

Using Learner Data to Improve Error Correction in Adjective-Noun Combinations

Ekaterina Kochmar and Ted Briscoe

The ALTA Institute, Computer Laboratory
University of Cambridge

Error detection and correction (ED&C): State-of-the-art

Attracted much attention recently:

- books [LEACOCK *et al.*, 2014; LEACOCK *et al.*, 2010]
- shared tasks [NG *et al.*, 2014; NG *et al.*, 2013; DALE *et al.*, 2012, DALE AND KILGARRIFF, 2011]
- multiple papers and dissertations
- multiple workshops (**10th** anniversary of **BEA!**)

However, so far:

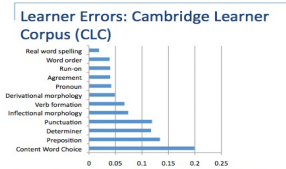
- major focus on grammatical errors, errors in articles and prepositions
- fewer address other error types [KOCHMAR AND BRISCOE, 2014; NG *et al.*, 2014; ROZOVSKAYA *et al.*, 2014; SAWAI *et al.*, 2013; DAHLMEIER AND NG, 2011]



Our work: Focus

Errors in content words (ANs in particular)

- Frequent error types [LEACOCK *et al.*, 2014; NG *et al.*, 2014]



← cover 20% of learner errors in the CLC [TETREAUULT AND LEACOCK, 2014]

- notoriously hard to master
- yet, important for successful writing [LEACOCK AND CHODOROW, 2003; JOHNSON, 2000; SANTOS, 1988]

Content word errors: Challenges

- Lack of strictly defined rules:
 - *powerful computer* ↔ *strong computer*
 - *powerful tea* ↔ *strong tea*
- Sources of confusion:
 - similarity in meaning:
 - ▷ *powerful* ~ *strong*
 - similarity in spelling:
 - ▷ *classic* ~ *classical*
 - overusing words with general meaning:
 - ▷ *big* vs *broad*|*wide*|*long*|...
 - L1-related confusions
 - ▷ *good humor* vs *good mood* (cf. French *bon humor*)



ED algorithms: General overview

Function vs Content Words

Function words

- ▷ Multi-class classification using number of possible alternatives
- ▷ Availability of finite confusion sets
- ▷ Error detection and correction – possible to do simultaneously

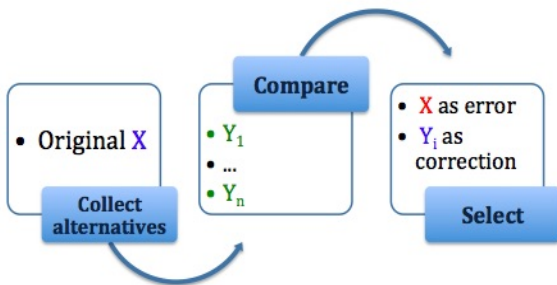
Content words

- ▷ What are the multiple classes?
- ▷ Confusion sets depend on the original word choice
- ▷ Error detection independent of error correction [KOCHMAR AND BRISCOE, 2014]

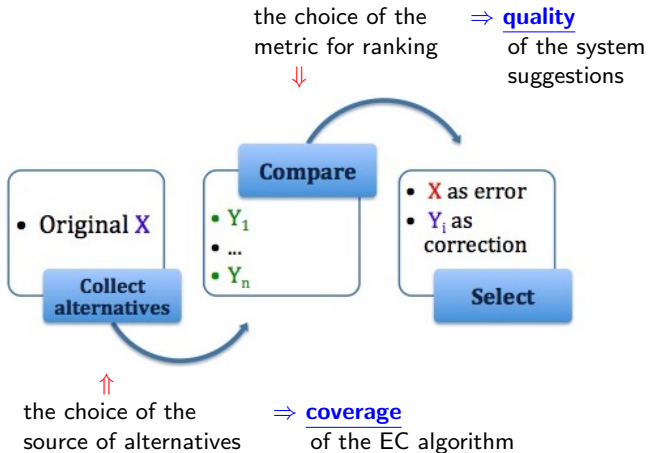
Basic EDC algorithm

Three-step algorithm [LEACOCK *et al.*, 2014]:

- 1 $\forall X$ look for more *fluent/native-like* Y 's
- 2 compare Y 's to X using some frequency-based measure
- 3 if $\exists Y_i$ more fluent than $X \Rightarrow X$ is an error, Y_i is a correction



