Embedding-based Query Spelling Correction

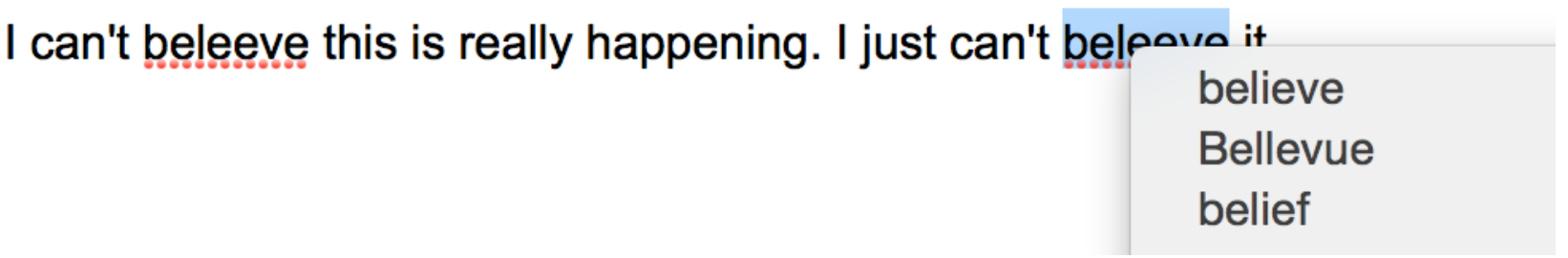
Ines Zelch[1,2] Gustav Lahmann[1] Matthias Hagen[1]

[1]Friedrich-Schiller-Universität Jena [2]Leipzig University

WOWS, March 28 2024

webis.de

Query Spelling Correction



[https://i.stack.imgur.com/DCUcD.png]

Query Spelling Correction

Type	Misspelling		Correction
Deletion	corosion	$\rightarrow$	corrosion
Insertion	occurr	$\rightarrow$	occur
Space	abouta	$\rightarrow$	about a
Special character	isnt	$\rightarrow$	isn't
Substitution	grammer	$\rightarrow$	grammar
Transposition	rewriet	$\rightarrow$	rewrite

[Hagen et al. 2017, SIGIR]

Query Spelling Correction

Type	Misspelling		Correction
Deletion	corosion	$\rightarrow$	corrosion
Insertion	occurr	$\rightarrow$	occur
Space	abouta	$\rightarrow$	about a
Special character	isnt	$\rightarrow$	isn't
Substitution	grammer	$\rightarrow$	grammar
Transposition	rewriet	$\rightarrow$	rewrite

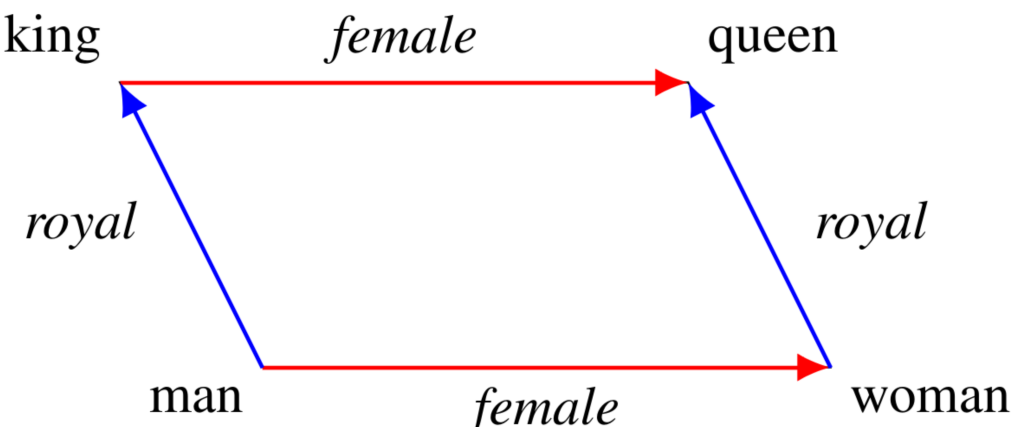
- real-word errors (e.g.: `their` instead of `there`)
- non-word errors (e.g.: `grammer`)

[Hagen et al. 2017, SIGIR]

Embeddings

- Blog post: embedding-based correction [Rushton, 2018]

  https://edrushton.medium.com/a-simple-spell-checker-built-from-word-vectors-9f28452b6f26

