A Comparative Evaluation of Deep and Shallow Approaches to the Automatic Detection of Common Grammatical Errors

Joachim Wagner, Jennifer Foster, and Josef van Genabith

EMNLP-CoNLL 28th June 2007

DCU IBM National Centre for Language Technology School of Computing, Dublin City University embarkinitiative

Talk Outline

- Motivation
- Background
- Artificial Error Corpus
- Evaluation Procedure
- Error Detection Methods
- Results and Analysis
- Conclusion and Future Work

Why Judge the Grammaticality?

- Grammar checking
- Computer-assisted language learning
  - Feedback
  - Writing aid
  - Automatic essay grading
- Re-rank computer-generated output
  - Machine translation

Why this Evaluation?

- No agreed standard
- Differences in
  - What is evaluated
  - Corpora
  - Error density
  - Error types

Talk Outline

- Motivation
- Background
- Artificial Error Corpus
- Evaluation Procedure
- Error Detection Methods
- Results and Analysis
- Conclusion and Future Work

Deep Approaches

- Precision grammar
- Aim to distinguish grammatical sentences from ungrammatical sentences
- Grammar engineers
  - Avoid overgeneration
  - Increase coverage
- For English:
  - ParGram / XLE (LFG)
  - English Resource Grammar / LKB (HPSG)

Shallow Approaches

- Real-word spelling errors
  - vs grammar errors in general
- Part-of-speech (POS) n-grams
  - Raw frequency
  - Machine learning-based classifier
  - Features of local context
  - Noisy channel model
  - N-gram similarity, POS tag set

7

Talk Outline

- Motivation
- Background
- Artificial Error Corpus
- Evaluation Procedure
- Error Detection Methods
- Results and Analysis
- Conclusion and Future Work

8

Common Grammatical Errors

- 20,000 word corpus
- Ungrammatical English sentences
  - Newspapers, academic papers, emails, …
- Correction operators
  - Substitute (48 %)
  - Insert (24 %)
  - Delete (17 %)
  - Combination (11 %)

9

Common Grammatical Errors

- 20,000 word corpus
- Ungrammatical English sentences
  - Newspapers, academic papers, emails, …
- Correction operators
  - Substitute (48 %) { Agreement errors / Real-word spelling errors
  - Insert (24 %)
  - Delete (17 %)
  - Combination (11 %)

10

Chosen Error Types

Agreement: She steered Melissa around a **corners.**

Real-word: She could **no** comprehend.

Extra word: Was that in the summer **in**?

Missing word: What the subject?

11

Automatic Error Creation

Agreement: replace determiner, noun or verb

Real-word: replace according to pre-compiled list

Extra word: duplicate token or part-of-speech, or insert a random token

Missing word: delete token (likelihood based on part-of-speech)

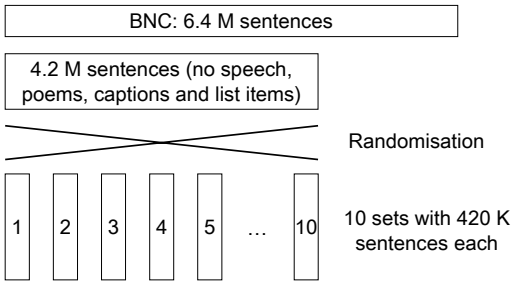
12

Talk Outline

- Motivation
- Background
- Artificial Error Corpus
- Evaluation Procedure
- Error Detection Methods
- Results and Analysis
- Conclusion and Future Work

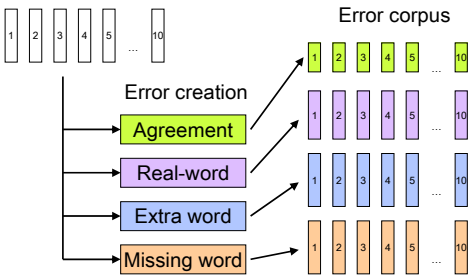
13

BNC Test Data (1)



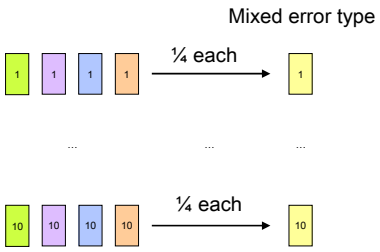
14

BNC Test Data (2)



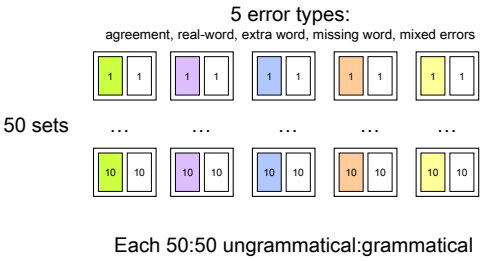
15

BNC Test Data (3)



