Language detection using character n-gram profiles

Inspiration from Cavnar and Trenkle (1994)

Atreya Shankar
Applying for: Scientific Researcher in NLP

July 6, 2021

Overview

Motivation

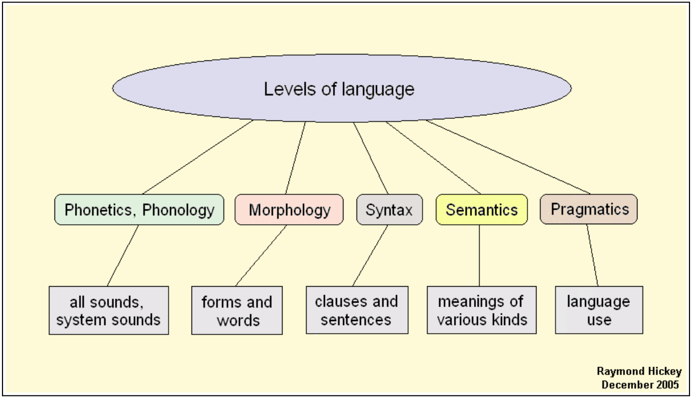


Figure 1: Levels/structures of languages; figure taken from Hickey (2005)

- Morphological profiling probably has lower data and compute requirements
- Makes sense given no external libraries are allowed

Motivation

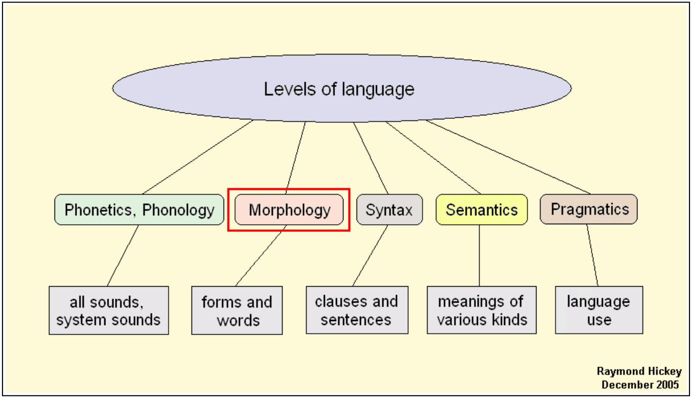


Figure 1: Levels/structures of languages; figure taken from Hickey (2005)

- Morphological profiling probably has lower data and compute requirements
- Makes sense given no external libraries are allowed

Motivation

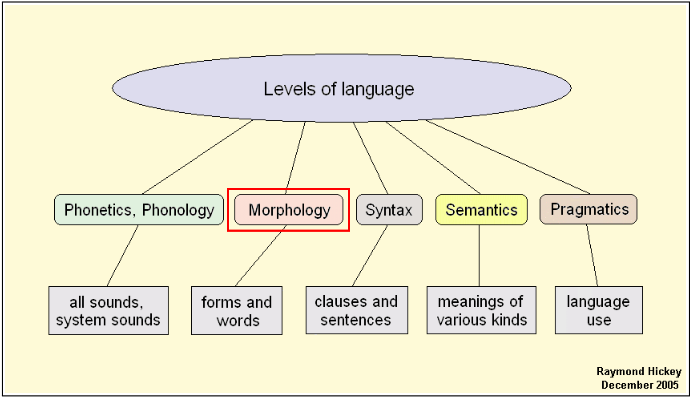


Figure 1: Levels/structures of languages; figure taken from Hickey (2005)

- Morphological profiling probably has lower data and compute requirements
- Makes sense given no external libraries are allowed

Methodology

- Character n-gram profiling technique from Cavnar, Trenkle, et al. (1994)
- WiLI-2018 data set for 235 languages with 235,000 paragraphs (Thoma, 2018)
- **Similarities:** Data is lower-cased and punctuation/special-tokens are removed
- **Differences:** Use vector-based difference norm instead of out-of-place distance
- **Two hyperparameters:** character n-gram length and ranked n-gram cutoff

N-Gram-Based Text Categorization

William B. Cavnar and John M. Trenkle
Environmental Research Institute of Michigan
P.O. Box 134001
Ann Arbor MI 48113-4001

Figure 2: Excerpt from Cavnar, Trenkle, et al. (1994)

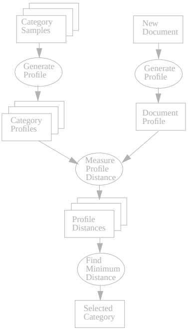
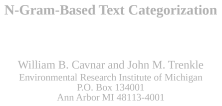


Figure 3: Flowchart from Cavnar, Trenkle, et al. (1994)

