Contents

1. Introduction
2. Previous work on L1 identification
3. Our baseline point: surface-based classification
   3.1. Features
   3.2. Method
3.3. Results on ICLEv2 and FALKO
4. Future work: Towards more linguistic modeling
5. References
Introduction
Introduction

Corpus in L2 = German

L1 = English
L1 = French
L1 = Russian

German
German
German
German

4
1. Introduction
2. Previous work on L1 identification
3. Our baseline approach: surface-based classification
   3.1. Features
   3.2. Method
   3.3. Results on ICLEv2 and FALKO
4. Future work: Towards more linguistic modeling
5. References
Previous work: Wong & Dras (2009)

Corpus: 665 ICLEv2 essays

Features:

- 3 error types (subject-verb disagreement, misuse of determiners, noun-number disagreement)
- 70/363/398 function words
- 300 letter n-grams, \( n \in \{1, 2, 3\} \)
- 450 POS n-grams, \( n \in \{2, 3\} \)

Method: SVM, 70 essays for training, 25 for testing

Result: 73.7% accuracy (combi)
Contents

1. Introduction
2. Previous work on L1 identification
3. Our baseline approach: surface-based classification
4. Future work: Towards adding linguistic modeling
5. References
Our baseline approach: Features

- n-grams of all occurring lengths, 2 ≤ n ≤ max-n(corpus)
- All n-grams occurring in ≥ 2 texts of the used corpus
  - n=6: die Studenten auf die wirkliche Welt
  - n=5: die Studenten auf die wirkliche Welt
  - n=4: was mich betrifft, von geringen Wert
  - n=3: was mich betrifft, von geringen Wert
  - n=2: und zwar, 30 Jahre

Examples (from FALKO):

- Features used: word-based recurring n-grams

...
Our baseline approach:

**Method:** Machine Learning

- **k-NN**, different distance metrics (Cosine, Dot Product)

**Metrics best for sparse vectors:**

**Testing:** leave-one-out

**Features:** as bit vectors (0=feature absent, 1=present)

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 3</th>
<th>Feature 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Feature bit vector:**

- Feature bit vector

- Features: as bit vectors (0=feature absent, 1=present)

- Testing: leave-one-out

- Cosine, Dot Product metrics best for sparse vectors

- Machine Learning: k-NN, different distance metrics
Baseline approach: ICLEv2 task

Replication of Wong & Dras (2009), i.e., we use the same dataset, but our own features & machine learning setup:

- Corpus: ICLEv2
- Feature set: word-based recurring n-grams:
  - Chinese, Japanese essays = 665 essays
- Seven L1 (Bulgarian, Czech, French, Russian, Spanish, ...)

Corpus: ICLEv2

Dataset, but our own features & machine learning setup: Replication of Wong & Dras (2009), i.e., we used same
### Baseline Approach: ICLEV2 Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>Distance</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>72.0%</td>
<td>97.14%</td>
</tr>
<tr>
<td>Dot Product</td>
<td>73.1%</td>
<td>3.97%</td>
</tr>
<tr>
<td>Cosine</td>
<td>73.7%</td>
<td>73.12%</td>
</tr>
<tr>
<td>Dot Product</td>
<td>74.3%</td>
<td>74.18%</td>
</tr>
<tr>
<td>Cosine</td>
<td>74.9%</td>
<td>75.02%</td>
</tr>
<tr>
<td>Dot Product</td>
<td>77.94%</td>
<td>77.69%</td>
</tr>
</tbody>
</table>

*Percentages*:
- 72.0% baseline accuracy
- 73.1% accuracy
- 73.7% accuracy
- 74.3% accuracy
- 74.9% accuracy

*Settings*:
- [2’, 29’]
- [2’, 3’]
- [2’, 3’]
- [2’, 3’]
- [2’, 3’]

*Type of n*:
- Success, low n drop
- Single n
- Success, high n drop
- Picked n
- Setting type of n

*Rank*:
- 5
- 4
- 3
- 2
- 2
- 2
- 1
Baseline approach: ICLEv2 results

Confusion matrix for the best result
Baseline approach: FALKO setup

Corpus: FALKO
– Subset with 6 L1 (Rus, Uzb, Fra, Eng, Dan, Tur) x 10 essays = 60 essays
– Feature set: recurring n-grams:
  – intervals $[2, n]$, $n \in [2, 6]$
  – Exploration of some other n-gram subsets