16

BNC Test Data (4)

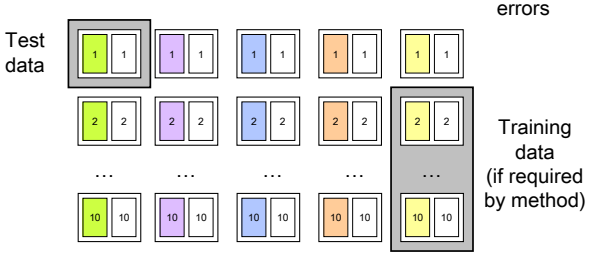


Each 50:50 ungrammatical:grammatical

17

BNC Test Data (5)

Example: 1st cross-validation run for agreement errors



18

Evaluation Measures

tp = true positive
tn = true negative
fp = false positive
fn = false negative

- Precision tp / (tp + fp)
- Recall tp / (tp + fn)
- F-score 2*pr*re / (pr + re)
- Accuracy (tp + tn) / total
- tp := ungrammatical sentences identified as such

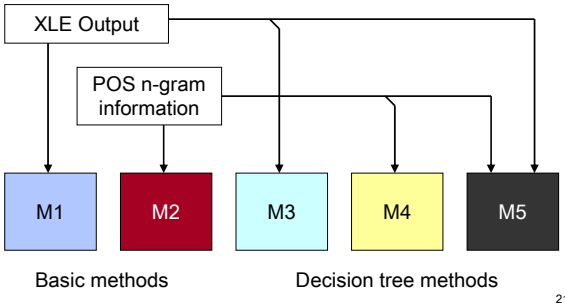
19

Talk Outline

- Motivation
- Background
- Artificial Error Corpus
- Evaluation Procedure
- Error Detection Methods
- Results and Analysis
- Conclusion and Future Work

20

Overview of Methods



21

M1

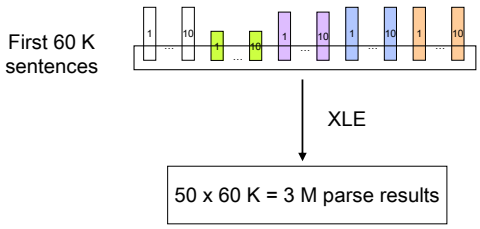
Method 1: Precision Grammar

- XLE English LFG
- Fragment rule
  - Parses ungrammatical input
  - Marked with *
- Zero number of parses
- Parser exceptions (time-out, memory)

22

M1

XLE Parsing



23

M2

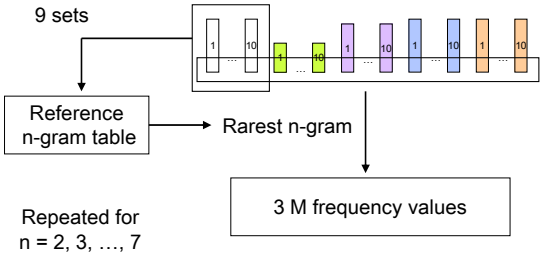
Method 2: POS N-grams

- Flag rare POS n-grams as errors
- Rare: according to reference corpus
- Parameters: n and frequency threshold
  - Tested n = 2, …, 7 on held-out data
  - Best: n = 5 and frequency threshold 4

24

M2

POS N-gram Information



25

M3

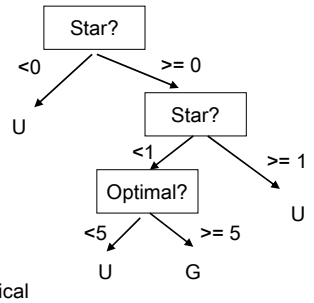
Method 3: Decision Trees on XLE Output

- Output statistics
  - Starredness (0 or 1) and parser exceptions (-1 = time-out, -2 = exceeded memory, …)
  - Number of optimal parses
  - Number of unoptimal parses
  - Duration of parsing
  - Number of subtrees
  - Number of words

26

M3

Decision Tree Example



U = ungrammatical
G = grammatical

27

M4

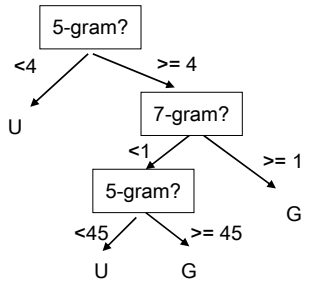
Method 4: Decision Trees on N-grams

- Frequency of rarest n-gram in sentence
- N = 2, …, 7
  - feature vector: 6 numbers

