*, https://www.aclweb.org/anthology/W19-1400, 2019 (very cool)
- Gabriel Bernier-Colborne, Cyril Goutte, Serge Léger*, Improving Cuneiform Language Identification with BERT*, Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, 2019, https://www.aclweb.org/anthology/W19-1402
- Tommi Sakari Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lind n. *Automatic language identification in texts: A survey*. Journal of Artificial Intelligence Research, 65:675 782, 2019.
- Martin Puttkammer, Roald Eiselen, Justin Hocking, and Frederik Koen. *NLP web services for resource-scarce languages.* In Proceedings of ACL 2018, System Demonstrations, pages 43 49, Melbourne, Australia, July 2018. Association for Computational Linguistics
- Jeanne E. Daniel at SUN did some language ident work for a poster at DLI2018. I'll try and get a reference for a related paper if any.
- B. Duvenhage, M. Ntini and P. Ramonyai published a paper at PRASA 2017. A GitHub repo with link to paper, corpus (derived from SADILAR) and source code is available at https://github.com/praekelt/feersum-lid-shared-task.
- A. Selamat and N. Akosu, "Word-length algorithm for language identification of under-resourced languages," Journal of King Saud University - Computer and Information Sciences, vol. 28, no. 4, pp. 457 – 469, 2016. [Online]. Available: http://www.sciencedirect.com/ science/article/pii/S1319157815000609.
- O. Giwa and M. H. Davel, "Language identification of individual words with joint sequence models," in Proceedings of the Annual Conference of the International Speech Communication Association INTER-SPEECH, 2014.
- "Developing text resources for ten south african languages." in In Proceedings of the 9th International Conference on Language Resources and Evaluation, 2014.
- O. Giwa and M. Davel, "N-gram based language identification of individual words," in Proceedings of the Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), 2013.
- G. Botha and E. Barnard, "Factors that affect the accuracy of text-based language identification," Comput. Speech Lang., vol. 26, no. 5, pp. 307–320, Oct. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.csl. 2012.01.004.
- G. Botha, V. Zimu, and E. Barnard, "Text-based language identification for the South African languages," in SAIEE Africa Research Journal, 2006.
- H. Combrinck and E. Botha, "Text-based automatic language identifi- cation." in In Proceedings of the 6th Annual Symposium of the Pattern Recognition Association of South Africa, 1994.

Challenges

Good training data for code switching and how people typically speak on twitter, WhatsApp, Facebook, etc.