Corpus: FALKO
Baseline approach: FALKO results

Best result: 63.3%, n = 2 (baseline

Word based n-grams
Baseline approach: FALKO results

Part-of-speech based n-grams
Word + open class (N.*, VV.*, ADJ.*, CARD classes) n-grams:

Baseline approach: FALKO results
Baseline approach: FALKO results

Word + open class POS (matching N.*, VV.*, ADJ.*, CARD):

Best result: 53.3%, n := [2, 5] (baseline #16.7%)

Graphs showing features and accuracy %.
Baseline approach: FALKO results

Word + ADJ. POS (ADJA, ADJD):

- Best result: 56.7%, n = 2 (baseline #16.7%)
Baseline approach: FALKO results

- Word + VV.* POS (VVFIN, VVIMP, VVINF, VVIZU, VVPP):

  - Best result: 53.3%, n = 2 (baseline #16.7 %)

```
<table>
<thead>
<tr>
<th>Features #</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2551</td>
<td>3699</td>
<td>3981</td>
<td>4090</td>
<td>4165</td>
</tr>
<tr>
<td>2000</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>4000</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>6000</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>8000</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>10000</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
</tr>
</tbody>
</table>
```

Accuracy %

Word + VV.* POS (VVFIN, VVIMP, VVINF, VVIZU, VVPP):
Baseline approach: FALKO results

Word + N.* POS (NN, NE)

Best result: 56.7%, n := \([2, 3]\) (baseline #16.7%), but: 58.3% u using n := \([2, 3, 6]\)
Baseline approach: FALKO results

Best results (accuracy baseline ≈ 16.7%)

Word based:
- \( n = 2 \), \( 3', 6 \), cosine, 4236 feat. (max. 5663):
  - \( N.* \), \( n \) subset \{2, 3', 6\}, cosine, 4236 feat. (max. 5663):
    - 58.3% accuracy

POS based:
- Word + open class POS based:
  - \( N.*, \ ADJ.*, VV.* \), \( n \) interval \( [2, 5] \), cosine, 7530 feat. (max. 12246):
    - 53.3% accuracy

POS based:
- \( n \) interval \( [2, 4] \), cosine, 6560 feat. (max. 12246):
  - 46.7% accuracy

POS based:
- \( n = 2 \) (single \( n \)), cosine, 2367 feat. (max. 3801):
  - 63.3% accuracy

Baseline approach: FALKO results
Contents

1. Introduction
2. Previous work on L1 identification
   3. Our baseline approach: surface-based classification
   3.1. Features
   3.2. Method
   3.3. Results on ICLEv2 and FALKO
3. Towards adding linguistic modeling
4. Future work: Towards adding linguistic modeling
5. References
Towards more linguistic modeling

Features: from surface to more linguistic modeling

• Modeling on different levels of abstraction:
  words, POS, lemmas, induced classes, ...

• Modeling on different levels of units: phrases, ...

Data Mining techniques

Evaluation method: Use of other Machine Learning and

...
<table>
<thead>
<tr>
<th>capacity</th>
<th>passenger</th>
<th>standard</th>
<th>living</th>
<th>adverse</th>
<th>proper</th>
<th>employment</th>
<th>prospects</th>
</tr>
</thead>
<tbody>
<tr>
<td>large</td>
<td>part</td>
<td>part</td>
<td>source</td>
<td>whole</td>
<td>whole</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>of</td>
<td>of</td>
<td>of</td>
<td>of</td>
<td>of</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example: L1=Chinese, L2=Russian

Example: Choice of Adj N vs. N typical?

Towards more linguistic modeling
References


Thank you for your attention!
Previous work

Corpus: ICLEV1, 5 L1 x 258 essays = 1290 essays

Features:
- 250 POS bi-grams
- 185 error types
- 200 char n-grams
- 400 function words

Method: SVM, 10-fold-cross-validation

Result: 80.2% accuracy (combi)

Koppel / Schler / Zigdon 2005;
Previous work

Tsur / Rappoport 2007;

Corpus: ICLEv1, 5 L1 x 258 essays = 1290 essays

Features:

– char n-grams, n:={1', 2', 3'}
– 460 function words

Motivation: Influence of syllable structure of L1 on the L2 lexis

Method: SVM, 10-fold-cross-validation

Result: 65.6% accuracy (bi-grams)