SUKI@VARDIAL2018

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén
THREE SHARED TASKS
ILI AND DFS

• Indo-Aryan Language Identification (ILI)
  • 5 closely related languages from the Indo-Aryan family
    – 10,000 to 18,000 lines of training and development data per language
    – 1,500 to 2,200 lines of test data per language
  • Open task

• Discriminating between Dutch and Flemish in Subtitles (DFS)
  • 2 varieties
    – 300,000 lines for training, 500 for development, and 20,000 for testing
    – Randomly selected lines from the SUBTIEL corpus
  • Open task
THREE SHARED TASKS
GDI

• (Swiss) German Dialect Identification (GDI)
  • 4 dialects
    – training and development data from 4-7 documents per dialect
    – Test data from 1-2 documents per dialect
  • 5th unknown dialect in test set
  • Closed task
• All our experiments and results in all three tasks are as in closed track, no other sources or information were used than training, development and test sets.
THE HELI METHOD OVERVIEW

- Language identification method for monolingual texts
- Each language is represented by several different language models
- Language models consist of probabilities of words and character n-grams, possibly utilizing different pre-processing schemes
  - Original words vs. Lowercased words
- Texts are tokenized into words using non-alphabetic and non-ideographic characters as delimiters
THE HELI METHOD
THE BACKOFF FUNCTION

• The text to be identified is processed one word at a time
• For each word the most specific language model is tried first
  • The most specific model is the one with the largest possible feature space
• If it cannot be applied the method backs off to more general language models
  • from words to higher order n-grams
  • from higher order n-grams to lower order n-grams.
THE HELI METHOD
CALCULATING THE SCORES

• A score is calculated for each word in the mystery text for each language known by the identifier
• Each word and character n-gram in a language model is associated with a value
• The values are negative logarithms of the relative frequencies in the training data
• If word-based language models are used for scoring the words, the scores are simply the values from the models
• When character n-grams are used for scoring a word, the score is the average of the values of the character n-grams
• If none of the character n-grams of a word are found from any of the language models, the method backs off to lower order n-grams
THE HELI METHOD
SCORING THE MYSTERY TEXT

• The whole mystery text gets the score equal to the average of the scores of the words
• The language having the lowest score is returned as the language with the maximum probability for the mystery text
THE HELI METHOD
CHOOSING THE LANGUAGE MODELS

The models to be used are decided upon their performance on the development data

- DFS
  - words and character n-grams from 1 to 8 (both original and lowercased)
- GDI
  - lowercased character 4-grams
- ILI
  - lowercased character n-grams from 1 to 6
THE HELI METHOD
THE RESULTS

• DFS
  • Macro F1: 0.613 (9th, 0.047 behind the best)
• GDI
  • Macro F1: 0.639 (“4th”, five best teams within 0.008)
• ILI
  • Macro F1: 0.887 (“4th”, three best teams within 0.015)
• Could we use the test data itself to improve the language models used by the identifier?

• Could we first identify the easy parts of an out-of-domain text and use them to improve the models in order to identify the more difficult parts?

• The method used is recursive and builds on the fact that we can process the same mystery text collection several times before providing the final labels

• New language material is incorporated into the language models while the mystery text collection is processed
• The development set is used to find the best parameters for the HeLI method
• The language of each line of the test data is identified using the HeLI method
• For each identification, a confidence score is calculated
  • The confidence score is the difference between the scores of the best and the second best identified language
• The line with the most confident identification is selected and the words and character n-grams from the line are added to the language models of the identified language
• The processed line is set as identified and the rest of the lines are re-identified using the adapted language models
LANGUAGE MODEL ADAPTATION
ADDING TO THE LANGUAGE MODELS

• Small modification was needed to the original HeLI implementation
  • The relative frequencies are now always re-calculated at the identification phase
  • The language models consist of counts of individual words and character n-grams as well as the total number of units of each type
• When new information is added to the model, the count of respective units and unit types is increased and new units introduced if needed
  • Very quick with generative language identification methods as there is no need to modify the models of other languages
  • It has not yet been tested with other LI methods than HeLI
LANGUAGE MODEL ADAPTATION
THE RESULTS

• DFS (in-domain)
  • Macro F1: 0.611 (“9th”, 0.002 decrease from basic HeLI)
• GDI (out-of-domain)
  • Macro F1: 0.686 (1st, 0.047 increase from basic HeLI)
• ILI (out-of-domain?)
  • Macro F1: 0.955 (“1st”, 0.068 increase from basic HeLI)
The adaptation process can be repeated.

In tests using the development sets, the additional accuracy gained was usually very small.

Used only in ILI submission, where it increased the F1-score to 0.9576 (0.0023 increase) using four epochs.
CONCLUSIONS

• Language model adaptation would seem to be especially usable in situations where the training material can be expected to be from a different domain than the material to be identified.

• The adaptation method proved to be very robust as it performed well even with the extra unknown dialect present in the GDI test set.
# RESULTS

## GDI

<table>
<thead>
<tr>
<th>System</th>
<th>F1 (macro)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HeLI with adaptive language models, run 2</td>
<td>0.6857</td>
</tr>
<tr>
<td>benf</td>
<td>0.6464</td>
</tr>
<tr>
<td>safina</td>
<td>0.6449</td>
</tr>
<tr>
<td>taraka_rama</td>
<td>0.6398</td>
</tr>
<tr>
<td><strong>The basic HeLI method, run 1</strong></td>
<td><strong>0.6386</strong></td>
</tr>
<tr>
<td>LaMa</td>
<td>0.6374</td>
</tr>
<tr>
<td>XAC</td>
<td>0.6336</td>
</tr>
<tr>
<td>GDI.classification</td>
<td>0.6203</td>
</tr>
<tr>
<td>dkosmajac</td>
<td>0.5909</td>
</tr>
<tr>
<td>Random Baseline</td>
<td>0.2521</td>
</tr>
</tbody>
</table>

## ILI

<table>
<thead>
<tr>
<th>Method (or team)</th>
<th>F1 (macro)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HeLI with iterative language model adaptation (run3)</strong></td>
<td><strong>0.9576</strong></td>
</tr>
<tr>
<td>HeLI with language model adaptation (run2)</td>
<td>0.9553</td>
</tr>
<tr>
<td>taraka_rama</td>
<td>0.9022</td>
</tr>
<tr>
<td>XAC</td>
<td>0.8933</td>
</tr>
<tr>
<td>ILIIdentification</td>
<td>0.8895</td>
</tr>
<tr>
<td><strong>HeLI (run1)</strong></td>
<td><strong>0.8873</strong></td>
</tr>
<tr>
<td>safina</td>
<td>0.8627</td>
</tr>
<tr>
<td>dkosmajac</td>
<td>0.8472</td>
</tr>
<tr>
<td>we_are_indian</td>
<td>0.8360</td>
</tr>
<tr>
<td>LaMa</td>
<td>0.8195</td>
</tr>
<tr>
<td>Random Baseline</td>
<td>0.2024</td>
</tr>
</tbody>
</table>