



ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ

Faculty of Mathematics and Computer Science
Information Systems Laboratory

The WikEd Error Corpus:

A Corpus of Corrective Wikipedia Edits and its Application to Grammatical Error Correction

PoITAL 2014

September 17, 2014

Roman Grundkiewicz
Marcin Junczys-Dowmunt



Introduction

Machine learning approaches are data-hungry.

Corpora for grammatical error correction:

- ▶ learner's corpora
- ▶ artificial errors
- ▶ text edition histories
- ▶ social networks



Introduction

Machine learning approaches are data-hungry.

Corpora for grammatical error correction:

- ▶ learner's corpora
- ▶ artificial errors
- ▶ text edition histories
- ▶ social networks



Outline

The WikEd Error Corpus:

- ▶ mining English Wikipedia
- ▶ corpus content and format

Grammatical error correction:

- ▶ English-as-a-Second-Language (ESL) learners' writings
- ▶ Statistical Machine Translation (SMT) system
- ▶ error selection procedure



Outline

The WikEd Error Corpus:

- ▶ mining English Wikipedia
- ▶ corpus content and format

Grammatical error correction:

- ▶ English-as-a-Second-Language (ESL) learners' writings
- ▶ Statistical Machine Translation (SMT) system
- ▶ error selection procedure



Part 1: The WikEd Error Corpus



Wikipedia Edits

Revision as of 00:52, 22 September 2002

([view source](#))

(*m:History of Wikipedia*)

← Previous edit

Revision as of 03:15, 22 September 2002

([view source](#))

m (*fixed misspell "srcratch"*)

Next edit →

Line 147:

which used an underlying [[MySQL]] database, added many features and was specifically written for the Wikipedia project by

[[user:Magnus Manske|Magnus Manske]]. (...) Then [[user:Lee Daniel Crocker|Lee Daniel Crocker]] rewrote the software from **srcratch**; the new version, a major improvement, has been running since July 2002.

The project has occasionally been visited by

Line 147:

which used an underlying [[MySQL]] database, added many features and was specifically written for the Wikipedia project by

[[user:Magnus Manske|Magnus Manske]]. (...) Then [[user:Lee Daniel Crocker|Lee Daniel Crocker]] rewrote the software from **scratch**; the new version, a major improvement, has been running since July 2002.

The project has occasionally been visited by



Extracting Edits

1 Iterate over each two adjacent revisions:

- ▶ skip vandalism reverts
- ▶ remove Wiki markup
- ▶ iterate over edited sentences

2 Extract sentences with edits:

- ▶ sentence length > 2 and < 120 tokens
- ▶ relative token-based edit distance < 0.3 :

$$\text{ed}(s_i, s_j) = \frac{\text{dist}(s_i, s_j) \min(|s_i|, |s_j|)}{\log_b \min(|s_i|, |s_j|)}$$



Extracting Edits

1 Iterate over each two adjacent revisions:

- ▶ skip vandalism reverts
- ▶ remove Wiki markup
- ▶ iterate over edited sentences

2 Extract sentences with edits:

- ▶ sentence length > 2 and < 120 tokens
- ▶ relative token-based edit distance < 0.3 :

$$\text{ed}(s_i, s_j) = \frac{\text{dist}(s_i, s_j) \min(|s_i|, |s_j|)}{\log_b \min(|s_i|, |s_j|)}$$



Collected Edits #1

▶ spelling error corrections:

You can use rsync to [-download-] {+download+} the database .

▶ grammatical error corrections:

There [-is-] {+are+} also [-a-] two computer games based on the movie .

▶ sentence rewordings and paraphrases:

These anarchists [-argue against-] {+oppose the+} regulation of corporations .



Collected Edits #2

- ▶ encyclopaedic style adjustments:

A [-local education authority-] {+Local Education Authority+} (LEA) is the part of a council in England or Wales

- ▶ information supplements:

Aphrodite is the Greek goddess of love {+, sex+} and beauty .