28

M4

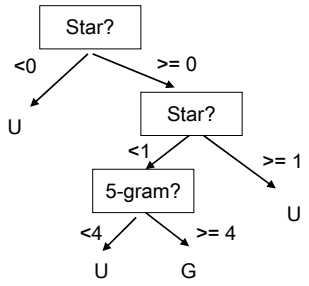
Decision Tree Example



29

M5

Method 5: Decision Trees on Combined Feature Sets



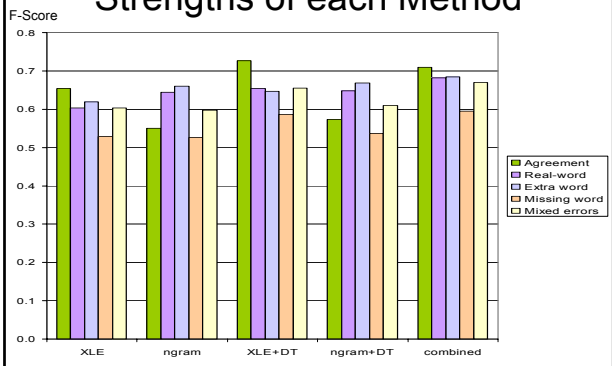
30

Talk Outline

- Motivation
- Background
- Artificial Error Corpus
- Evaluation Procedure
- Error Detection Methods
- **Results and Analysis**
- Conclusion and Future Work

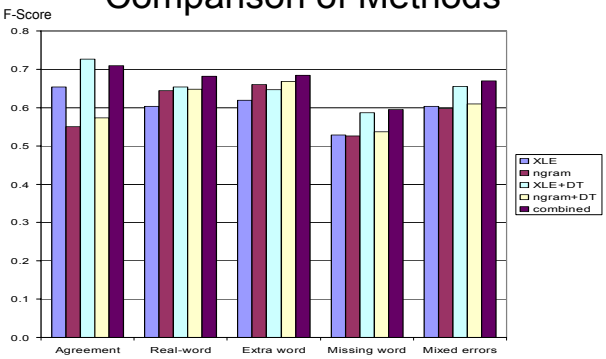
31

Strengths of each Method



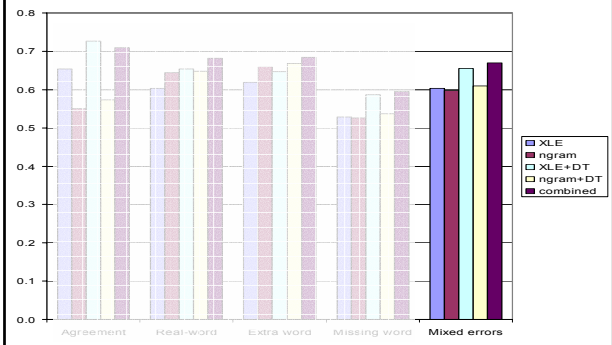
32

Comparison of Methods



33

Results: F-Score



34

Talk Outline

- Motivation
- Background
- Artificial Error Corpus
- Evaluation Procedure
- Error Detection Methods
- Results and Analysis
- **Conclusion and Future Work**

35

Conclusions

- Basic methods surprisingly close to each other
- Decision tree effective with deep approach
- Combined approach best on all but one error type

36

Future Work

- Error types:
  - Word order
  - Multiple errors per sentence
- Add more features
- Other languages
- Test on MT output
- Establish upper bound

37

Thank You!

ICHEC
Irish Centre for High-End Computing

embarkinitiative
Investing in People and Ideas

Djamé Seddah
(La Sorbonne University)

BRITISH NATIONAL CORPUS

National Centre for Language Technology
School of Computing, Dublin City University

embarkinitiative

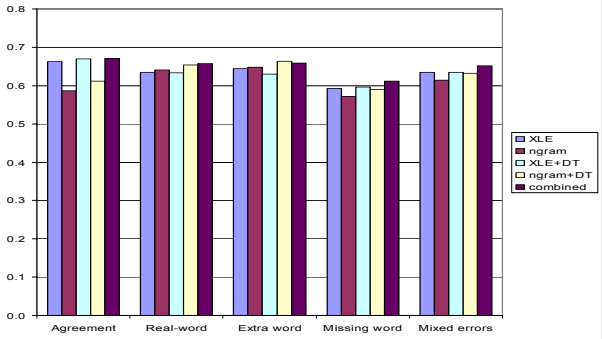
38

Extra Slides

- P/R/F/A graphs
- More on why judge grammaticality
- Precision Grammars in CALL
- Error creation examples
- Variance in cross-validation runs
- Precision over recall graphs (M3)
- More future work