Basic EDC algorithm performance



Different sources

- Reference databases of known learner errors and their corrections [WIBLE *et al.*, 2003; SHEI AND PAIN, 2000]
- Semantically related: WordNet, dictionaries and thesauri [ÖSTLING AND KNUTSSON, 2009; FUTAGI *et al.*, 2008; SHEI AND PAIN, 2000]
- Spelling alternatives and homophones [DAHLMEIER AND NG, 2011]
- L1-specific confusion sets [DAHLMEIER AND NG, 2011; CHANG *et al.*, 2008; LIU, 2002]
- Wikipedia revisions [MADNANI AND CAHILL, 2014]

Our work: Contributions

In this work

We treat error detection and error correction as separate steps, and focus on [error correction](#)

Contributions

- 1 Explore different ways to construct the correction sets and to rank the alternatives
- 2 Demonstrate how error patterns extracted from learner text can be used to improve the ranking of the alternatives
- 3 Present an EDC system for AN combinations
- 4 Explore the usefulness of augmenting sets of alternatives for an EC system

Datasets

- 1 the AN dataset extracted from the *Cambridge Learner Corpus* (CLC) and annotated with respect to the learner errors
<http://ilexir.co.uk/media/an-dataset.xml>
- 2 the AN dataset extracted from the *CLC-FCE* dataset
<http://ilexir.co.uk/applications/adjective-noun-dataset/>
- 3 the AN dataset extracted from the *CoNLL-2014* Shared Task on Grammatical Error Correction training and development sets
<http://www.comp.nus.edu.sg/~nlp/conll14st.html>

Datasets

Annotated dataset

- **340** unique errors
- annotated with the error types for adjectives and nouns (S, F and N)

CLC-FCE dataset

- **456** ANs that have adjective–noun combinations as corrections
- no annotation for error types

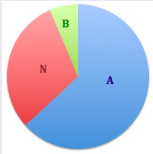


NUCLE dataset

- **369** ANs that have adjective–noun combinations as corrections
- no annotation for error types
- smaller number of L1s, different set of topics, etc.

All datasets

Distribution of errors in the choice of adjectives (A), nouns (N) or both words

Word	Ann. data	CLC-FCE	NUCLE
A	63.24%	43.20%	34.15%
N	30.29%	52.63%	60.16%
Both	6.47%	4.17%	5.69%

		
-----------------------------------------------------------------------------------	-----------------------------------------------------------------------------------	------------------------------------------------------------------------------------

Error Correction Algorithm

Key points

- 1 Explore resources to retrieve alternatives and report **coverage**
 - **coverage** – proportion of gold standard corrections covered by the resources
- 2 Rank AN alternatives and assess the **quality** of ranking (*MRR*)
 - **quality** – ability of the algorithm to rank the more appropriate corrections higher than the less appropriate ones
- 3 Use confusion sets extracted from the learner data

Resources

- Levenshtein distance (**Lv**): form-related confusions, F
E.g.: ***electric** society → **electronic** society
important ***costumer** → important **customer**
- WordNet (**WN**): semantically related confusions, S
E.g.: ***heavy** decline → **steep** decline
good ***fate** → good **luck**
- Confusion pairs from the **CLC**: cover L1-related confusions, N
E.g.: ***strong** noise → **loud** noise
historical ***roman** → historical **novel**

Coverage

Coverage of different sets of alternatives

Setting	Ann. data	CLC-FCE	NUCLE
Lv	0.1588	0.0833	0.0897
WN	0.4353	0.3904	0.2880
CLC	0.7912	0.8684	0.5625
CLC+Lv	0.7971	0.8706	0.5951
CLC+WN	0.8558	0.8904	0.6141
All	0.8618	0.8925	0.6467

Alternative ANs

ANs generation

$$\{\textit{alternative ANs}\} = \cup \left(\begin{array}{l} \{\textit{alternative adjs}\} \times \textit{noun} \\ \textit{adj} \times \{\textit{alternative nouns}\} \end{array} \right) \quad (1)$$

Evaluation

$$MRR = \frac{1}{|N|} \sum_{i=1}^{|N|} \frac{1}{rank_i} \quad (2)$$

N – total number of erroneous ANs

Ranking

Ranking measures

- 1 *Frequency* in the BNC+ukWaC
- 2 *Normalised pointwise mutual information (NPMI)*:

$$NPMI(AN) = \frac{PMI(AN)}{-\log_2(P(AN))} \quad (3)$$

where

$$PMI(AN) = \frac{\log P(AN)}{P(A)P(N)} \quad (4)$$

Quality (I)

MRR for the alternatives ranking (I)