Methodology

- Character n-gram profiling technique from Cavnar, Trenkle, et al. (1994)
- WiLI-2018 data set for 235 languages with 235,000 paragraphs (Thoma, 2018)
- **Similarities:** Data is lower-cased and punctuation/special-tokens are removed
- **Differences:** Use vector-based difference norm instead of out-of-place distance
- **Two hyperparameters:** character n-gram length and ranked n-gram cutoff


N-Gram-Based Text Categorization

William B. Cavnar and John M. Trenkle
Environmental Research Institute of Michigan
P.O. Box 134001
Ann Arbor MI 48113-4001

Figure 2: Excerpt from Cavnar, Trenkle, et al. (1994)

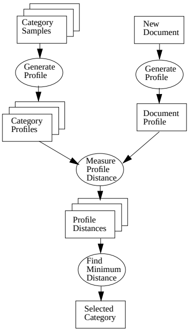


Figure 3: Flowchart from Cavnar, Trenkle, et al. (1994)

Results

N-gram length	N-gram cutoff	Weighted Test F_1	Best Language	Worst Language
2	100	0.865	Navajo	Konkani
2	300	0.893	Navajo	Pampanga
3	100	0.859	Dhivehi	Chavacano
3	**300**	**0.898**	Navajo	Chavacano

Table 1: Tabular summary of model performances; MLP from Thoma (2018) achieved an accuracy of 0.883

Language	N_1	N_2	N_3	N_4	N_5
English	*the*	*and*	*ing*	*ion*	*ent*
Deutsch	*der*	*sch*	*die*	*ein*	*che*
Italiano	*del*	*ell*	*ent*	*ion*	*lla*

Table 2: Tabular summary of top five character trigrams with highest relative frequency per language

Results

N-gram length	N-gram cutoff	Weighted Test F_1	Best Language	Worst Language
2	100	0.865	Navajo	Konkani
2	300	0.893	Navajo	Pampanga
3	100	0.859	Dhivehi	Chavacano
3	**300**	**0.898**	Navajo	Chavacano

Table 1: Tabular summary of model performances; MLP from Thoma (2018) achieved an accuracy of 0.883

Language	N_1	N_2	N_3	N_4	N_5
English	*the*	*and*	*ing*	*ion*	*ent*
Deutsch	*der*	*sch*	*die*	*ein*	*che*
Italiano	*del*	*ell*	*ent*	*ion*	*lla*

Table 2: Tabular summary of top five character trigrams with highest relative frequency per language

Discussion

Gold language	Utterance	Predicted language
English	*What is this?*	Cantonese
Deutsch	*Was ist das?*	Chavacano
Italiano	*Cos'è questo?*	Asturian

Table 3: Examples of erroneous language detection for short phrases

Plenty of failing cases:

- Short phrases where language profile cannot converge
- Slang, colloquial or borrowed words
- Transliteration from non-Latin to Latin script

Plenty of workarounds:

- Word-level language identification with large-enough vocabulary
- Complex modeling over sequential subwords, for example using neural networks; such as in Bartz et al. (2017)

Discussion

Gold language	Utterance	Predicted language
English	*What is this?*	Cantonese
Deutsch	*Was ist das?*	Chavacano
Italiano	*Cos'è questo?*	Asturian

Table 3: Examples of erroneous language detection for short phrases

Plenty of failing cases:

- Short phrases where language profile cannot converge
- Slang, colloquial or borrowed words
- Transliteration from non-Latin to Latin script

Plenty of workarounds:

- Word-level language identification with large-enough vocabulary
- Complex modeling over sequential subwords, for example using neural networks; such as in Bartz et al. (2017)

Conclusions

- Portable and lightweight character n-gram profiling technique from Cavnar, Trenkle, et al. (1994)
- Trained and tested on WiLI-2018 (Thoma, 2018)
- Best character trigram model achieved **89.8%** weighted F_1 test score
- Works well for medium-long length documents, likely robust to previously unseen words and spelling errors
- Known limitations on short length documents

Conclusions

- Portable and lightweight character n-gram profiling technique from Cavnar, Trenkle, et al. (1994)
- Trained and tested on WiLI-2018 (Thoma, 2018)
- Best character trigram model achieved **89.8%** weighted F_1 test score
- Works well for medium-long length documents, likely robust to previously unseen words and spelling errors
- Known limitations on short length documents

Bibliography I

Bartz, Christian, Tom Herold, Haojin Yang, and Christoph Meinel (2017). "Language Identification Using Deep Convolutional Recurrent Neural Networks". In: *CoRR* abs/1708.04811. arXiv: 1708.04811. URL: http://arxiv.org/abs/1708.04811.

Cavnar, William B, John M Trenkle, et al. (1994). "N-gram-based text categorization". In: *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. Vol. 161175. Citeseer.

Hickey, Raymond (2005). "Levels of language". In: *Universität Duisburg-Essen*.

Thoma, Martin (2018). "The WiLI benchmark dataset for written language identification". In: *arXiv preprint arXiv:1801.07779*.