Using Language Models to Detect Errors in Second-Language Learner Writing

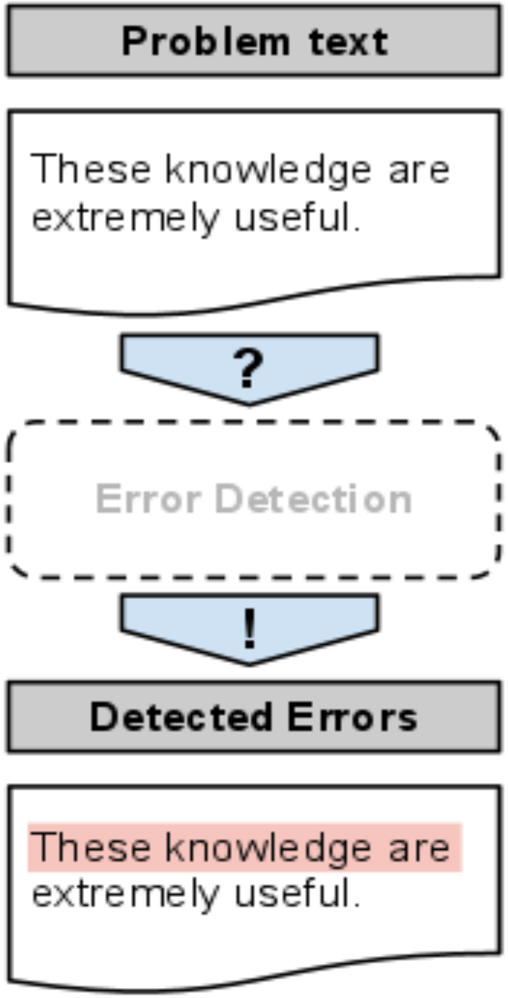
Nils Rethmeier

Bauhaus Universität Weimar
Web Technology and Information Systems Group

Motivation

Problem: We wrote a text but do not know if and **where we made errors**.

Task: **Find** the **errors** in the text.



Agenda

Error Detection Background

- Error Types
- Language Model, Class-based Language Model
- Combination Models

Detection Performance Measures

- Precision, recall
- Sentence and word level

Test Collections to determine performance

- English learner errors and artificially generated errors

Evaluation Results

- Influence of algorithmic parameters on detection results
- Comparison to error detection performed by humans

Summary

Error Detection Background

Error Categories

There is **no standardized definition** for **writing errors**.
However, we organized errors into one of four general categories.

Grammar and Word Usage Errors[1]
- Wrong articles, faulty wording, word countability problems (detected)
- Wrong word order, punctuation mistakes (partially detected)

Spelling Errors[2]
- Non-word errors, e.g. "Wykipedia" (detected)
- Real-word errors, e.g. "their", instead of "there" (detected)

Semantic Errors
- Are errors in meaning, e.g. bees are mammals (not detected)

Style Errors
- Writing that hinders understanding and reading, e.g. grandiloquence, overlong sentences (not detected)

1 C. Leacock, "Automated Grammatical Error Detection for Language Learners," Synthesis Lectures on Human Language Technologies, 2010
2 D. Fossati and B. Di Eugenio, "A mixed Trigrams Approach for Context Sensitive Spell Checking", 2010

Error Detection Background

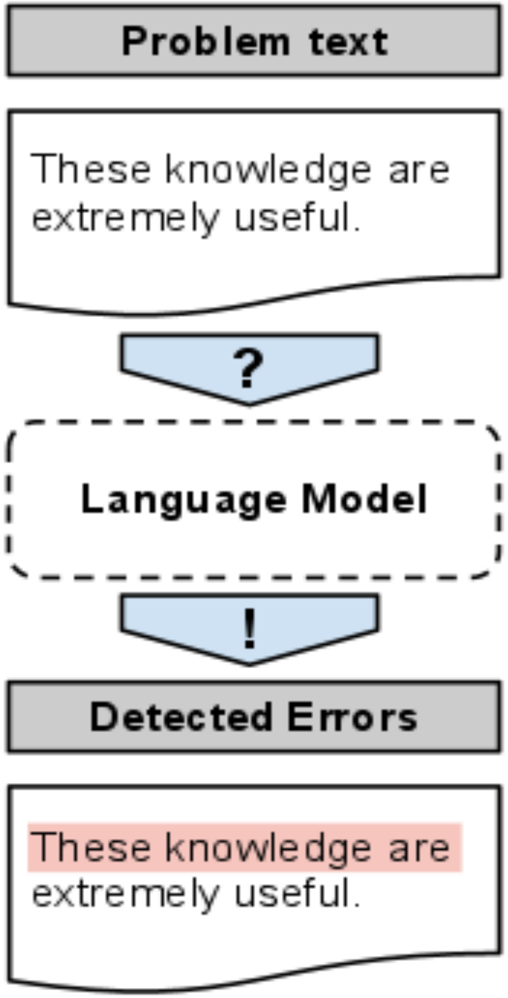
Error Detection Approaches

Human Annotation

- Professionals (Proofreading Services)
- Laymen (Friends, Mechanical Turk[1])

Computational Error Detection

- Rule based
  - Formal grammars[2]
- Statistical
  - **Word language models**
  - **Class-based language models**
  - **Combinations of both**

