Does Automated Feedback in a Proofreading Tool Help an English Language Learner?

Claudia Leacock, Butler Hill Group
Michael Gamon, Microsoft Research
Chris Brocket, Microsoft Research
... and
William B. Dolan, Microsoft Research
Jianfeng Gao, Microsoft Research
Dmitriy Belenko, Microsoft Research
Lucy Vanderwende, Microsoft Research
Alexandre Klementiev, University of Illinois at Urbana Champaign
ESL Assistant

• March 2008: CALICO Workshop: Gamon et al.
  – System Description & Evaluation. No user action.
  – System performance is state-of-the-art

• June 24, 2008: ESL Assistant goes live!

• 2009 CALICO Workshop Presentation
  – System Usage
  – Evaluation
Most frequent errors made by East Asian non-native speakers

**Noun Related**: Articles (inclusion & choice), Noun Number, Noun of Noun

- I think it’s *a/the best way to resolve issues like this.
- Conversion always takes a lot of *efforts/effort.
- Please send the *feedback of customer/customer feedback to me by mail.

**Preposition Related**: inclusion & choice

- It seems ok and I did not pay much attention *on/to it.
- I should *to ask/ask a rhetorical question.

**Verb Related**: Gerund/Infinitive Confusion, Auxiliary Verb Error, Verb Formation Errors (6), Cognate/Verb confusion, Irregular Verbs

- On Saturday, I with my classmate went *eating/to eat.
- Hope you will *happy/be happy in Taiwan.
- I *teached/taught him all the things I know.

**Adjective Related**: Adjective Confusion (4), Adjective Order

- She is very *interesting/interested in the problem.
- So *Korea/Korean Government is intensely fostering trade.
Users and Data Collection
ESL Assistant User Interface

Microsoft Research | ESL Assistant | Check | Download Outlook add-in

I am interested of this movie.

Hover your mouse over suggestions to see examples of use. Click to accept.
Microsoft confidential. Terms of use and privacy notice.

Tell us what you think!

Original phrase « Hide

Web results 1-10 of 4,530
See also: Images, Video, News, Maps, MSN, MSDN

Update on v1.9.3 status and talk of v1.9.1 - YAF Announcements - Yet ...

I'm interested of this: YAFPro Theme
Created by Jabon Cargman (Tiny Gecko)
Powered by Yet Another Forum.net version 1.9.1 (NET v2.0) - 2007/3/19
forum.yetanotherforum.net/yaf_post3784_Update-on-v193-status-and-talk-of-v191.aspx · Cached page

Phrase with suggestion

Web results 1-10 of 9,680,000
See also: Images, Video, News, Maps, MSN, MSDN

Why are you interested in this job? In our company? » Ian Christie...
Ian Christie's Bold Career Blog Insights, ideas, tools and a firm push on your career development.
boldcareer.com/blog/archives/2005/09/15/why_are_you_interested_in_this_job_in_our_company..... · Cached page

Interested in this creative and
Interested in this creative and personally
Page Views per Day

Traffic via website links:
- Windows Live Translator: 35%
- Chinese MSN: 13%
- Taiwan MSN: 11%
- Korean MSN: 7%

Beijing Olympics
Live Translator snafu
## User Location

<table>
<thead>
<tr>
<th>country</th>
<th>visits</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>51,285</td>
<td>26.80%</td>
</tr>
<tr>
<td>United States</td>
<td>28,916</td>
<td>15.10%</td>
</tr>
<tr>
<td>Taiwan</td>
<td>25,753</td>
<td>13.40%</td>
</tr>
<tr>
<td>Korea - South</td>
<td>12,934</td>
<td>6.80%</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>8,826</td>
<td>4.60%</td>
</tr>
<tr>
<td>Brazil</td>
<td>4,648</td>
<td>2.40%</td>
</tr>
<tr>
<td>Canada</td>
<td>3,917</td>
<td>2.00%</td>
</tr>
<tr>
<td>Germany</td>
<td>3,077</td>
<td>1.60%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>2,928</td>
<td>1.50%</td>
</tr>
<tr>
<td>Japan</td>
<td>2,581</td>
<td>1.30%</td>
</tr>
<tr>
<td>Italy</td>
<td>2,579</td>
<td>1.30%</td>
</tr>
<tr>
<td>Spain</td>
<td>2,557</td>
<td>1.30%</td>
</tr>
<tr>
<td>Russian Federation</td>
<td>2,448</td>
<td>1.30%</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>2,021</td>
<td>1.10%</td>
</tr>
</tbody>
</table>
Growth of the Database: Users and Sessions
Repeat users

Return frequency

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Percentage of Total Visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Once only</td>
<td>100%</td>
</tr>
<tr>
<td>2 times or more</td>
<td>0%</td>
</tr>
<tr>
<td>3 times or more</td>
<td>0%</td>
</tr>
<tr>
<td>4 times or more</td>
<td>0%</td>
</tr>
<tr>
<td>5 times or more</td>
<td>0%</td>
</tr>
</tbody>
</table>
Collected Data

Writing Domains: By Number of Sentences

- Email: 53%
- Non-technical: 24%
- Technical: 14%
- Other: 5%
- Unrelated: 5%
Frequent Users (2/10/09)

- Frequent Users: 578
- Sessions: 5,305
- Session-Unique Sentences: 39,944
- Grammatical Error Flags: 17,832
User interactions
Users Examine 87% of Suggestions

**Conclusion:** A significant number of users are inspecting the suggested rewrites and making a deliberate choice to accept it or not accept it.
Do users make the right choices?