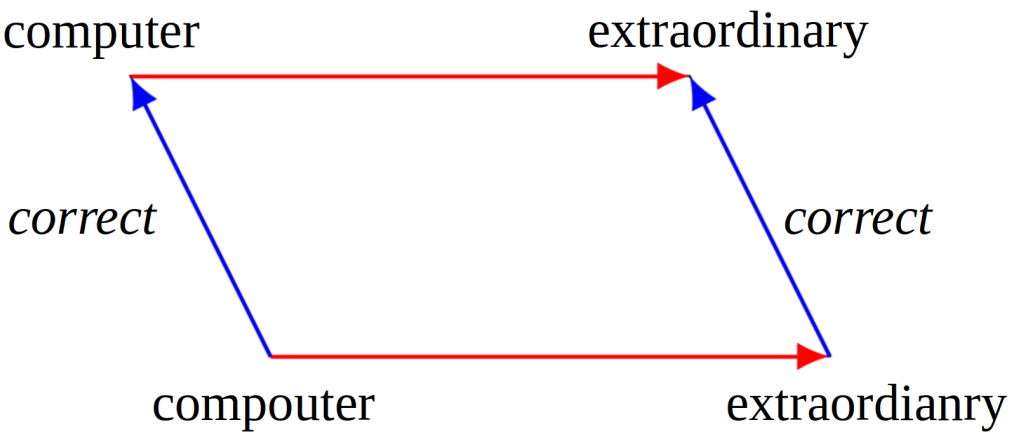


$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

[Ethayarajh et al. 2019, ACL]

Query Spelling Correction with Embeddings

- Misspelling property encoded in embeddings



$$\overrightarrow{computer} - \overrightarrow{compouter} + \overrightarrow{extraordianry} \approx \overrightarrow{extraordinary}$$

[Rushton, 2018]

Query Spelling Correction with Embeddings

- 300-dim GloVe embeddings nlp.stanford.edu/projects/glove/
- Compute correction vector $\overrightarrow{v_{correct}}$ (Oxford dict of common misspellings)
- Add $\overrightarrow{v_{correct}}$ to each GloVe embedding
- Save nearest neighbor
- Look up correction for each query word

Evaluation – Retrieval Effectiveness

None

Rushton

GPT-3.5

Hunspell

pyspellchecker

Evaluation – Retrieval Effectiveness

	ANTIQUE		TREC DL '19		TREC DL '20	
Spell Correction	BM25	PL2	BM25	PL2	BM25	PL2
None						
Rushton						
GPT-3.5						
Hunspell						
pyspellchecker						

nDCG@10

Evaluation – Retrieval Effectiveness

	ANTIQUE		TREC DL '19		TREC DL '20	
Spell Correction	BM25	PL2	BM25	PL2	BM25	PL2
None	0.51	0.49	0.48	0.47	0.49	0.48
Rushton	0.51	0.49	0.48	0.47	0.49	0.48
GPT-3.5						
Hunspell						
pyspellchecker						

nDCG@10

Evaluation – Retrieval Effectiveness

	ANTIQUE		TREC DL '19		TREC DL '20	
Spell Correction	BM25	PL2	BM25	PL2	BM25	PL2
None	0.51	0.49	0.48	0.47	0.49	0.48
Rushton	0.51	0.49	0.48	0.47	0.49	0.48
GPT-3.5	0.49	0.47	0.48	0.47	0.48	0.47
Hunspell	0.48	0.46	0.36	0.34	0.42	0.41
pyspellchecker	0.43	0.41	0.31	0.30	0.33	0.32

nDCG@10

Evaluation – Corrections

	ANTIQUE				
	#err.	corrected			
Error Type		R	H	P	G
Deletion	14	4	6	5	12
Insertion	5	1	1	1	4
Space	13	1	1	2	11
Special char.	19	-	2	6	9
Substitution	8	1	2	2	8
Transposition	4	2	2	1	4
Incorr. changes		0	52	75	42

R – Rushton
H – Hunspell
P – pyspellchecker
G – GPT3.5

Evaluation – Corrections Examples

	ANTIQUE				
	#err.	corrected			
Error Type		R	H	P	G
Deletion	14	4	6	5	12
Insertion	5	1	1	1	4
Space	13	1	1	2	11
Special char.	19	-	2	6	9
Substitution	8	1	2	2	8
Transposition	4	2	2	1	4
Incorr. changes		0	52	75	42

Original query:

`why is gordon ramsey so popular`

Rushton: —

Hunspell: —

pyspellchecker:

`why is cordon raise so popular`

GPT-3.5:

`why is gordon ramsay so popular?`

Evaluation – Corrections Examples

	ANTIQUE				
	#err.	corrected			
Error Type		R	H	P	G
Deletion	14	4	6	5	12
Insertion	5	1	1	1	4
Space	13	1	1	2	11
Special char.	19	-	2	6	9
Substitution	8	1	2	2	8
Transposition	4	2	2	1	4
Incorr. changes		0	52	75	42