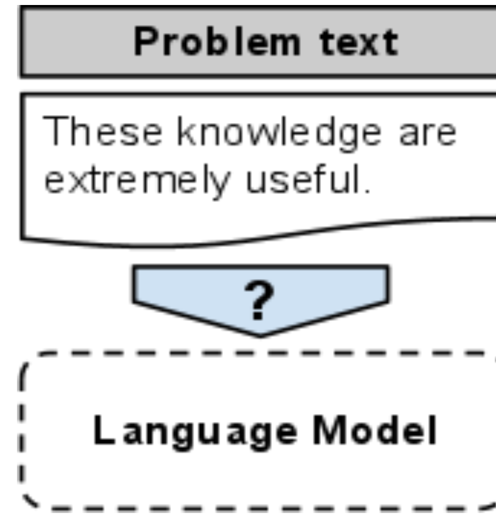


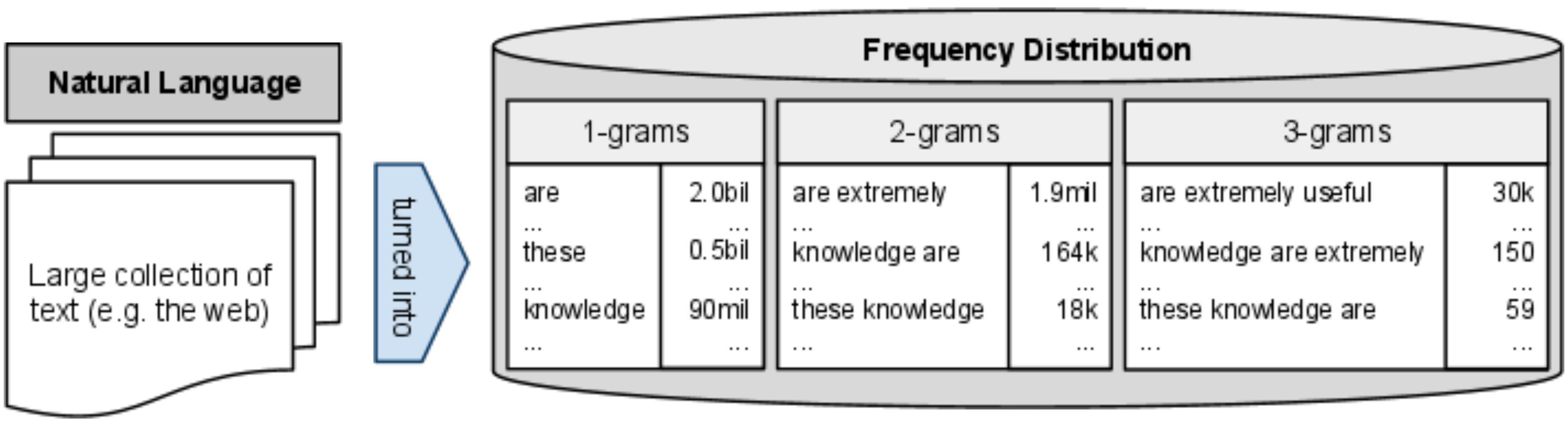
1 Amazon Mechanical Turk, https://www.mturk.com, as of Septemper 9, 2011
2 J. Wagner, A Comparative Evaluation of Deep and Shallow Approaches to the Automatic Detection of Common Grammatical Errors, 2007

Error Detection Background

Language Model: Frequency

A Language Model represents a natural language as a **frequency distribution** of word sequences (**word n-grams**).

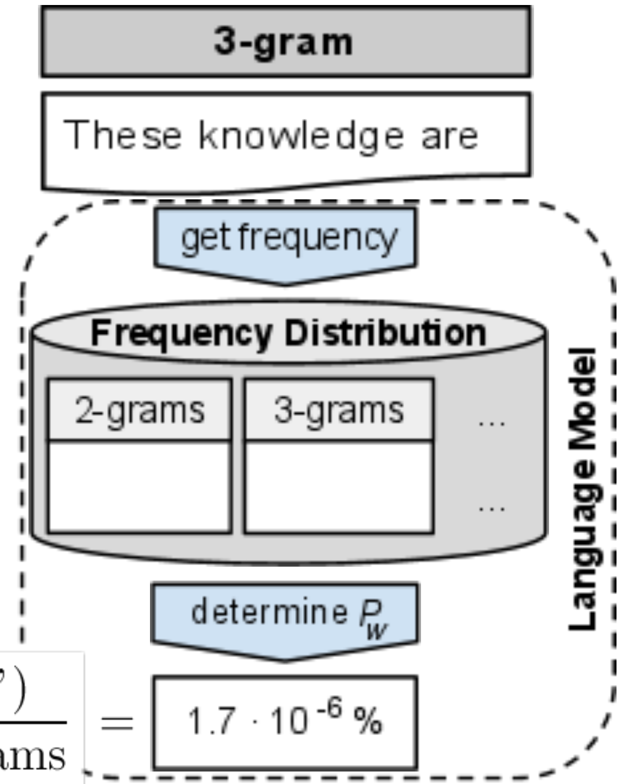




Error Detection Background

Language Model: Probability

How probable P_w is the 3-gram "these knowledge are" in the English language.



$$P_w(\text{"these knowledge are"}) = \frac{\text{Frequency}(\text{"these knowledge are"})}{\text{Total number of corpus word 3-grams}} =$$

Error Detection Background

Language Model: Backoff

For some 3-grams $P_w = 0.0\%$, because the frequency is 0.

Problem:

We do not know if the language model is missing the frequency because:

- The n-gram is incorrect language
- Our text collection is incomplete, i.e. does not contain this part of the language

Solution: Estimate a probability using Backoff[1]

$$P_w(\text{"these knowledge were"}) = 0.0\%$$

$$P_w(\text{"these knowledge were"}) \approxeq 0.4 \cdot P_w(\text{"knowledge were"})$$

