Faking Errors to Avoid Making Errors: Very Weakly Supervised Learning for Error Detection in Writing

Jonas Sjöbergh and Ola Knutsson

September 6, 2005
Goal

- Find errors in text
- Languages with few resources
- Little manual work
Overview

- Treat grammar checking as tagging (OK/ERROR)
- Generate errors automatically
- Use off the shelf machine learning
Example

I  OK
enjoy  OK
watching  OK
a  ERROR
movies  OK
.  OK
Why machine learning?

- Writing rules takes time
- Manual work is expensive
- Machine learning complements linguists
Why generate errors?

- Error free text easily available
- Easier to corrupt than correct
- A lot of work to annotate real errors
- Generated errors trivial to annotate
- Correct contexts mitigates unrealistic generation
More details

- Collect text with no (few) errors
- Write simple program to generate errors and mark as ERROR
- Mark all other words as OK
- Train machine learner
- Finished
Example generator program

(1) Read lemma lexicon (or stems)
(2) Read PoS-tags with agreement constraints
(3) Run PoS-tagger
(4) For each tagged sentence:
(5) Pick random word with agreement constraint
(6) Get lemma (lexicon)
(7) Get random word with this lemma (lexicon)
(8) If not exact same word:
(9) Change word, mark as error

30 minutes vs. 1,000 hours
Many generated “errors” will not lead to ungrammatical sentences.

This is not a problem, since there will also be many examples of grammatical text.

Non-errors will (generally) not be learned by machine learner.
What are split compounds?

“quick sand” or “quicksand”

“en lånhårig sjukgymnast” = “a physical therapist with long hair”
“en lång hårig sjuk gymnast” = “a tall, hairy and sick gymnast” (still grammatical)

“ett personnummer” = “social security number”
“ett person nummer” = “one person number” (ungrammatical, agreement)

Hard to detect, handled by few existing grammar checkers. Well suited to our method.
How to generate

Simple:
- Word is made of two short words in corpus?
- If so, split into these and mark with ERROR
- Else, no change and mark as OK

Slightly better:
- We use spelling checker, which does this but handles inflections
Machine learning features

- Word n-grams
- PoS n-grams (150 tags, statistical tagger)
- ERROR/OK n-grams
- combinations of these

Most useful was PoS n-grams
A small improvement

Observation: all split compounds have at least 2 parts

Idea: filter all alarms with no neighboring ERROR

Slightly better: use spelling checker on neighboring ERRORs (also removes detections of misspelled split compounds)
What are agreement errors?

In Swedish, for instance noun, adjective and determiner have to agree in gender, number etc.

Similar to “I play the baseball.”

Precision relatively easy, recall harder.

Best covered error type of existing grammar checkers.
How to generate

- Select random word from class with agreement restrictions
- Check in lexicon for other inflectional form of word
- Output other form, mark as ERROR
- Mark all other words as OK
Machine learning features

- gender n-grams (if applicable)
- number n-grams (if applicable)
- definiteness n-grams (if applicable)
- PoS n-grams
- ERROR/OK n-grams
- combinations of these

Most useful was PoS + gender n-grams
## Split compounds

<table>
<thead>
<tr>
<th></th>
<th>Gr.</th>
<th>SnålGr. + filter</th>
<th>SnålGr. + filter</th>
<th>Basel. + filter</th>
<th>Basel. + filter</th>
<th>Union</th>
<th>Intersection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detected</td>
<td>322</td>
<td>588</td>
<td>535</td>
<td>331</td>
<td>120</td>
<td>582</td>
<td>275</td>
</tr>
<tr>
<td>False neg</td>
<td>490</td>
<td>224</td>
<td>277</td>
<td>481</td>
<td>692</td>
<td>230</td>
<td>537</td>
</tr>
<tr>
<td>False pos</td>
<td>6</td>
<td>49</td>
<td>24</td>
<td>162</td>
<td>6</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>Precision</td>
<td>98%</td>
<td>92%</td>
<td>96%</td>
<td>67%</td>
<td>95%</td>
<td>95%</td>
<td>100%</td>
</tr>
<tr>
<td>Recall</td>
<td>40%</td>
<td>72%</td>
<td>66%</td>
<td>41%</td>
<td>15%</td>
<td>72%</td>
<td>34%</td>
</tr>
</tbody>
</table>