Setting	Ann. set	CLC-FCE	NUCLE
CLC _{freq}	0.3806	0.3121	0.2275
CLC _{NPMI}	0.3752	0.2904	0.1961
(CLC+Lv) _{freq}	0.3686	0.3146	0.2510
(CLC+Lv) _{NPMI}	0.3409	0.2695	0.1977
(CLC+WN) _{freq}	0.3500	0.2873	0.2267
(CLC+WN) _{NPMI}	0.3286	0.2552	0.1908
All _{freq}	0.3441	0.2881	0.2468
All _{NPMI}	0.3032	0.2407	0.1943

Exploitation of confusion probabilities

Use the **confusion probabilities (CP)** from the CLC – probabilities associated with the words used as corrections given the original (incorrect) word choice

Formula refinement

$$M' = M \times CP(a_{orig} \rightarrow a_{alt}) \times CP(n_{orig} \rightarrow n_{alt}) \quad (5)$$

- M – a measure of choice
- $CP(a_{orig} \rightarrow a_{alt=orig})$ and $CP(n_{orig} \rightarrow n_{alt=orig})$ set to 1.0

Example: **big enjoyment* → *great pleasure*

CLC confusion pairs

Original	Alternatives	CP(orig → alt)
<i>big</i>	<i>great</i>	0.0144
	<i>large</i>	0.0141
	<i>wide</i>	0.0043

	<i>significant</i>	$5.1122 * 10^{-5}$
<i>enjoyment</i>	<i>pleasure</i>	0.0938
	<i>entertainment</i>	0.0313
	<i>fun</i>	0.0104
	<i>happiness</i>	0.0052

Example: *big enjoyment → great pleasure

Basic ranking algorithm (raw frequency)

System: great fun (7759 in the native corpus)

GS: great pleasure (2829 in the native corpus)

Refined ranking algorithm (frequency')

System & GC: great pleasure ($Freq' = 3.8212$)

great fun ($Freq' = 1.1620$)

Freq' vs freq

fluency in the native data + *appropriateness* of a correction

Quality (II)

MRR for the alternatives ranking (II)

Setting	Ann. set	CLC-FCE	NUCLE
CLC _{freq}	0.3806	0.3121	0.2275
CLC _{NPMI}	0.3752	0.2904	0.1961
(CLC+Lv) _{freq}	0.3686	0.3146	0.2510
(CLC+Lv) _{NPMI}	0.3409	0.2695	0.1977
(CLC+WN) _{freq}	0.3500	0.2873	0.2267
(CLC+WN) _{NPMI}	0.3286	0.2552	0.1908
All _{freq}	0.3441	0.2881	0.2468
All _{NPMI}	0.3032	0.2407	0.1943
All _{freq} '	0.5061	0.4509	0.2913
All _{NPMI} '	0.4843	0.4316	0.2118

Further analysis of the results

- 1 Breakdown of the results
 - Top N coverage
 - Error types
- 2 System augmentation
- 3 Error detection + correction

% of errors covered by top N alternatives

Top N	Ann. data	CLC-FCE	NUCLE
1	41.18	34.21	21.20
2	49.12	45.18	27.99
3	56.77	50.88	33.70
4	61.77	55.04	38.04
5	65.29	58.55	40.49
6	66.18	61.40	42.39
7	67.35	62.28	43.21
8	68.53	63.60	44.29
9	69.71	65.35	45.38
10	71.18	66.45	46.20
Not found	25.29	19.96	48.64

Error type analysis for the annotated dataset

Type	S	F	N
MRR_{found}	0.6007	0.8486	0.6507
Not found	0.1990	0.1705	0.5410

Some observations

- type N (non-related confusion) – the hardest to correct (not surprisingly...)
- type F (form-related) – the easiest (smaller confusion sets)
e.g., $MRR = 0.875$ for the ANs with *elder* :
elder → *elderly* or *older*

NUCLE results

- 35% of the GS corrections not covered by any sets of alternatives
- confusion sets from the CLC can only cover about 56%
- more limited number of L1s
- different set of topics and learner levels
- more of the type N?

**architectural* development → *infrastructural* development
medical **debt* → medical *bill*

(6)

Augmenting sets of alternatives

Method

- Add *bill* to the set of alternatives for *debt*
- Add *infrastructural* to the set of alternatives for *architectural*
-
- Check whether the results of the error correction system improve

Augmented sets of alternatives results

Setting	Ann. set	CLC-FCE	NUCLE
CLC	<u>0.3806</u>	0.3121	0.2275
CLC+Lv	0.3686	<u>0.3146</u>	<u>0.2510</u>
Augm	0.4420	0.3533	0.2614