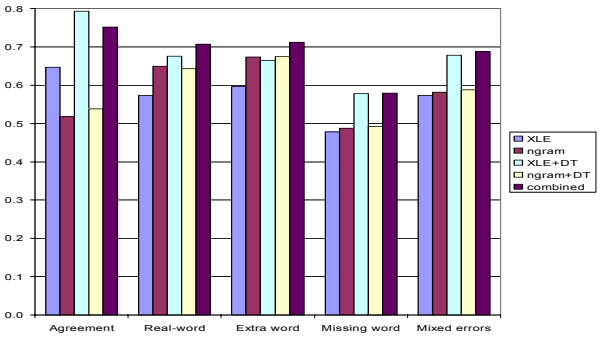
39

Results: Precision



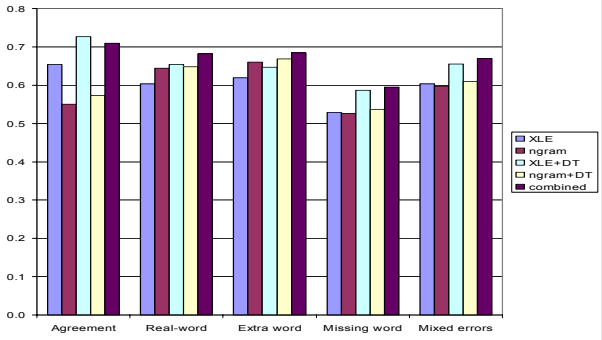
40

Results: Recall



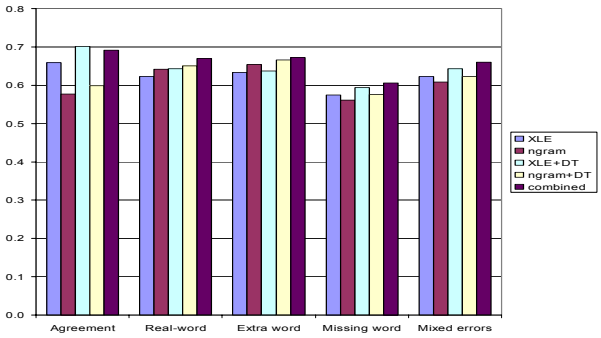
41

Results: F-Score



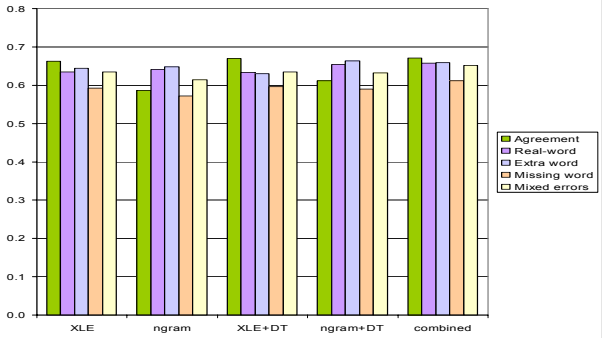
42

Results: Accuracy



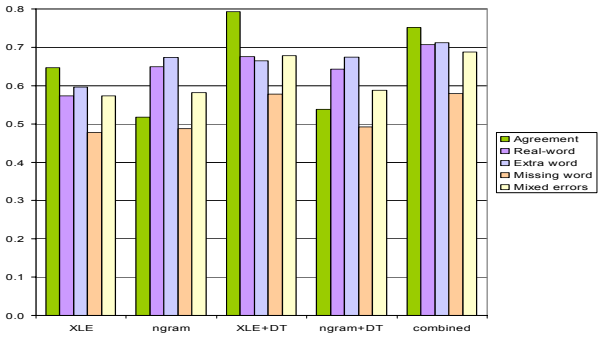
43

Results: Precision



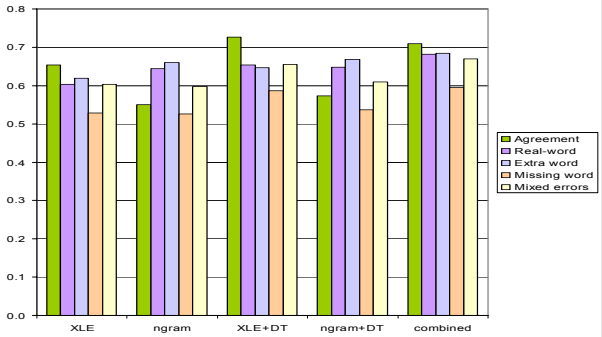
44

Results: Recall



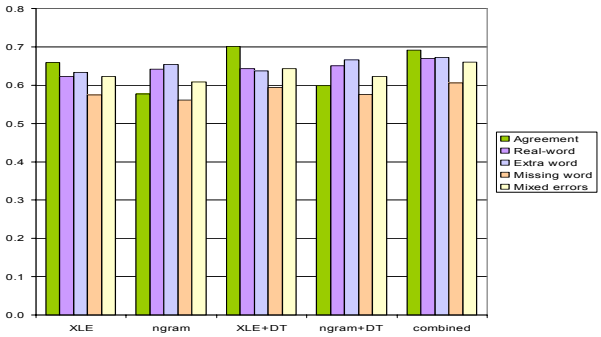
45

Results: F-Score



46

Results: Accuracy



47

Why Judge Grammaticality? (2)

- Automatic essay grading
- Trigger deep error analysis
  - Increase speed
  - Reduce overflagging
- Most approaches easily extend to
  - Locating errors
  - Classifying errors

48

Precision Grammars in CALL

- Focus:
  - Locate and categorise errors
- Approaches:
  - Extend existing grammars
  - Write new grammars

49

Grammar Checker Research

- Focus of grammar checker research
  - Locate errors
  - Categorise errors
  - Propose corrections
  - Other feedback (CALL)

50

N-gram Methods

- Flag unlikely or rare sequences
  - POS (different tagsets)
