Broad-Scope Language Identification

Susanne Fertmann, Guy Emerson, Liling Tan

Background

What language is this?

i) ადამიანის უფლებათა საყოველთაო დეკლარაცია

ii) Nou tou imen nou'n ne dan laliberte ek legalite

Most previous approaches to language identification only deal with a small number of languages, which neglects low-resource languages entirely.

Baldwin and Lui (2010) have shown that the task is difficult when the number of possible languages is large, or when the input text is short.

They consider 67 languages, which is the broadest existing system we are aware of.

Objectives

- *Produce a language identification system that can deal with a wide range of languages.*
- *Compile a corpus for training and evaluation.*

Model

We used frequencies of character n-grams and words as features, and tested two types of model, assuming all languages were equally likely:

Cosine similarity (if vectors are normalised to unit length):

$$\text{Sim}(f, x) = \sum x_i f_i$$

Multinomial Naive Bayes (if f is normalised to sum to one):

$$\log P(f \mid x) \propto \sum x_i \log f_i$$

To avoid infinities, we applied Simple Good-Turing smoothing, which reserves some probability mass for unseen items.

Data Collection

We crawled and cleaned data from:

***Omniglot** – Multilingual phrases and babel story translation*

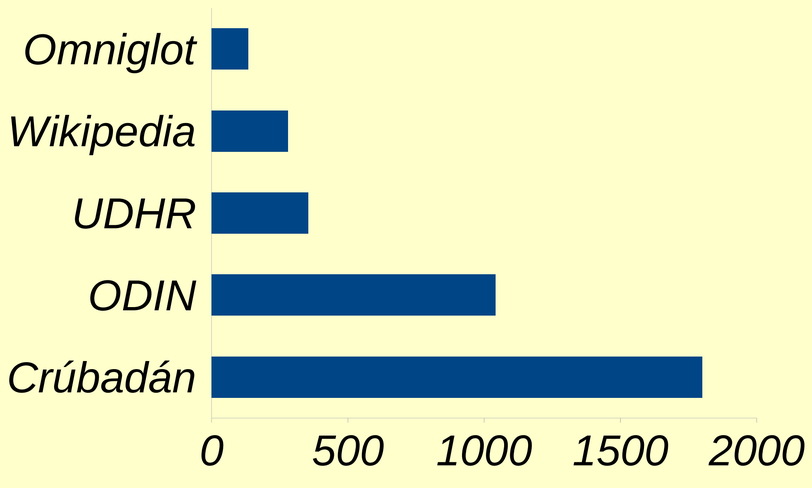
***Wikipedia** – Web encyclopedia*

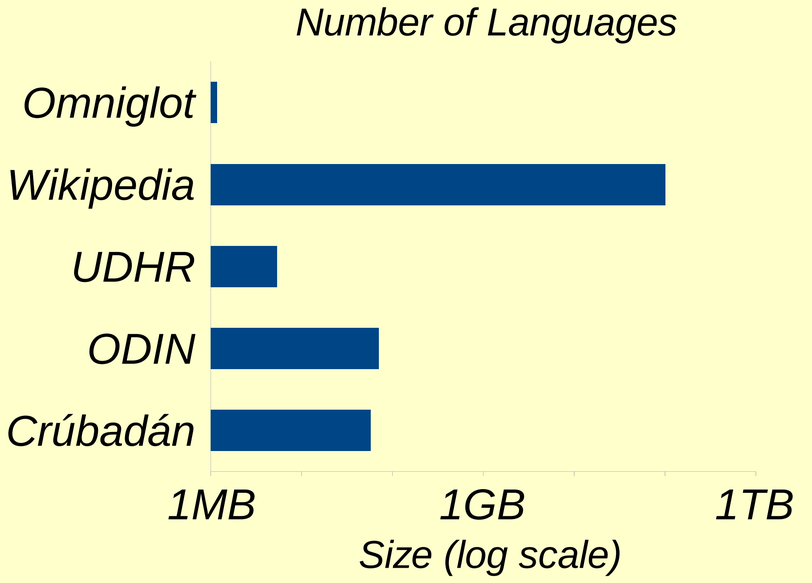
***UDHR** (Universal Declaration of Human Rights)*

