Short text language identification for under resourced languages

Bernardt Duvenhage

Feersum Engine, Praekelt Consulting, Johannesburg, South Africa
bernardt@praekelt.com

Abstract. The paper presents a hierarchical naive Bayesian and lexicon based classifier for short text language identification (LID) useful for under resourced languages. The algorithm is evaluated on short pieces of text for the 11 official South African languages some of which are similar languages. [1]

Keywords: Language identification · Similar languages.

1 Background

Accurate language identification (LID) is the first step in many natural language processing and machine comprehension pipelines. LID is further also an important step in harvesting scarce language resources. Availability of data is still one of the big roadblocks for applying data driven approaches like supervised machine learning in developing countries.

An in depth survey of algorithms, features, datasets, shared tasks and evaluation methods may be found in [5]. The datasets for the DSL 2015 & DSL 2017 shared tasks [8] are often used in LID benchmarks. The NCHLT text corpora [1] may be used for a shared LID task for the South African languages. The DSL 2017 paper [8] gives an overview of the solutions of all of the teams that competed on the shared task and the winning approach [2] used an SVM with character n-gram, parts of speech tag features and some other engineered features. The winning approach for DSL 2015 [7] used an ensemble naive Bayes classifier. The fasttext classifier [6] is perhaps one of the best known efficient 'shallow' text classifiers that have been used for LID [2]. Hierarchical stacked classifiers (including lexicons) have also been proposed that would for example first classify a piece of text by language group and then by exact language [4][3].

2 Methodology and results

The proposed LID algorithm[3] builds on the work in [3] and [7]. We apply a naive Bayesian classifier with character (2, 4 & 6)-grams, word unigram and

[1] Full paper presented at NeurIPS 2019 Workshop on Machine Learning for the Developing World.
[2] https://fasttext.cc/blog/2017/10/02/blog-post.html
[3] Available at https://github.com/praekelt/feersum-lid-shared-task.

word bigram features with a hierarchical lexicon based classifier. The algorithm is evaluated against recent approaches using existing test sets from previous works on South African languages as well as the Discriminating between Similar Languages (DSL) 2015 and 2017 shared tasks.

The naive Bayesian classifier is trained to predict the specific language label of a piece of text, but used to first classify text as belonging to either the Nguni family, the Sotho family, English, Afrikaans, Xitsonga or Tshivenda. The lexicon based classifier is then used to predict the specific language within a language group. If the lexicon prediction of the specific language has high confidence then its result is used as the final label else the naive Bayesian classifier's specific language prediction is used as the final result. The lexicon is built over all the data and includes the vocabulary from both the training and testing sets.

Table 1. LID Accuracy - The models we executed ourselves are marked with *. The results that are not available from our own tests or the literature are indicated with '—'.

Model	Algorithm	NCHLT	DSL '15	DSL '17
Joulin et al. 2017 [6] *	fasttext	93.30	93.20	88.60
Bestgen 2017 (DSL winner) [2]	SVM	—	—	92.74
Malmasi & Dras 2015 (DSL winner) [7]	NB ensemble	—	95.54	—
Duvenhage et al. 2017 [3] *	NB+Lex	94.59	—	—
Naive-Bayes only *	NB	94.36	94.98	91.89
Stacked model *	NB+Lex	**96.12**	**99.34**	**98.70**
Stacked model (50% lex dropout) *	NB+Lex	94.90	98.06	96.21

The average classification accuracy results are summarised in Table 1. The accuracies reported are for classifying a piece of text by its specific language label. The accuracy of the proposed algorithm seems to be dependent on the support of the lexicon. Without a good lexicon a non-stacked naive Bayesian classifier might even perform better.

3 Conclusion

LID of short texts, informal styles and similar languages remains a difficult problem which is actively being researched. We would like to investigate the value of a lexicon in a production system and how to possibly maintain it using self-supervised learning. We are investigating the application of deeper language models some of which have been used in more recent DSL shared tasks. We would also like to investigate data augmentation strategies to reduce the amount of training data that is required.

Further research opportunities include data harvesting, building standardised datasets and shared tasks for South Africa as well as the rest of Africa. In general, the support for language codes that include more languages seems to be growing, discoverability of research is improving and paywalls seem to no longer be a big problem in getting access to published research.

References

1. NCHLT text corpora (2014), available from http://www.nwu.ac.za/ctext
2. Bestgen, Y.: Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). pp. 115–123. Association for Computational Linguistics, Valencia, Spain (Apr 2017). https://doi.org/10.18653/v1/W17-1214, https://www.aclweb.org/anthology/W17-1214
3. Duvenhage, B., Ntini, M., Ramonyai, P.: Improved text language identification for the south african languages. 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech) pp. 214–218 (2017)
4. Goutte, C., Léger, S., Carpuat, M.: The NRC system for discriminating similar languages. In: Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects. pp. 139–145. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (Aug 2014). https://doi.org/10.3115/v1/W14-5316, https://www.aclweb.org/anthology/W14-5316
5. Jauhiainen, T.S., Lui, M., Zampieri, M., Baldwin, T., Lindén, K.: Automatic language identification in texts: A survey. Journal of Artificial Intelligence Research **65**, 675–782 (2019)
6. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://www.aclweb.org/anthology/E17-2068
7. Malmasi, S., Dras, M.: Language identification using classifier ensembles. In: Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects. pp. 35–43. Association for Computational Linguistics, Hissar, Bulgaria (Sep 2015), https://www.aclweb.org/anthology/W15-5407
8. Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., Aepli, N.: Findings of the VarDial evaluation campaign 2017. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). pp. 1–15. Association for Computational Linguistics, Valencia, Spain (Apr 2017). https://doi.org/10.18653/v1/W17-1201, https://www.aclweb.org/anthology/W17-1201