

IES
Institute of Education Sciences

INSTITUTE FOR INTELLIGENT SYSTEMS
Department of Psychology

THE UNIVERSITY OF MEMPHIS

Spelling Mistacks & Typos: Can Your ITS Handle Them?

Adam M. Renner^a
Philip M. McCarthy^b
Chutima Boonthum^c
Danielle S. McNamara^a

^aUniversity of Memphis, Psychology / Institute for Intelligent Systems
^bUniversity of Memphis, English / Institute for Intelligent Systems
^cHampton University, Computer Science

THE UNIVERSITY OF MEMPHIS

Intelligent Tutoring Systems

- Provide assessment of user input
- Guided feedback based on user's response
- Many ITSs use conversational dialogue
- NLP for assessment and determines feedback
 - Input matched to benchmark
 - Assessed for similarity
- Assessment limited to proficiency of user
 - High school students or younger
 - Make typing errors/spelling mistakes
- What the student *intended*

THE UNIVERSITY OF MEMPHIS

ITS User-Language

- Contains high rate of typographical & grammatical errors
 - Not a new issue in NLP
- Traditional spellchecking not suitable (e.g., MS Word, email)
- ITSs necessitate **automatic** corrections
 - Why2-Atlas (VanLehn et al., 2002)
 - CIRCSIM-Tutor (Elmi & Evens, 1998)
 - Many more just **ignore** errors
- NLP tools thought resistant to errors
 - LSA (Landauer et al., 2007) – semantic overlap across two **whole** texts
 - Short responses?
 - Responses with multiple errors?
 - NLP tools trained on edited text
 - When used in ITS, similarity assessment inevitably affected

THE UNIVERSITY OF MEMPHIS

Problems with Evaluating User-Language

- Lack of "colloquial" paraphrase corpora
 - Microsoft Research Paraphrase Corpus (Dolan, Quirk, & Brockett, 2004)
 - Only binary rating (*is/is not* a paraphrase)
 - Echo Chamber (Brockett & Dolan, 2005)
 - Paraphrase Game (Chklovski, 2005)
- Limitations in "cleaning" ITS input
 - Datasets artificially created (Fossati & Di Eugenio, 2008)
 - Target populations are relatively proficient
 - Why2-Atlas: College undergraduates
 - CIRCSIM-Tutor: 1st year medical students
 - Use lexicons; computationally expensive

THE UNIVERSITY OF MEMPHIS

User-Language Paraphrase Corpus

- 1998 target sentence/student response pairs
- Paraphrase attempts by high school students
 - During interactions with iSTART (McNamara, Levinstein, & Boonthum, 2004)
- Paraphrases evaluated on widely used computational indices
 - Latent semantic analysis (LSA; Landauer, McNamara, Dennis, & Kintsch, 2007)
 - Entailment (Rus et al, 2007)
 - Type-Token Ratio (TTR; Graesser, McNamara, et al., 2004)
 - Mean Edit Distance (MED; McCarthy et al., 2007)
- Paraphrases also evaluated by trained experts on 10 dimensions w/ Likert ratings

THE UNIVERSITY OF MEMPHIS

Research Questions

- How are established computational indices affected by the types of errors found in typed user-language?
- Do user errors affect NLP assessment and feedback produced by an established ITS?
- Does correcting user errors improve the capacity for ITS assessment to correspond to human ratings?

THE UNIVERSITY OF MEMPHIS

iSTART

- High school students (U.S. grades 9-12)
- Reading strategy training
 - Paraphrasing, Elaboration, Making Bridging Inferences, Comprehension Monitoring
- Paraphrase the following:
 - *Over two thirds of the heat generated by a resting human is created by organs of the thoracic and abdominal cavities and the brain.*
 - *a lot of heat made by a lazy person is made by systems of your stomach and thinking box.*

iSTART Evaluation Process

- Based on match between paraphrase and target sentence
- Respond to or remove Frozen expression
 - e.g., *I think this is saying...*
- Word & Soundex matching against benchmark for length, relevance, & similarity
 - Irrelevant (IRR) – too few words match
 - Too short (SH) – response is shorter than specified threshold
 - Too similar (SIM1) – length and word match is close to benchmark
- Word match & LSA cosines for quality
 - Adequate paraphrase (SIM2)
 - Better than a paraphrase (OK)

Detailed formulae – McNamara, Boonthum, et al. (2007)

Soundex

- Compensates for misspellings (Christian, 1998)
- Vowels removed
- Like-sounding consonants mapped onto same symbol
 - e.g., *b, f, p, v*
- Lexicon-free
- Word frequency problem
 - Students make more mistakes on new or uncommon words