  - Tokens
  - Raw frequency vs. mutual information
- Most publications are in the area of context-sensitive spelling correction
  - Real word errors
  - Resulting sentence can be grammatical

51

Test Corpus - Example

- Missing Word Error

She didn't **want** to face him

↓

She didn't to face him

52

Test Corpus – Example 2

- Context-sensitive spelling error

I love **them** both

↓

I love **then** both

53

Cross-validation

- Standard deviation below 0.006
- Except Method 4: 0.026
- High number of test items
- Report average percentage

54

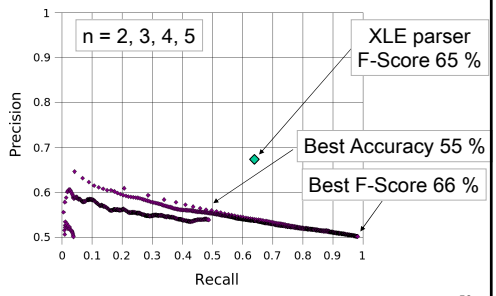
Example

Run	F-Score
1	0.654
2	0.655
3	0.655
4	0.655
5	0.653
6	0.652
7	0.653
8	0.657
9	0.654
10	0.653
Stdev	**0.001**

Method 1 – Agreement errors:
65.4 % average F-Score

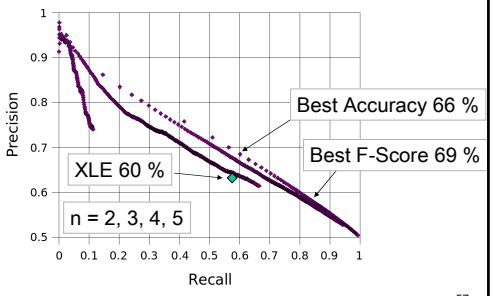
55

POS n-grams and Agreement Errors



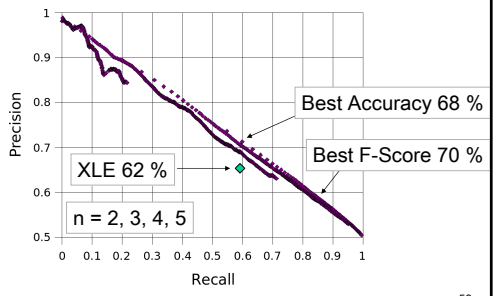
56

POS n-grams and Context-Sensitive Spelling Errors



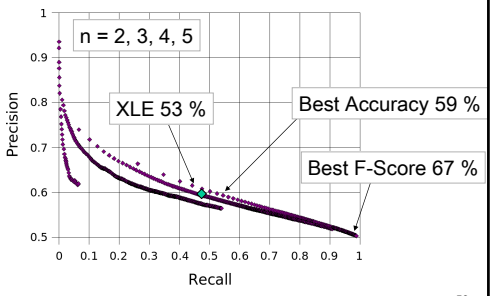
57

POS n-grams and Extra Word Errors



58

POS n-grams and Missing Word Errors



59