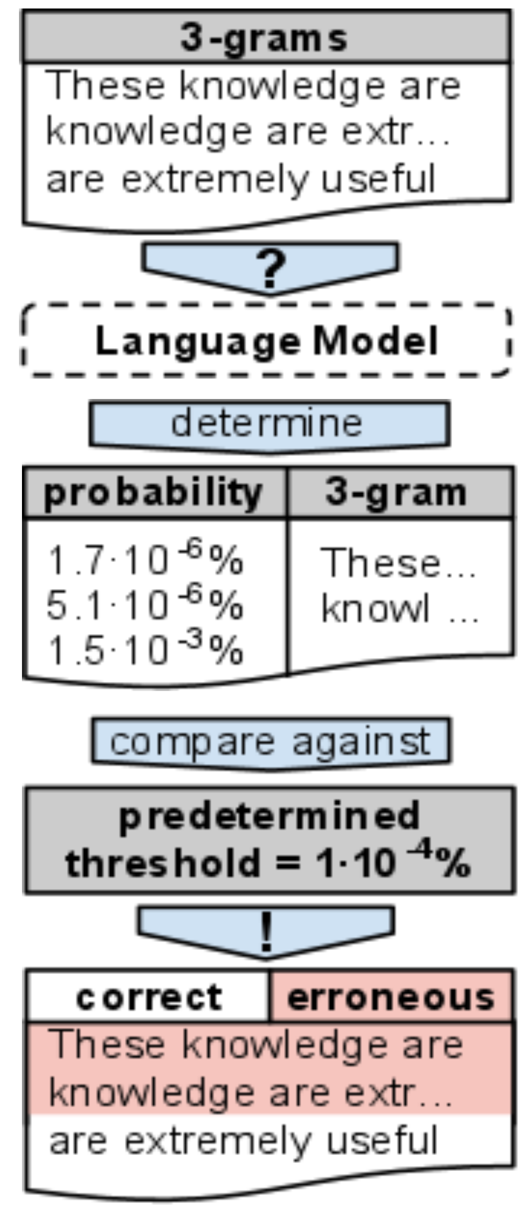
$$P_w(\text{"these knowledge were"}) \approx 0.4 \cdot 7.5 \cdot 10^{-4} = .3 \cdot 10^{-4}$$

1 Google's Stupid Backoff technique from: "Brants, T and Popat, A.C., Large language models in machine translation, 2007"

Error Detection Background

Probabilities for binary text classification:

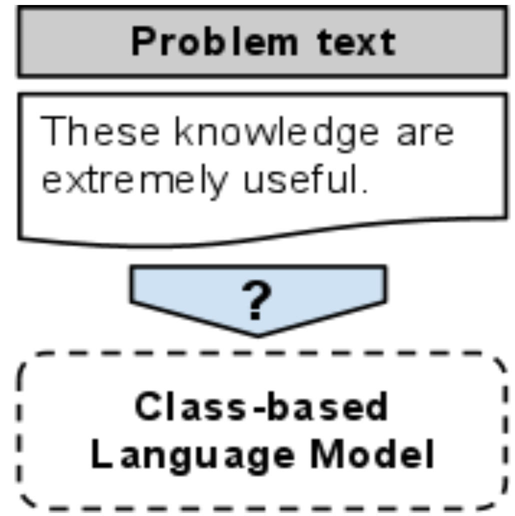
Comparing a text's n-gram probabilities against a predetermined threshold classifies these n-grams into correct and erroneous.



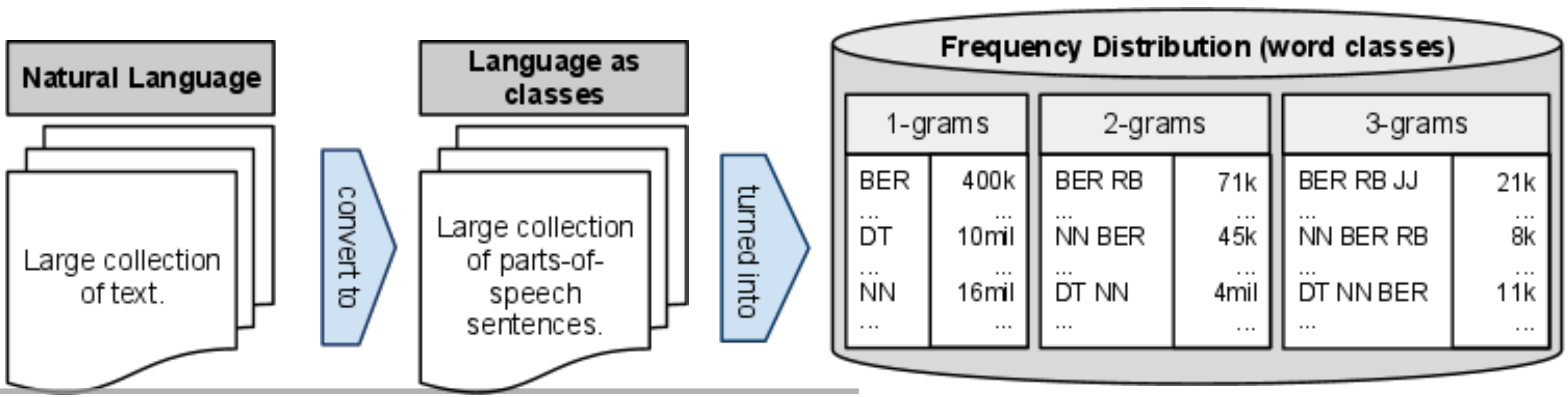
Error Detection Background

Class-based Language Model: Frequency

A model that represents language as a **frequency distribution** of word class sequences (**class n-grams**).



Example:
"These knowledge are" has the word classes "DT NN BER"

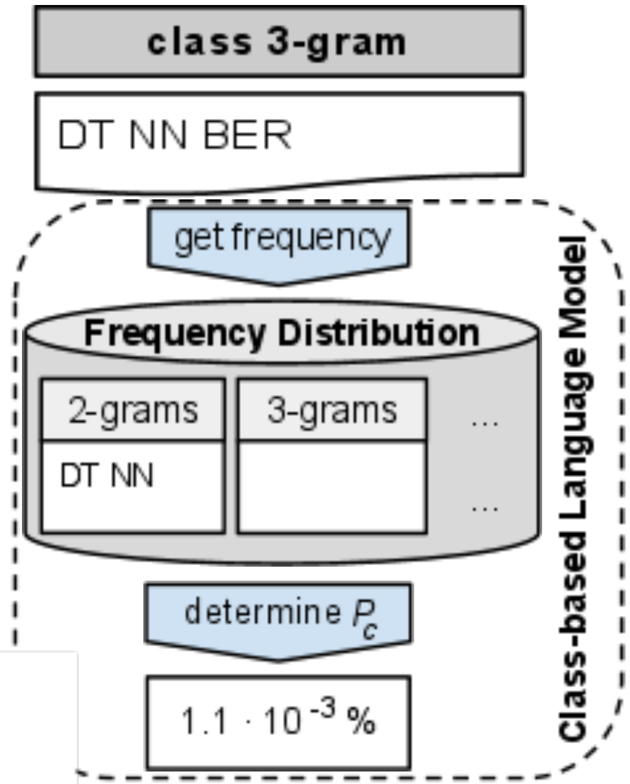


QTag parts-of-speech tags: DT = determiner, NN = noun, singular, BER = are, JJ = adjective, RB = adverb

Error Detection Background

Class-based Language Model: Probability

How probable P_c is the class 3-gram "DT NN BER" in the English language.



$$P_c(\text{"DT NN BER"}) = \frac{\text{Frequency("DT NN BER")}}{\text{Total number of corpus class 3-grams}} =$$

Combing Models:

Problem:

No Language Model represents a language exactly. This model sparseness leads to false detections.

