Computer-assisted stylistic revision with incomplete and noisy feedback
A pilot study

Christian M. Meyer and Johann Frerik Koch

The 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)
June 16, 2016. San Diego, CA, USA.
Vision

• Bryant & Ng (2015): best grammar correction software achieves only 73% of human performance

• Our vision: research new useful approaches to intelligent writing assistance with a focus on German native speakers
Goal of this work

There won’t be perfect systems! 😞

How do users deal with incorrect and incomplete feedback?

• Pilot user study
• German L1 text revision task
• focus on stylistic issues
Previous Work

Data-driven evaluation
• Shared tasks: HOO, CoNLL, BEA,…
• meaningful system comparison?
• interpretation of evaluation metrics?
• reliability of the reference data?

User-driven evaluation
• (Manual) feedback by teachers and peers
• Variation of feedback granularity, extent & formulation, time
• Nagata & Nakatani (2010): “precision-oriented error detection is better than recall-oriented”
Hypotheses

**H1** If users receive **correct feedback**, they will more likely revise the corresponding section
Hypotheses

**H1**  
If users receive **correct feedback**, they will more likely revise the corresponding section

**H2**  
If users receive **incorrect feedback**, they will more likely revise the corresponding section  
– although it would not be necessary
Hypotheses

H1 If users receive **correct feedback**, they will more likely revise the corresponding section

H2 If users receive **incorrect feedback**, they will more likely revise the corresponding section – although it would not be necessary

H3 If users receive **incomplete feedback**, they will more likely miss issues not highlighted to them

I look forward your response.

A

to
Hypotheses

H1  If users receive **correct feedback**, they will more likely revise the corresponding section

H2  If users receive **incorrect feedback**, they will more likely revise the corresponding section – although it would not be necessary

H3  If users receive **incomplete feedback**, they will more likely miss issues not highlighted to them

H4  Providing automatic feedback does not affect the required **time to complete the task**
Experimental Setup: Data

$T_1$ News item
206 words

$T_2$ Wikipedia article
183 words
Experimental Setup: Data

11 text positions 8 introduced issues
Experimental Setup: Data

TP: correct feedback
FP: incorrect feedback
FN: incomplete feedback
### Experimental Setup: Data

<table>
<thead>
<tr>
<th>Issue</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Collocation</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Variation</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Experimental Setup: Population

Experimental Group ($N = 15$) w/ feedback

Control Group ($N = 11$) w/o feedback

within-subject variable

between-subject variable

$T_1$

$T_2$
Experimental Setup: Tool

**Experimental Group** (N = 15) w/ feedback

**Control Group** (N = 11) w/o feedback

**User Study**

New tool: InViEdit
https://github.com/UKPLab/naacl-bea2016-writing-study
Writing Assistance Software

https://github.com/UKPLab/naacl-bea2016-writing-study

System Usability Scale
SUS = 76.3
> 68.0 “acceptable”
> 71.4 “good”
Experimental Setup: Analysis

**User Study**

- **Experimental Group** ($N = 15$) w/ feedback
- **Control Group** ($N = 11$) w/o feedback

**New tool: InViEdit**

https://github.com/UKPLab/naacl-bea2016-writing-study

- Revised vs. not revised positions
- Revised vs. not revised positions
Data Analysis

11 positions (TP/FP/FN) 
× 26 participants 
= 286 data points

Data point $x = (\text{revised vs. not revised})$

<table>
<thead>
<tr>
<th></th>
<th>min($x$)</th>
<th>$\bar{x}$</th>
<th>SE</th>
<th>max($x$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EG</td>
<td>2</td>
<td>5.86</td>
<td>0.53</td>
<td>10</td>
</tr>
<tr>
<td>CG</td>
<td>0</td>
<td>3.18</td>
<td>0.74</td>
<td>8</td>
</tr>
</tbody>
</table>

Unpaired two sample Student’s $t$ test  
with significance level $\alpha = 0.05$  ($P \leq 0.05$)
H1: Correct Feedback helps

**Expectation:** $\mu_{\text{EG(TP)}} \neq \mu_{\text{CG(TP)}}$

**Arithmetic mean:**
\[ \bar{x}_{\text{EG(TP)}} = 4.13 \quad (\text{SE} = 0.23) \]
\[ \bar{x}_{\text{CG(TP)}} = 1.63 \quad (\text{SE} = 0.51) \]

**Test statistic:**
\[ t_{H1} = 4.85 \]
\[ |t_{H1}| > 2.06 \quad (P < 0.0001) \]

• reject null hypothesis at 5% level
• significant difference b/w groups
H2: Incorrect Feedback causes unnecessary revisions

**Expectation:** $\mu_{EG(FP)} \neq \mu_{CG(FP)}$

**Arithmetic mean:**
- $\bar{x}_{EG(FP)} = 1$ (SE = 0.25)
- $\bar{x}_{CG(FP)} = 0.18$ (SE = 0.12)

**Test statistic:**
- $t_{H2} = 2.55$
- $|t_{H2}| > 2.06$ ($P = 0.017$)

- reject null hypothesis at 5% level
- significant difference b/w groups
H3: Incomplete Feedback causes users to miss similar issues

Expectation: $\mu_{EG(FN)} \neq \mu_{CG(FN)}$

Arithmetic mean:
$\bar{x}_{EG(FN)} = 0.73$ (SE = 0.28)
$\bar{x}_{CG(FN)} = 1.36$ (SE = 0.36)

Test statistic:
$t_{H3} = -1.39$
$|t_{H3}| < 2.06$ ($P = 0.17$)

• cannot reject null hypothesis
• cannot find significant difference ×
**H4: Task completion time similar**

**Expectation:** $\mu_{EG(\tau)} = \mu_{CG(\tau)}$

**Arithmetic mean:**

$\bar{x}_{EG(\tau)} = 13 \text{ min, } 3 \text{ sec (SE = 104 sec)}$

$\bar{x}_{CG(\tau)} = 13 \text{ min, } 27 \text{ sec (SE = 144 sec)}$

**Test statistic:**

$t_{H4} = -0.14$

$|t_{H4}| < 2.06 \ (P = 0.89)$

- cannot reject null hypothesis
- cannot find significant difference
Conclusion

What we learned from this work:

• correct feedback helps (H1)
• incorrect feedback problematic (H2) – overtrust?
  • but: demand for adaptive systems!
• tendency to miss FNs, but not significant (H3)
  • confirms “precision more important than recall”
• using feedback didn’t take longer (H4)

Software for evaluating writing assistance tools:
https://github.com/UKPLab/naacl-bea2016-writing-study
Kontakt / Contact

Christian M. Meyer
Technische Universität Darmstadt
Ubiquitous Knowledge Processing Lab

✉ Hochschulstr. 10, 64289 Darmstadt, Germany
📞 +49 (0)6151 16–25293
📧 +49 (0)6151 16–25295
✉ meyer (at) ukp.informatik.tu-darmstadt.de

Rechtliche Hinweise

Legal Issues
The slides are intended for personal use by the audience of the talk. Photographies, illustrations, tradedmarks, or logos are property of the holder of rights. To avoid any misconceptions, I would strongly recommend to get in touch before reusing or redistributing the slides or any additional material of the talk. The same applies if you consider your rights infringed – please let me know to initiate further clarification.