***ODIN** (Online Database of Interlinear Glossed Text) – IGTs from linguistics papers (Lewis and Xia, 2010)*

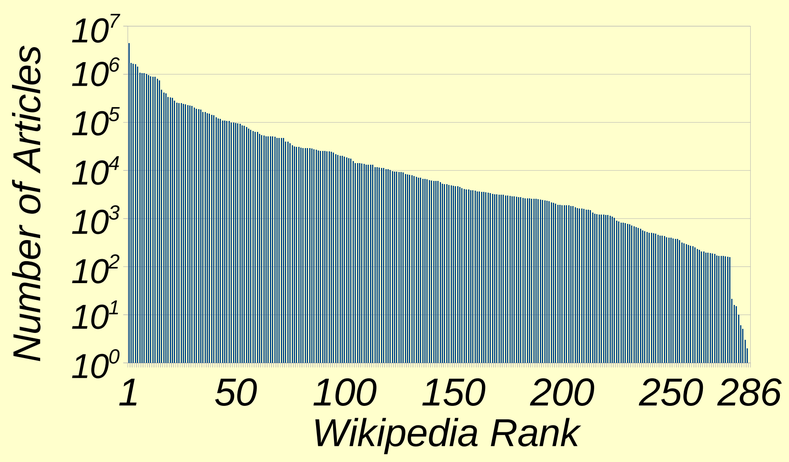
***An Crúbadán** – Character n-gram and word frequencies collected through web crawling (Scannell, 2007)*

Comparison of Size of Datasets:





Challenge: Imbalanced Data



Challenge: Standardisation

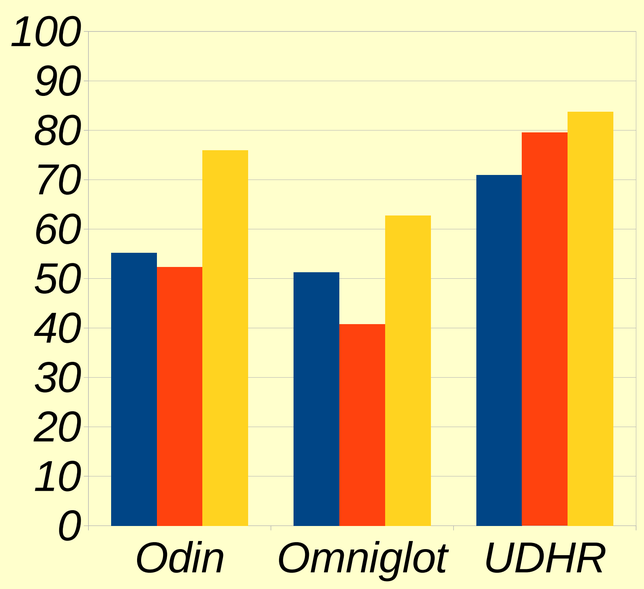
- *Different data resources use different language codes.*
- *We built an automatic mapping to ISO-639-3/5.*

Results

We performed ten-fold cross-validation for the cosine model, comparing different three different feature sets:

- *Character 1-5 grams*
- *Words*
- *Combined*

Average Accuracy:



- *Accuracy depends heavily on the dataset being used.*
- *Neither character n-grams nor words consistently outperforms the other.*
- *The combined model outperforms using only words or character n-grams, for all datasets.*

Conclusion

- *Data is available but difficult to access and standardise.*
- *We crafted a corpus with >1000 languages from different resources.*
- *Our language identification system can deal with >1000 languages.*
- *Results are competitive with other existing systems.*
- *Corpus and models are open-source.*

Future Work

- *Evaluation of the Multinomial Naive Bayes model*
- *Feature selection*
- *Language family identification*

References

Baldwin, Timothy, and Marco Lui (2010). "Language identification: The long and the short of the matter." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

Lewis, William D., and Fei Xia (2010). "Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages." In *Literary and Linguistic Computing 25.3*

Scannell, Kevin P (2007). "The Crúbadán Project: Corpus building for under-resourced languages." In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop. Vol. 4. 2007.*

UNIVERSITÄT DES SAARLANDES