Improvement:

Class-based models are less sparse[1] and can reduce false detections[2] when combined with word language models.

Combination methods[2] for P_c and P_w:

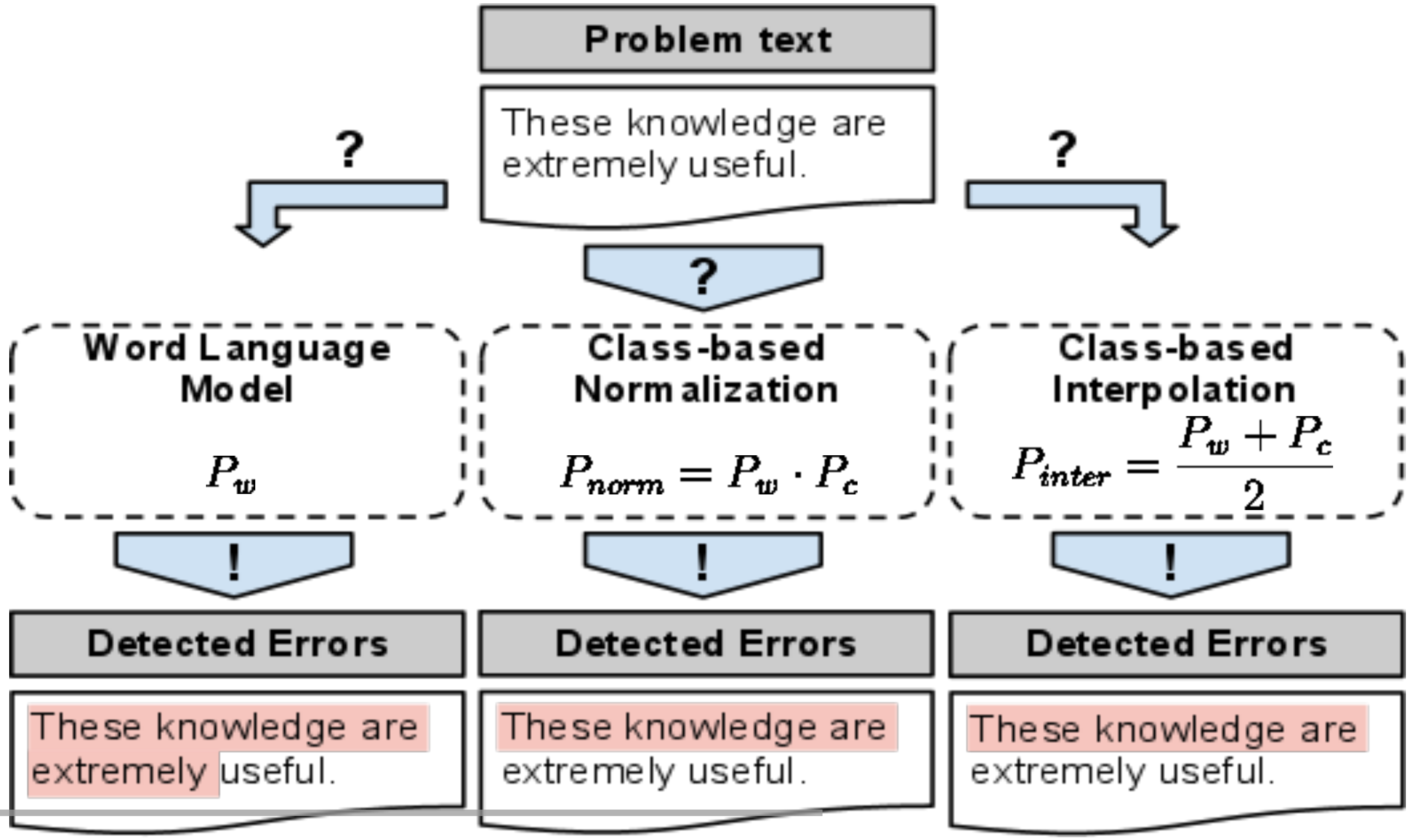
Normalization: $$P_{norm} = P_w \cdot P_c$$

Interpolation: $$P_{inter} = \frac{P_w + P_c}{2}$$

1 D. Jurafsky, Speech and Language Processing. Prentice Hall, 2 ed., May 2008
2 C. Samuelsson, "A class-based language model for large-vocabulary speech recognition extracted from part-of-speech statistics," 1999

Language Model Summary:

We looked at three different types of language models.



1 Detection results may differ by model. The above detections are only examples.

Agenda

Error Detection Background

- Error Types
- Language Model, Class-based Language Model
- Combination Models

Detection Performance Measures

- Precision, recall
- Sentence and word level

Test Collections to determine performance

- English learner errors and artificially generated errors

Evaluation Results

- Influence of algorithmic parameters on detection results
- Comparison to error detection performed by humans

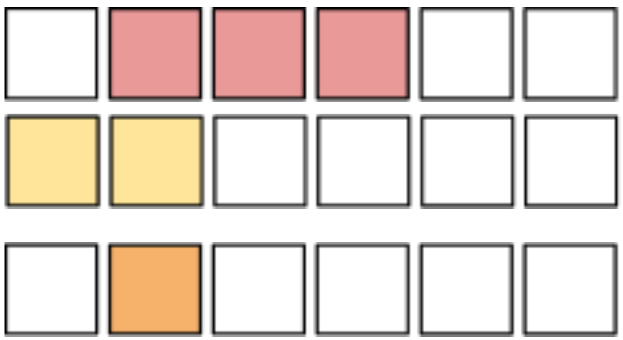
Summary

- Summary

Detection Performance Measures

Performance Measures

Recall measures what percentage of reference errors was detected. **Precision** measures how many error detections were indeed detected correctly.