Baseline: more common in compound than as a word

Union: Granska or our method
Intersection: Granska and our method

Word (75) and ProbGranska (225) do not look for split compounds
### Agreement errors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Detected</td>
<td>71</td>
<td>17</td>
<td>101</td>
<td>88</td>
<td>100</td>
<td>134</td>
<td>54</td>
</tr>
<tr>
<td>False neg</td>
<td>155</td>
<td>-</td>
<td>125</td>
<td>138</td>
<td>126</td>
<td>92</td>
<td>172</td>
</tr>
<tr>
<td>False pos</td>
<td>1</td>
<td>-</td>
<td>5</td>
<td>15</td>
<td>143</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>Precision</td>
<td>99%</td>
<td>-</td>
<td>95%</td>
<td>85%</td>
<td>41%</td>
<td>88%</td>
<td>98%</td>
</tr>
<tr>
<td>Recall</td>
<td>31%</td>
<td>-</td>
<td>45%</td>
<td>39%</td>
<td>44%</td>
<td>60%</td>
<td>24%</td>
</tr>
</tbody>
</table>

Baseline: first consecutive words with feature difference

Union: Granska or our method
Intersection: Granska and our method
<table>
<thead>
<tr>
<th></th>
<th>Word</th>
<th>ProbGr.</th>
<th>Granska</th>
<th>SnålGr.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>All detected</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>All false pos</td>
<td>92</td>
<td>36</td>
<td>35</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>Spelling</td>
<td>8</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>False</td>
<td>89</td>
<td>-</td>
<td>20</td>
<td>-</td>
<td>101</td>
</tr>
<tr>
<td>Grammatical</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>False</td>
<td>3</td>
<td>36</td>
<td>15</td>
<td>50</td>
<td>99</td>
</tr>
<tr>
<td>Agreement</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Split compounds</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Hard to reach high precision with few existing errors.
### Learner essays 10,000 words

<table>
<thead>
<tr>
<th></th>
<th>Word</th>
<th>ProbGr.</th>
<th>Granska</th>
<th>SnålGr.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>All detected</td>
<td>392</td>
<td>101</td>
<td>411</td>
<td>122</td>
<td>592</td>
</tr>
<tr>
<td>All false pos</td>
<td>21</td>
<td>19</td>
<td>13</td>
<td>19</td>
<td>67</td>
</tr>
<tr>
<td>Spelling</td>
<td>334</td>
<td>34</td>
<td>293</td>
<td>26</td>
<td>362</td>
</tr>
<tr>
<td>False</td>
<td>18</td>
<td>-</td>
<td>5</td>
<td>-</td>
<td>21</td>
</tr>
<tr>
<td>Grammatical</td>
<td>58</td>
<td>67</td>
<td>118</td>
<td>96</td>
<td>230</td>
</tr>
<tr>
<td>False</td>
<td>3</td>
<td>19</td>
<td>8</td>
<td>19</td>
<td>46</td>
</tr>
<tr>
<td>Agreement</td>
<td>32</td>
<td>9</td>
<td>49</td>
<td>43</td>
<td>74</td>
</tr>
<tr>
<td>Split compounds</td>
<td>5</td>
<td>8</td>
<td>20</td>
<td>27</td>
<td>35</td>
</tr>
</tbody>
</table>

Easy to achieve high precision, many errors to detect.
## Introduction

Jonas Sjöbergh and Ola Knutsson

## Our Method

## Evaluation

### Error collections

### Real texts

Student essays 10,000 words

<table>
<thead>
<tr>
<th></th>
<th>Word</th>
<th>ProbGr.</th>
<th>Granska</th>
<th>SnålGr.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>All detected</td>
<td>38</td>
<td>23</td>
<td>48</td>
<td>28</td>
<td>90</td>
</tr>
<tr>
<td>All false pos</td>
<td>31</td>
<td>45</td>
<td>13</td>
<td>31</td>
<td>111</td>
</tr>
<tr>
<td>Spelling</td>
<td>24</td>
<td>3</td>
<td>17</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>False</td>
<td>28</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>28</td>
</tr>
<tr>
<td>Grammatical</td>
<td>14</td>
<td>20</td>
<td>31</td>
<td>27</td>
<td>65</td>
</tr>
<tr>
<td>False</td>
<td>3</td>
<td>45</td>
<td>13</td>
<td>31</td>
<td>83</td>
</tr>
<tr>
<td>Agreement</td>
<td>5</td>
<td>0</td>
<td>11</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>Split compounds</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Contains many quotes, from “strange” domains. Problematic for statistical methods.
Conclusions

- Very little manual work
- Very few NLP resources (PoS tagger, possibly spelling checker)
- Unannotated text
- Not as good as state of the art
- Good enough to be useful
- Complements traditional checkers