Procedure

- Identified, coded, & corrected all errors
 - Based on validated models of grammar (e.g., Foster & Vogel, 2004)
- Interrater agreement for subset ($n = 200$)
 - Kappa = .70, $p < .001$
 - Single rater coded entire corpus
- 83% of responses contained some form of error
- 52% had some form of spelling error
- 63% of spelling errors were *internal* to target sentence

Error types & frequencies

Spelling (internal)	665 (33%)
Spelling (external)	386 (19%)
Capitalization	1157 (58%)
S-V Agreement	367 (18%)
Article agreement	75 (4%)
Preposition agreement	53 (3%)
Determiner agreement	59 (3%)
Spacing	174 (9%)
Punctuation	344 (17%)
Conjunction agreement	43 (2%)
Possessive agreement	71 (4%)
Extra/omitted/substitute	230 (12%)

Results

- Significant effect of error correction on computational similarity indices
 - Partial Eta² =
 - LSA .178
 - Entailment .268
 - TTR .240
 - MED .111
- *Spelling internal* accounts for large portion of variance
 - Adjusted R² =
 - LSA .35
 - Entailment .45
 - TTR .46
 - MED .17

Example

Target Sentence:

An increase in temperature of a substance is an indication that it has gained heat energy.

LSA
.54 → .90
Entailer
.41 → .78
TTR
.86 → .62
MED
.78 → .60

Student response:

increase in tempiture has gaind heat energy.

Revised response:

Increase in temperature has gained heat energy.

Results

➤ Table 1: Crosstabulation of iSTART responses to user paraphrases

		iSTART response – corrected						Total
		Too Better			Too Short			
iSTART response original paraphrase		Better	Good	Similar	Short	Irrelevant	Frozen	
	Better		691	45	37	4	0	0
Good		12	194	98	0	0	0	304
Too Similar		7	7	527	0	0	0	541
Too Short		11	0	1	206	2	1	221
Irrelevant		6	0	0	6	120	7	139
Frozen		0	0	0	0	0	16	16
Total		727	245	663	216	122	24	1998

➤ Cramer's V = .849, $p < .001$
➤ Marginal Homogeneity (MH) = 5.892, $p < .001$

Results

- Compared iSTART feedback's correspondence to human ratings of Paraphrase Quality
- Removed cases that required no correction or were entirely garbage
 - $n = 328$
- Separate ANOVAs for original and corrected
 - Dependent – Paraphrase Quality
 - Fixed Factor – iSTART response
 - Original paraphrases, $F(5, 1636) = 53.324, p < .001$
 - Corrected paraphrases, $F(5, 1636) = 58.543, p < .001$

Results

➤ Separate pairwise comparisons of Paraphrase Quality

		Original			Corrected		
		Mean Diff.	SE	Sig. ^a	Mean Diff.	SE	Sig. ^a
Frozen	Irrelevant	.152	.402	1	.081	.361	1
	Too short	-.776	.370	.581	-.922	.299	.032
	Too Sim	-1.955	.363	< .001	-2.176	.288	< .001
	Good	-2.071	.366	< .001	-2.421	.297	< .001
Irrelevant	Better	-1.897	.361	< .001	-2.106	.288	< .001
	Too short	-.918	.209	< .001	-1.002	.245	.001
	Too Sim	-2.107	.196	< .001	-2.257	.231	< .001
	Good	-2.223	.203	< .001	-2.502	.242	< .001
Too short	Better	-2.0249	.192	< .001	-2.187	.231	< .001
	Too Sim	-1.189	.115	< .001	-1.255	.112	< .001
	Good	-1.305	.127	< .001	-1.500	.133	< .001
	Better	-1.131	.111	< .001	-1.185	.111	< .001
Too similar	Good	-.116	.103	1	-.245	.107	.331
	Better	.058	.082	1	.070	.077	1
Good	Better	.174	.097	1	.315	.106	.044

^a Adjustment for multiple comparisons: Bonferroni.

Discussion

- ITS feedback algorithms may be optimized if user-language can be filtered prior to processing
 - Misclassification OK for motivation
 - Accuracy not OK: simple rewording can pass for good paraphrase; paraphrase can pass for better
- Established NLP approaches not as robust to user-language as believed
 - Response length not enough to wash out individual errors
 - ULPC represents types & amount of errors *real* students make
- Most variance accounted for by internal misspellings
 - Provides direction for future research
 - Automatic spelling corrections only for words in the benchmark
 - Will be silent & computationally light

Thank you!

- We would like to thank:
 - Vasile Rus
 - Ben Duncan
 - John Myers
 - Rebekah Guess

- Research supported by:



IIS-0735682



Institute of Education Sciences
R305A080589



INSTITUTE for INTELLIGENT SYSTEMS