Precision P

$$P = \frac{\text{Number of matches}}{\text{Number of detected errors}}$$

Recall R

$$R = \frac{\text{Number of matches}}{\text{Number of reference errors}}$$

Here

$$P = \frac{1 \cdot \square}{3 \cdot \square} = 0.33$$

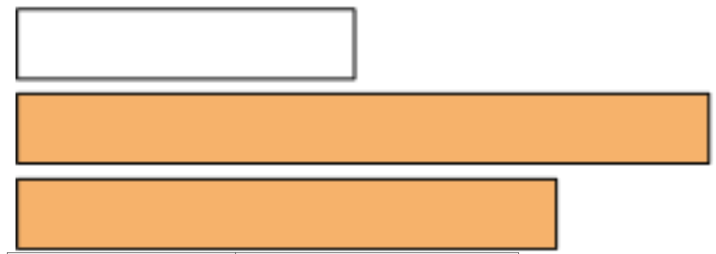
$$R = \frac{1 \cdot \square}{2 \cdot \square} = 0.50$$

Detection Performance Measures

Detection Granularity

Sentence level:

- Flags whole sentence as either grammatical or ungrammatical
- Common for detection evaluation
- No specific error locations



Precision	= 1.0
Recall	= 1.0

Word level:

- Each word is either grammatical or ungrammatical
- Measures specific error matches

Precision	= 0.2
Recall	= 0.5

Agenda

Error Detection Background
- Error Types
- Language Model, Class-based Language Model
- Combination Models

Detection Performance Measures
- Precision, recall
- Sentence and word level

Test Collections to determine performance
- English learner errors and artificially generated errors

Evaluation Results
- Influence of algorithmic parameters on detection results
- Comparison to error detection performed by humans

Summary

Test Collections

English Learner Corpora

Are collections of manually error annotated language learner writing.
We use them by extracting reference error positions from each corpus.

MELD[1]

- 58 learner essays (6,553 words)
- Sentences related
- Only a simple {error, correction} notation, no types

Artificially generated errors

10% British National Corpus of generated Errors (BNCd)[2]

- 9,413,338 words
- Each sentence contains one of four error types, e.g. spelling errors

1 E. Fitzpatrick and M. Seegmiller, "The **M**ontclair **E**lectronic **L**anguage **D**atabase project," Language and Computers, 2004
2 Wagner J., A Comparative Evaluation of Deep and Shallow Approaches to Automatic Error Detection, 2007

Agenda

Error Detection Background
- Error Types
- Language Model, Class-based Language Model
- Combination Models

Detection Performance Measures
- Precision, recall
- Sentence and word level

Test Collections to determine performance
- English learner errors and artificially generated errors

Evaluation Results
- Influence of algorithmic parameters on detection results
- Comparison to error detection performed by humans

Summary

Evaluation Results

Evaluation Framework:

- **Performance measures** (precision, recall)
- Trainingset 80% BNCd[1]
  - Trained a probability threshold that classify text n-grams with maximum overall performance (F1-score)
- **Testsets**
  - 10% BNCd (9.4mil words), artificial errors
  - MELD[2] (6.5k words), learner errors

Influence of algorithmic parameters on detection performance (BNCd)**:**

- N-gram length (3, 4-grams)
- Best detection model (language model, normalization, interpolation)
- Text error density (percent of errors in a text)

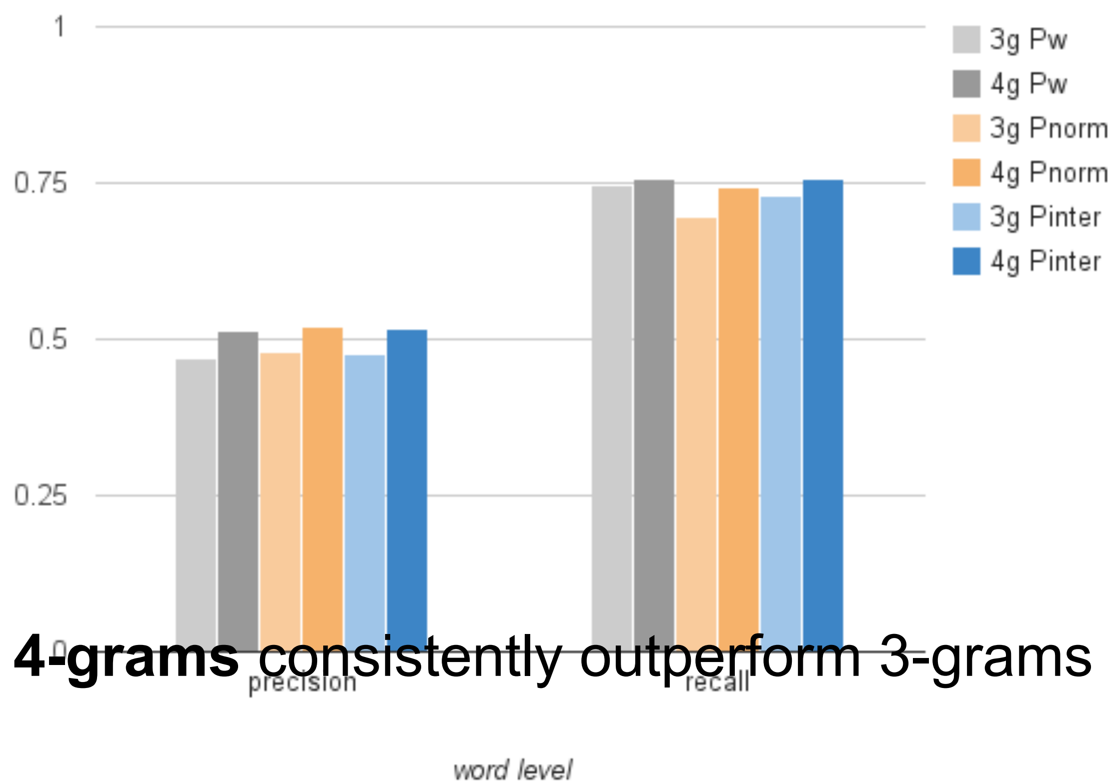
Detection performance comparison

- algorithmic detection vs. professional annotators (MELD)