To answer, need human evaluation:

• Time consuming, costly
• Inter-rater agreement (Tetreault & Chodorow)

BUT ... necessary for system development

• Single Annotator
• Internally consistent to measure relative performance during system development
To answer:
Do users make the right choices?

• Evaluated user data to date:
  34% of frequent user sessions: 6K flags

• From Evaluated Flags:
  1. Calculate performance for ALL suggestions.
  2. Calculate system performance for ONLY suggestions that were accepted.
  3. Compare ratios of good and bad flags.
## Evaluation Categories

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>SubEval</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>Correct Flag</td>
<td>The correction fixes a problem in the user input.</td>
</tr>
<tr>
<td></td>
<td>Both Good</td>
<td>The suggestion is a legitimate alternative of a well-formed original input. Ex: I like working/to work.</td>
</tr>
<tr>
<td></td>
<td>Misdiagnosis</td>
<td>The original input contained an error but the suggested rewrite neither improves nor further degrades the user input. Ex: If you have fail machine on hand.</td>
</tr>
<tr>
<td>Neutral</td>
<td>Both Wrong</td>
<td>An error type is correctly diagnosed but the suggested rewrite does not correct the problem. Ex: “can you give me ^ suggestion” insert the instead of a</td>
</tr>
<tr>
<td></td>
<td>Non-ascii</td>
<td>A non-ascii or text processing mark-up character is in the immediate context. (Only applies to user data)</td>
</tr>
<tr>
<td>Bad</td>
<td>False Flag</td>
<td>The suggestion resulted in an error or would otherwise lead to a degradation over the original user input.</td>
</tr>
</tbody>
</table>
Error Type: Are users accepting the right suggestions?

<table>
<thead>
<tr>
<th>Noun-related</th>
<th>Prep-related</th>
<th>Verb-related</th>
<th>Adj-related</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All Suggestions</strong></td>
<td><strong>All Suggestions</strong></td>
<td><strong>All Suggestions</strong></td>
<td><strong>All Suggestions</strong></td>
</tr>
<tr>
<td>bad 16%</td>
<td>bad 24%</td>
<td>bad 6%</td>
<td>bad 23%</td>
</tr>
<tr>
<td>neut 28%</td>
<td>neut 39%</td>
<td>neut 32%</td>
<td>neut 32%</td>
</tr>
<tr>
<td>good 56%</td>
<td>good 37%</td>
<td>good 62%</td>
<td>good 45%</td>
</tr>
<tr>
<td><strong>Accepted</strong></td>
<td><strong>Accepted</strong></td>
<td><strong>Accepted</strong></td>
<td><strong>Accepted</strong></td>
</tr>
<tr>
<td>bad 11%</td>
<td>bad 13%</td>
<td>bad 3%</td>
<td>bad 9%</td>
</tr>
<tr>
<td>neut 26%</td>
<td>neut 42%</td>
<td>neut 25%</td>
<td>neut 28%</td>
</tr>
<tr>
<td>good 63%</td>
<td>good 45%</td>
<td>good 72%</td>
<td>good 63%</td>
</tr>
</tbody>
</table>
Domains: Are users accepting the right suggestions?

<table>
<thead>
<tr>
<th>Email</th>
<th>Non-technical</th>
<th>Technical</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Suggestions</strong></td>
<td><strong>Suggestions</strong></td>
<td><strong>Suggestions</strong></td>
</tr>
<tr>
<td>bad 15%</td>
<td>bad 12%</td>
<td>bad 34%</td>
</tr>
<tr>
<td>neutral 32%</td>
<td>neutral 32%</td>
<td>neutral 28%</td>
</tr>
<tr>
<td>good 53%</td>
<td>good 56%</td>
<td>good 38%</td>
</tr>
<tr>
<td><strong>Accepted</strong></td>
<td><strong>Accepted</strong></td>
<td><strong>Accepted</strong></td>
</tr>
<tr>
<td>bad 9%</td>
<td>bad 10%</td>
<td>bad 19%</td>
</tr>
<tr>
<td>neutral 28%</td>
<td>neutral 34%</td>
<td>neutral 29%</td>
</tr>
<tr>
<td>good 63%</td>
<td>good 56%</td>
<td>good 52%</td>
</tr>
</tbody>
</table>
What do users do with neutral flags?

Neutral Categories: “both wrong” and “misdiagnosis” 78% of neutral flags

- Misdiagnosis 64%
- Both ok 15%
- Both wrong 14%
- Non ascii 7%

Inspect >15.5K Flags to Accept 6.4K
Neutral Flags not accepted but sentence edited to produce no flag

- I don't know that you knew or not, this early morning I got a from head office ...
  - suggestion: delete “from”
  I don't know that you knew or not, this early morning I heard from the head office ...

- Please play with the software and Friday I will be by to work with any questions you may regarding it.
  - suggestion: regarding → regard
  Please play with the software and Friday I will be by to work with any questions you may have regarding it.

From 1,349 sentences with neutral flags found 215 subsequently submitted “similar” strings with no error flag. Users not accept suggestion but did something ELSE to make the flag go away.
Users improve 40% of the time

Not Accept Suggestion but Revise Sentence
Revise and not improve 16%
Typed in suggestion 44%
Revise and improve 40%

Identifying the location of an error can help the user.
Conclusions

• Traffic: There is an interest in ESL proofing tools
• Even current state-of-the-art error correction can be useful for ELLs:
  ➢ Users do not accept proposed corrections blindly – they are selective in their behavior
  ➢ Users make informed choices – they can distinguish correct suggestions from incorrect ones
  ➢ Sometimes just identifying the location of an error enables the users to repair the problem themselves
www.eslassistant.com