SIGTYP 2020 Shared Task

Rudolf Rosa, Martin Vastl, Daniel Zeman

September 24, 2020
ACL Special Interest Group on Typology
https://sigtyp.github.io/
(founded 2020)
Number of Genders
192 WALS Features

- 1A Consonant Inventories
- 2A Vowel Quality Inventories
- ...
- 20A Fusion of Selected Inflectional Formatives
- 26A Prefixing vs. Suffixing in Inflectional Morphology
- 27A Reduplication
- 28A Case Syncretism
- ...
- 81A Order of Subject Object and Verb
- 85A Order of Adposition and Noun Phrase
- 107A Passive Constructions
- 120A Zero Copula for Predicate Nominals
- ...
- 129A Hand and Arm
- 131A Numeral Bases
- 138A Tea
7 General (Non-linguistic) Features

- WALS language code: cze
- Language name: Czech
- Family: Indo-European
- Genus: Slavic
- Latitude: 50.0
- Longitude: 15.0 ... somewhere east of Kouřim :-)
- Country codes: CZ
Gender in WALS

- Lexical category of nouns
- Agreement or cross-reference elsewhere:
  - Pronouns
  - Adjectives, determiners (inflection)
  - Verbs (inflection)
  - ... or a subset thereof

- Data:
  - Ukrainian and Russian: 3 genders (not 4, with animacy)
  - Czech and Slovak not shown at all
  - English: 3 genders; although only in pronouns!
Sparse Data

- 2662 languages in WALS (only 1357 used in the shared task)
- 7 general features
- 192 linguistic features (only 185 in the shared task)
2662 languages in WALS (only 1357 used in the shared task)
7 general features
192 linguistic features (only 185 in the shared task)

Within the shared task training data:
Features per language
- maximum (English): 159 linguistic features
- median: 28 linguistic features
- minimum (2 languages): 4 linguistic features
Sparse Data

- 2662 languages in WALS (only 1357 used in the shared task)
- 7 general features
- 192 linguistic features (only 185 in the shared task)

**Within the shared task training data:**

**Features per language**
- maximum (English): 159 linguistic features
- median: 28 linguistic features
- minimum (2 languages): 4 linguistic features

**Languages per feature**
- maximum (87A Order of Object and Verb): 785 languages
- median: 168 languages
- minimum (2 features): 8 languages
The Shared Task

- How well could the missing feature values be guessed?

- 1125 training languages
- 83 development languages
- 149 test languages

- Some features (about 50% randomly picked) are masked: the value is ?

- Task: based on the remaining features in this or other languages, predict the values of the masked features
The ÚFAL Submission

- Probabilistic System
- Neural System
- Combined System

- + a number of other dead-end attempts
Greenberg Universals

- Greenberg (1963): there are correlations
- e.g. universal (17):
  with overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun
Greenberg Universals

- Greenberg (1963): there are correlations
- e.g. universal (17):
  with overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun

- WALS features:
  - 81A Order of Subject, Object and Verb
    - = 3 VSO
  - 87A Order of Adjective and Noun
    - = 2 Noun-Adjective
Greenberg Universals

- Greenberg (1963): there are correlations
- e.g. universal (17):
  with overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun

- WALS features:
  - 81A Order of Subject, Object and Verb
    - = 3 VSO
  - 87A Order of Adjective and Noun
    - = 2 Noun-Adjective
  - In training data (given that 81A = 3 VSO):
    - 54 languages have both features filled
    - 28 languages (52%) = 2 Noun-Adjective
    - 19 languages (35%) = 1 Adjective-Noun
    - 7 languages (13%) = 3 No dominant order
Greenberg Universals

- Greenberg (1963): there are correlations
- e.g. universal (17):
  *with overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun*

WALS features:
- 81A Order of Subject, Object and Verb
  - $= 3$ VSO
- 87A Order of Adjective and Noun
  - $= 2$ Noun-Adjective
- In training data (regardless of 81A):
  - 713 languages have 87A filled
  - 455 languages (64%) = 2 Noun-Adjective
  - 200 languages (28%) = 1 Adjective-Noun
  - 53 languages (7%) = 3 No dominant order
  - 5 languages (1%) = 4 Only internally-headed relative clauses
Probabilistic System

\[ \text{score}(s_i = x, t_j = y) = P(t_j = y|s_i = x) \times \log c(s_i = x, t_j = y) \times I(s_i, t_j) \] (1)

- Single best signal (no voting among source features)
- Country codes: ignore \textit{US}
- Latitude and longitude: group into zones
Neural System

- **Language embeddings** based on their feature values
- Latitude and longitude: cluster the points via k-Means, use cluster id as feature
- Training neural network:
  - Pick language
  - Pick feature value (50% probability that it belongs to the language)
  - Goal of the network: predict that a feature value belongs to the language
- Prediction
  - For a language and a masked feature:
  - Pass all possible values to the network
  - Pick the value with the highest output probability
Combined System

- Development data:
  - Neural system is slightly better (74.49% > 73.81%)
  - The systems make different errors (oracle → 81%)

- System-internal confidence scores:
  - Probabilistic system → cond. prob. × log count × mutual info
  - Neural system → output feature prob. from the network

- Empirically found thresholds $T_N$ and $T_P$
  - If neural confidence $> T_N$, use neural system
  - Else: if probabilistic confidence $> T_P$, use probabilistic system
  - Else use again neural system
### Table 1: Accuracy of various models on the development and test data.

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>53.45</td>
<td>51.39</td>
</tr>
<tr>
<td>Probabilistic</td>
<td>73.81</td>
<td>71.08</td>
</tr>
<tr>
<td>Neural</td>
<td>74.49</td>
<td>69.80</td>
</tr>
<tr>
<td>Combined</td>
<td>75.50</td>
<td>70.75</td>
</tr>
<tr>
<td>Feed-forward</td>
<td>56.45</td>
<td></td>
</tr>
<tr>
<td>kNN-Hamming</td>
<td>62.28</td>
<td></td>
</tr>
<tr>
<td>kNN-LangEmbed</td>
<td>68.10</td>
<td></td>
</tr>
</tbody>
</table>

*Results System Test Baseline 53.45 51.39 Probabilistic 73.81 71.08 Neural 74.49 69.80 Combined 75.50 70.75 Feed-forward 56.45 kNN-Hamming 62.28 kNN-LangEmbed 68.10*
SIGTYP Shared Task 2020 Rankings

Task: Constrained
1. ÚFAL
2. NEMO_system2
3. NEMO_system1

Task: Unconstrained
1. CrossLingference
4. Panlingua_rule
5. Panlingua_hybrid
6. Panlingua
7. baseline_frequency
8. baseline_knn-imputation
9. NUIG
Thanks!
Díky!