Combined algorithm results

Algorithm

- **Error detection** [KOCHMAR AND BRISCOE, 2014]:
 $P = 0.6850$, $R = 0.5849$ on the incorrect examples in the annotated dataset
- + **Error correction** step:
 - $MRR = 0.2532$ on the set of detected errors
 - 24.28% cases GS correction not found
 - $MRR_{found} = 0.6831$

Conclusions

In this work we:

- focused on EC in adjective–noun combinations
- experimented with 3 publicly available datasets
- looked at the coverage of resources and the quality of suggestions

and we showed:

- 1 the confusion patterns from the learner data provide the **highest coverage** and improve the **overall ranking**
- 2 error correction system can reach an *MRR* of **0.5061**
- 3 correction set augmentation is **helpful**
- 4 *MRR* of **0.2532** on the set of errors identified by ED algorithm

Thank you!

Contact: Ekaterina.Kochmar@cl.cam.ac.uk

Data:

- annotated AN dataset
<http://ilexir.co.uk/media/an-dataset.xml>
- the AN dataset extracted from the CLC-FCE
<http://ilexir.co.uk/applications/adjective-noun-dataset/>

References

Y.-C. Chang, J. S. Chang, H.-J. Chen and H.-C. Liou, 2008. *An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology.* *Computer Assisted Language Learning*, 21(3)

D. Dahlmeier and H. T. Ng, 2011. *Correcting Semantic Collocation Errors with L1-induced Paraphrases.* In *Proceedings of the EMNLP 2011*

R. Dale and A. Kilgarriff, 2011. *Helping Our Own: The HOO 2011 Pilot Shared Task.* In *Proceedings of the ENLG 2011*

R. Dale, I. Anisimoff and G. Narroway, 2012. *HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task.* In *Proceedings of the BEA 2012*

Y. Futagi, P. Deane, M. Chodorow, and J. Tetreault, 2008. *A computational approach to detecting collocation errors in the writing of non-native speakers of English.* *Computer Assisted Language Learning*, 21(4)

D. Johnson, 2000. *Just the Right Word: Vocabulary and Writing.* In R. Indrisano & J. Squire (Eds.), *Perspectives on Writing: Research, Theory, and Practice*



E. Kochmar and T. Briscoe, 2014. *Detecting Learner Errors in the Choice of Content Words Using Compositional Distributional Semantics*. In Proceedings of the COLING 2014

C. Leacock, M. Chodorow, and J. Tetreault, 2014. *Automated Grammatical Error Detection for Language Learners, Second Edition Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers

C. Leacock, M. Chodorow, and J. Tetreault, 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool Publishers

C. Leacock and M. Chodorow, 2003. *Automated Grammatical Error Detection*. In M. D. Shermis and J. C. Burstein (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*,

A Liu, 2002. *A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learners English*. Master's thesis

N. Madnani and A. Cahill, 2014. *An Explicit Feedback System for Preposition Errors based on Wikipedia Revisions*. In Proceedings of the BEA 2014

H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault, 2013. *The CoNLL-2013 Shared Task on Grammatical Error Correction*. In Proceedings of the CoNLL 2013

H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant, 2014. *The CoNLL-2014 Shared Task on Grammatical Error Correction*. In Proceedings of the CoNLL 2014

D. Nicholls, 2003. *The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT*. In Proceedings of the Corpus Linguistics conference

R. Östling and O. Knutsson, 2009. *A corpus-based tool for helping writers with Swedish collocations*. In Proceedings of the Workshop on Extracting and Using Constructions in NLP

A. Rozovskaya, K.-W. Chang, M. Sammons, D. Roth, and N. Habash, 2014. *Correcting Grammatical Verb Errors*. In Proceedings of the EACL 2014

T. Santos, 1988. *Professors' reaction to the academic writing of nonnative speaking students*. TESOL Quarterly, 22(1)

Y. Sawai, M. Komachi, and Y. Matsumoto, 2013. *A Learner Corpus-based Approach to Verb Suggestion for ESL*. In Proceedings of the ACL 2013

C.-C. Shei and H. Pain, 2000. *An ESL Writer's Collocation Aid*. Computer Assisted Language Learning, 13(2)

J. Tetreault and C. Leacock, 2014. *Automated Grammatical Error Correction for Language Learners*. Tutorial, COLING 2014

D. Wible, C.-H. Kuo, N.-L. Tsao, A. Liu and H.-L. Lin, 2003. *Bootstrapping in a language-learning environment*. Journal of Computer Assisted Learning, 19(4)

H. Yannakoudakis, T. Briscoe, and B. Medlock, 2011. *A New Dataset and Method for Automatically Grading ESOL Texts*. In Proceedings of the ACL: HLT 2011