- ▶ changes made by vandals:

David Zuckerman is a writer and [-producer-] {+poopface+} for television shows



The WikEd Corpus

The Wiked Error Corpus version 0.9:

- ▶ edits from English Wikipedia
- ▶ <http://romang.home.amu.edu.pl/wiked/wiked.html>
- ▶ ca. 12 million of sentences



Lang-8 Corpus

Lang-8 Learner Corpus v1.0:

- ▶ true ESL corpus
- ▶ scraped from <http://lang-8.com>
- ▶ ca. 2.5 million of sentences



Corpora Statistics

Statistics	WikEd 0.9	L8-NAIST
# sentences	12.13	2.57
# tokens (source side)	292.57	28.51
# edits	16.01	3.41
# edits per sentence	1.32/sent.	1.33/sent.
% sentences with ≥ 1 edit	91.79%	53.86%

* millions



Part 2: Grammatical Error Correction



Task Description

Similar to the CoNLL 2014 Shared Task:

- ▶ create a grammatical error correction system
- ▶ training set: NUCLE Corpus ver. 3.0 (57 thousands of sentences)
- ▶ 28 different error categories
- ▶ evaluation with $F_{0.5}$ on test set (ST-2013)



System Description

SMT approach: “bad” English → “good” English

- ▶ Moses SMT system
- ▶ the NUCLE corpus as training set
- ▶ tuning on evaluation metric with 4-fold cross validation
- ▶ language model from Common Crawl



Baseline

System	4×2-CV	ST-2013
NUCLE	22.10	27.62
+WikEd	18.21	23.63
+L8-NAIST	24.44	34.06



Corpus Adaptation

The WikEd Error Corpus is not an ESL error corpus

- ▶ select errors that resemble mistakes from NUCLE



Corpus Adaptation

The WikEd Error Corpus is not an ESL error corpus

- ▶ select errors that resemble mistakes from NUCLE



Error Selection

Extracting error patterns from NUCLE

- ▶ compute a sequence of deletions/insertions/substitutions for each sentence pair
- ▶ concatenate adjacent words to a phrase
- ▶ generalize substitutions if consist of common substrings, e.g. `sub(«(\w{3,})d»,«\1»)`



Extracted Patterns

Pattern	Freq.	Example
sub(«(\w{3,})»,«\1s»)	2864	burning of fuels [-emit-] {+emits+} various gases
ins(«the»)	2494	{+The+} 21st century will be ...
del(«the»)	1772	Lastly , [-the-] engineers will put
sub(«(\w{3,})s»,«\1»)	1317	the amount of [-supports-] {+support+}
ins(«,»)	971	
ins(«a»)	679	puzzled with {+a+} lack of
sub(«(\w{3,})»,«\1d»)	300	technology that has been [-shape-] {+shaped+} by
del(«,»)	266	
sub(«(\w{3,})»,«\1ed»)	252	technology [-develop-] {+developed+} through research



Adapted Corpora

Statistics	WikEd	+Select	L8-NAIST	+Select
# sentences	12.13		2.57	
# tokens (source side)	292.57	294.97	28.51	34.35
# edits	16.01	5.32	3.41	1.07
# edits per sentence	1.32/sent.	0.44/sent.	1.33/sent.	0.42/sent.
% sentences with ≥ 1 edit	91.79%	32.62%	53.86%	28.15%

* millions



Results

System	4×2-CV	ST-2013
NUCLE	22.10	27.62
+WikEd	18.21	23.63
+Select	24.33	30.06
+L8-NAIST	24.44	34.06
+Select	26.40	34.15



Results

System	4×2-CV	ST-2013
NUCLE	22.10	27.62
+WikEd	18.21	23.63
+Select	24.33	30.06
+L8-NAIST	24.44	34.06
+Select	26.40	34.15
Joint-Translation	26.01	32.33
Composition	26.63	35.79



Summary

The WikEd Error Corpus:

- ▶ ca. 56 millions of sentences in version 1.0
- ▶ can be adapted to various tasks
- ▶ publicly available

<http://romang.home.amu.edu.pl/wiked/wiked.html>



Thank you for your attention

The WikEd Error Corpus:

**A Corpus of Corrective Wikipedia Edits
and its Application
to Grammatical Error Correction**

PoITAL 2014
September 17, 2014

Roman Grundkiewicz
Marcin Junczys-Dowmunt