1 Wagner J., *A Comparative Evaluation of Deep and Shallow Approaches to Automatic Error Detection*, 2007
2 E. Fitzpatrick and M. Seegmiller, "The **M**ontclair **El**ectronic **L**anguage **D**atabase project," Language and Computers, 2004

Evaluation Results

N-Gram Length (drawn from BNCd)

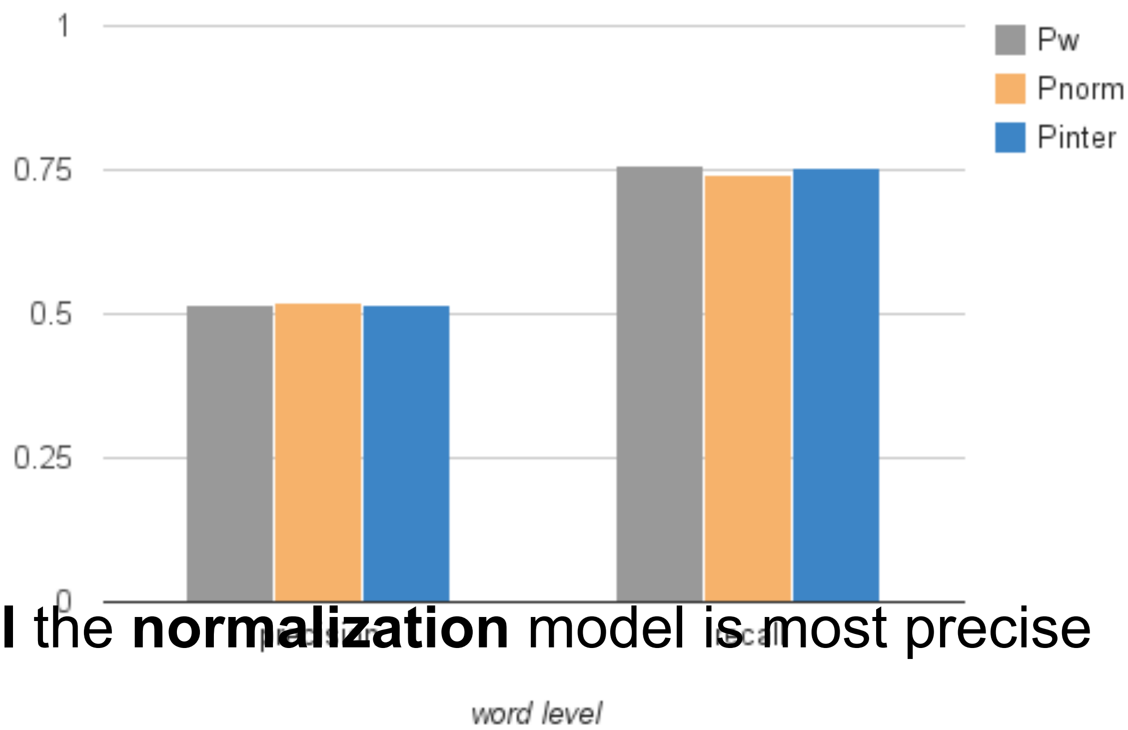


Conclusion:

- at **word level 4-grams** consistently outperform 3-grams

Evaluation Results

Standard vs. Combination Model (BNCd)