Original query:

`why is gordon ramsey so popular`

Rushton: —

Hunspell: —

pyspellchecker:

`why is cordon raise so popular`

GPT-3.5:

`why is gordon ramsay so popular?`

Limitation Rushton: GloVe vocabulary

Evaluation – Corrections Examples

	ANTIQUE				
	#err.	corrected			
Error Type		R	H	P	G
Deletion	14	4	6	5	12
Insertion	5	1	1	1	4
Space	13	1	1	2	11
Special char.	19	-	2	6	9
Substitution	8	1	2	2	8
Transposition	4	2	2	1	4
Incorr. changes		0	52	75	42

Original query:

`[...] copies of letters of commendation prsented [...]`

Rushton, Hunspell, pyspell., GPT-3.5:

`[...] copies of letters of commendation presented [...]`

Evaluation – Corrections Examples

	ANTIQUE				
	#err.	corrected			
Error Type		R	H	P	G
Deletion	14	4	6	5	12
Insertion	5	1	1	1	4
Space	13	1	1	2	11
Special char.	19	-	2	6	9
Substitution	8	1	2	2	8
Transposition	4	2	2	1	4
Incorr. changes		0	52	75	42

Original query:

`how can i keep my rabit indoors`

Rushton:

`how can i keep my rabbit indoors`

Hunspell:

`how can i keep my habit indoors`

pyspellchecker:

`how can i keep my barit indoors`

GPT-3.5:

`how can i keep my rabbit indoors`

Evaluation – Corrections Examples

	ANTIQUE				
	#err.	corrected			
Error Type		R	H	P	G
Deletion	14	4	6	5	12
Insertion	5	1	1	1	4
Space	13	1	1	2	11
Special char.	19	-	2	6	9
Substitution	8	1	2	2	8
Transposition	4	2	2	1	4
Incorr. changes		0	52	75	42

Original query:

`what is wifi vs bluetooth`

Rushton, GPT-3.5: —

Hunspell, pyspellchecker:

`what is wife vs bluetooth`

Evaluation – Corrections Examples

	ANTIQUE				
	#err.	corrected			
Error Type		R	H	P	G
Deletion	14	4	6	5	12
Insertion	5	1	1	1	4
Space	13	1	1	2	11
Special char.	19	-	2	6	9
Substitution	8	1	2	2	8
Transposition	4	2	2	1	4
Incorr. changes		0	52	75	42

Original query:

```
why hot air raise up
```

GPT-3.5:

```
why does hot air raise up?
```

Original query:

```
is cdg airport in main paris
```

GPT-3.5:

```
[...] corrected search query:
is cdg airport in main paris?
```

Evaluation – Corrections Examples

	ANTIQUE				
	#err.	corrected			
Error Type		R	H	P	G
Deletion	14	4	6	5	12
Insertion	5	1	1	1	4
Space	13	1	1	2	11
Special char.	19	-	2	6	9
Substitution	8	1	2	2	8
Transposition	4	2	2	1	4
Incorr. changes		0	52	75	42

Original query:

`how can we get concentration onsomething`

Rushton: —

Hunspell:

`how can we get concentration something`

pyspellchecker:

`how can we get concentration`

GPT-3.5:

`how can we get concentration on something?`

Limitation Rushton: Spaces, real-word errors

Evaluation – Corrections Examples

	ANTIQUE					TREC DL '19					TREC DL '20				
	#err.	corrected				#err.	corrected				#err.	corrected			
Error Type		R	H	P	G		R	H	P	G		R	H	P	G
Deletion	14	4	6	5	12	-	-	-	-	-	-	-	-	-	-
Insertion	5	1	1	1	4	1	-	1	1	1	-	-	-	-	-
Space	13	1	1	2	11	-	-	-	-	-	-	-	-	-	-
Special char.	19	-	2	6	9	1	-	-	-	-	3	-	-	-	-
Substitution	8	1	2	2	8	1	-	1	1	1	1	-	-	-	-
Transposition	4	2	2	1	4	-	-	-	-	-	-	-	-	-	-
Incorr. changes		0	52	75	42		1	12	15	5		0	11	23	15

Wrap-Up

- Considered spell correctors worsen retrieval results (on average)
- Embedding-based approach better than common spell correctors

Wrap-Up

- Considered spell correctors worsen retrieval results (on average)
- Embedding-based approach better than common spell correctors

Limitations

- GloVe vocabulary
- Single words
- No context

Wrap-Up

- Considered spell correctors worsen retrieval results (on average)
- Embedding-based approach better than common spell correctors

Limitations

- GloVe vocabulary
- Single words
- No context

Future Work

- Other correction approaches
- Other embeddings
- Create misspelling query variants