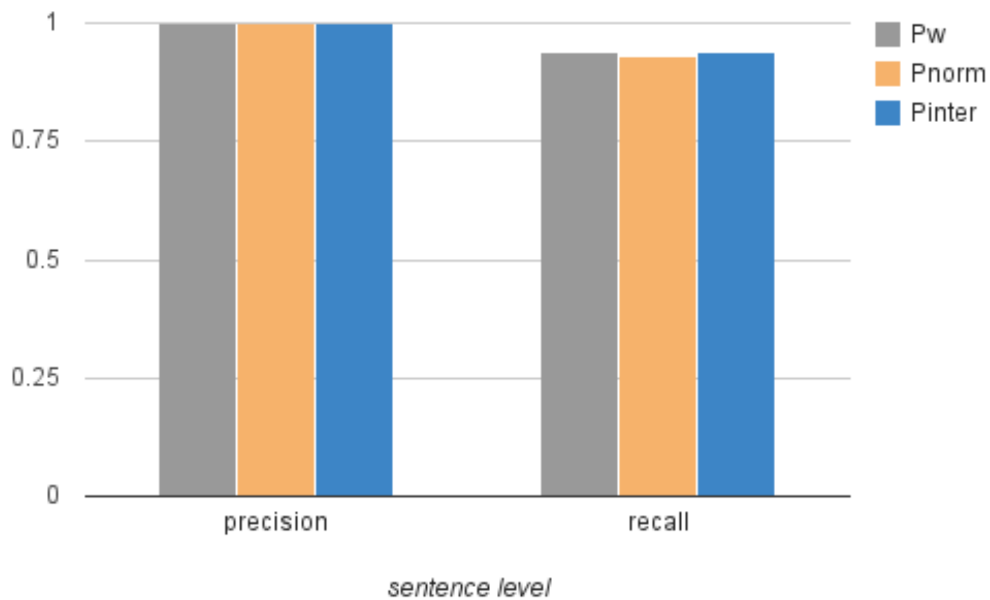


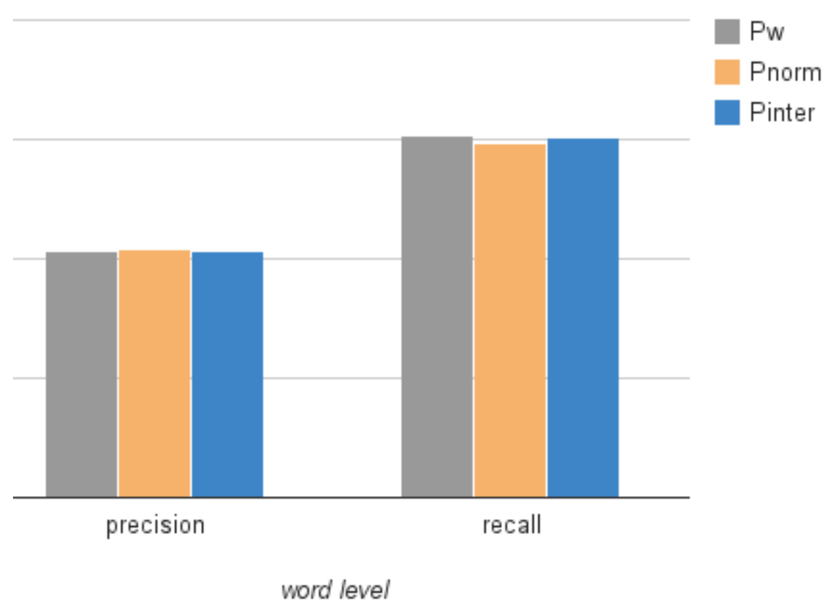
Conclusion:

- at **word level** the **normalization** model is most precise

Evaluation Results

Problems at sentence level (BNCd)

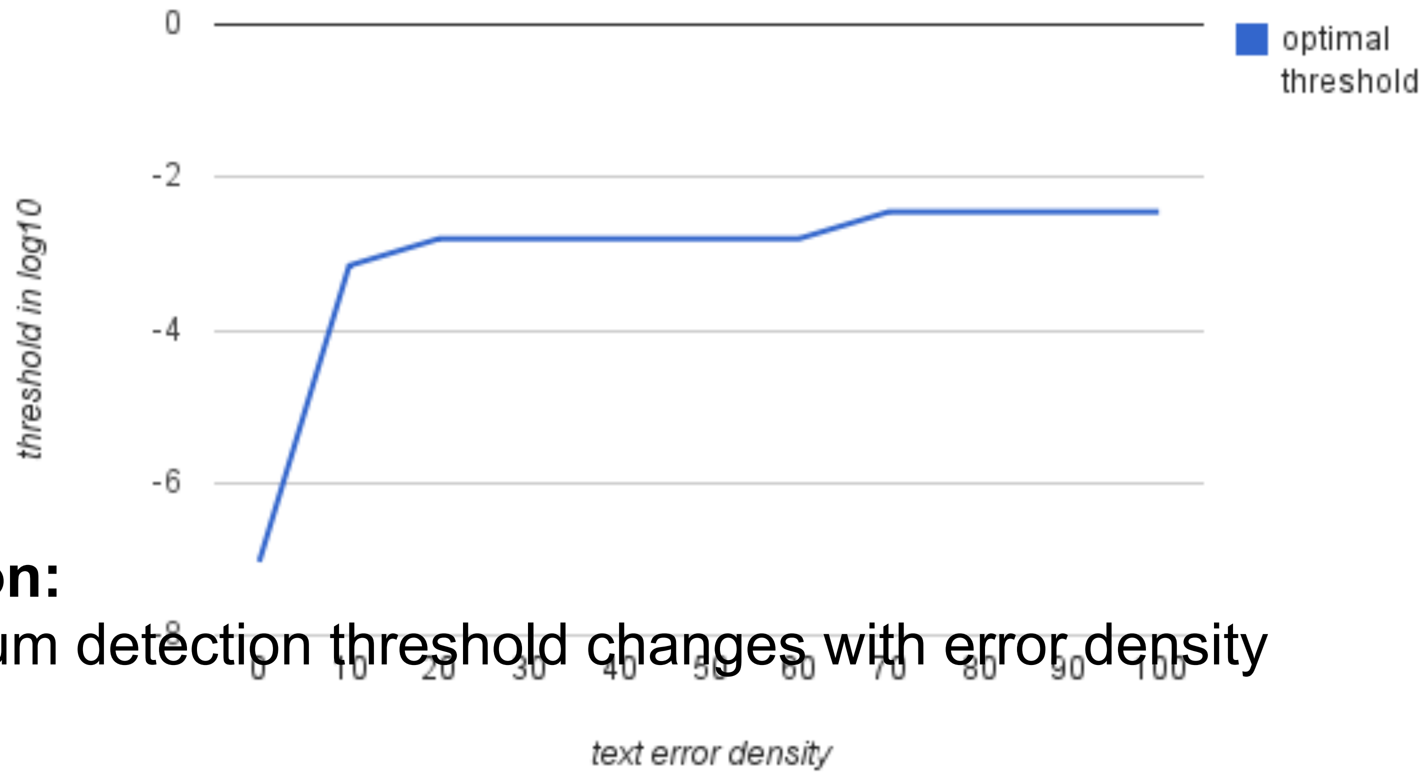




- Sentence level detection is not a good indicator of quality

Evaluation Results

Optimal threshold in relation to a text's error density.



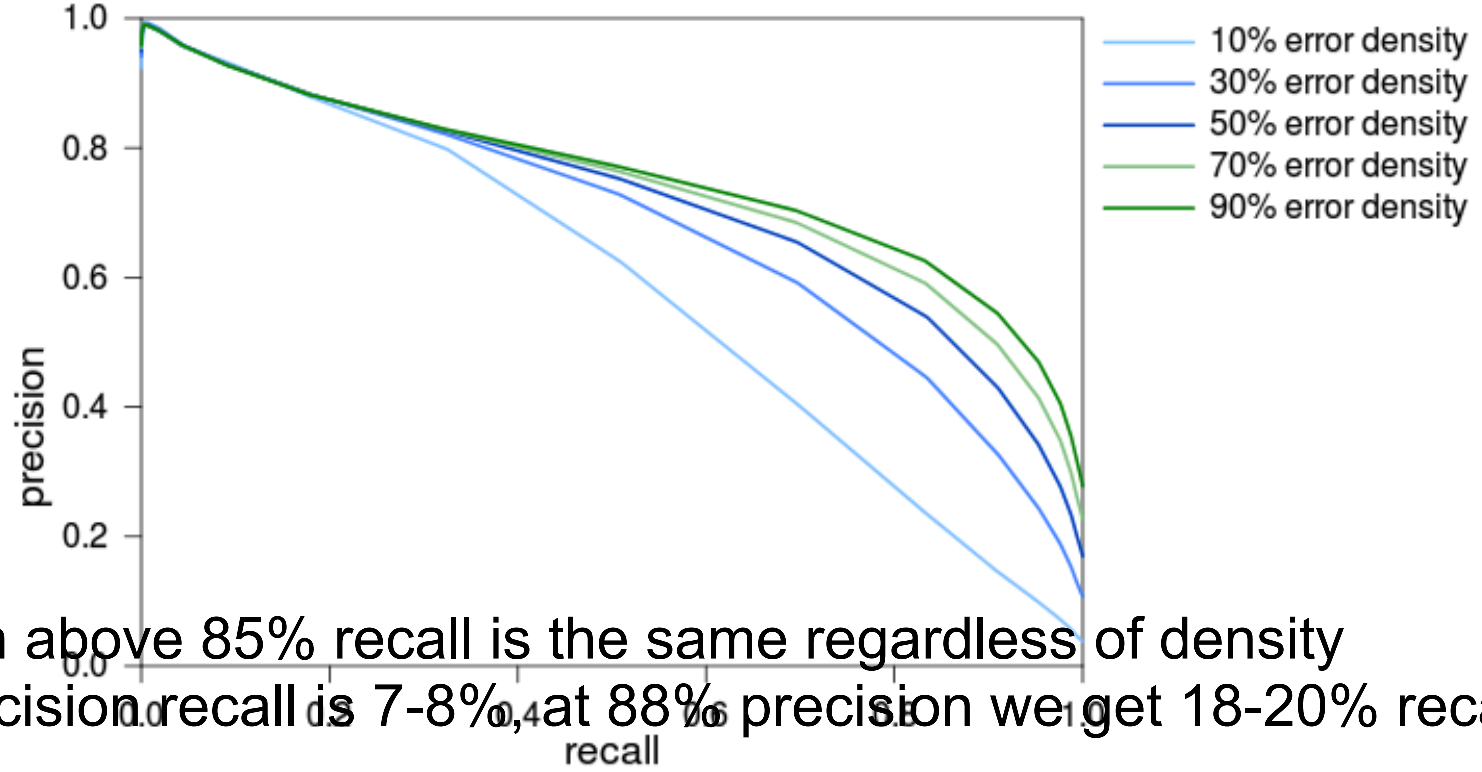
Conclusion:

- Optimum detection threshold changes with error density

Shown model uses linear **interpolation** to combine **word** and **part-of-speech** probabilities. Model with highest precision.

Evaluation Results

Precision in relation to recall.



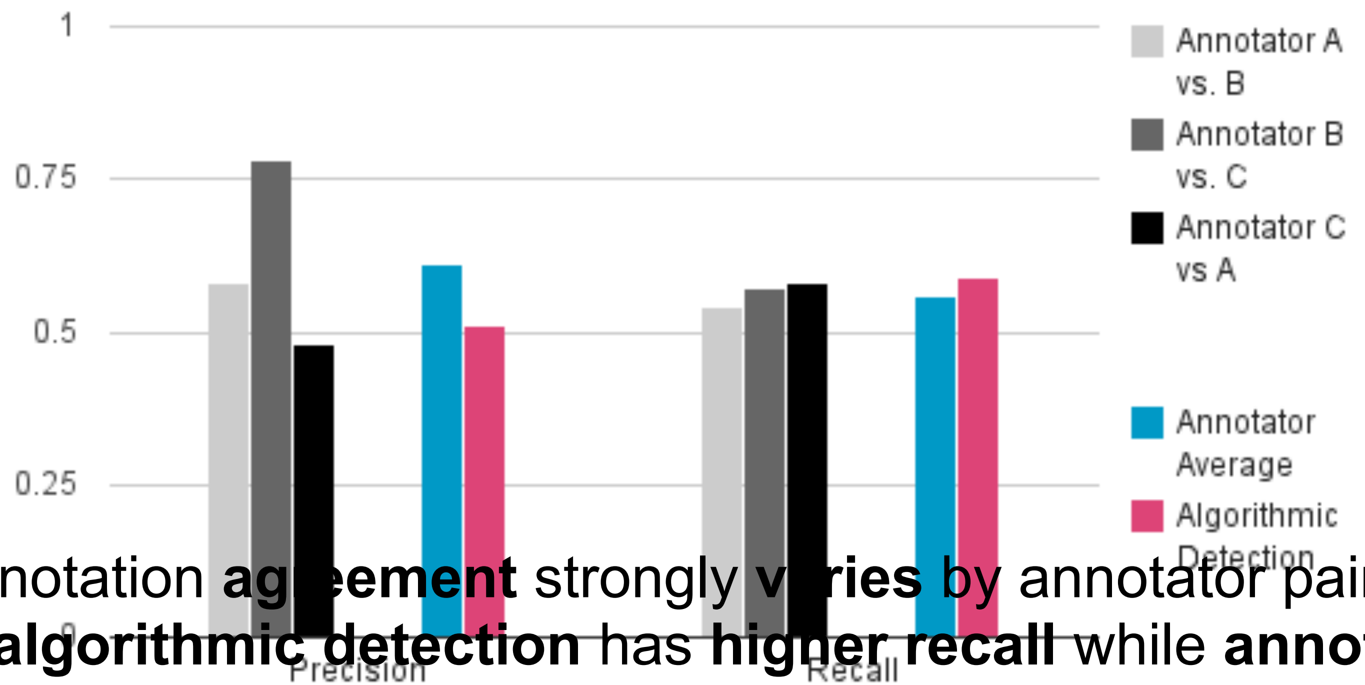
Conclusion:

- At precision above 85% recall is the same regardless of density
- At 95% precision recall is 7-8%, at 88% precision we get 18-20% recall

Shown model uses linear **interpolation** to combine **word** and **part-of-speech** probabilities. Model with highest precision.

Evaluation Results

Agreement between professional annotators vs. algorithmic detection (MELD)



Conclusion:

- **Human** annotation **agreement** strongly **varies** by annotator pairs
- On MELD **algorithmic detection** has **higher recall** while **annotators** achieve significantly **higher precision** on average

Summary

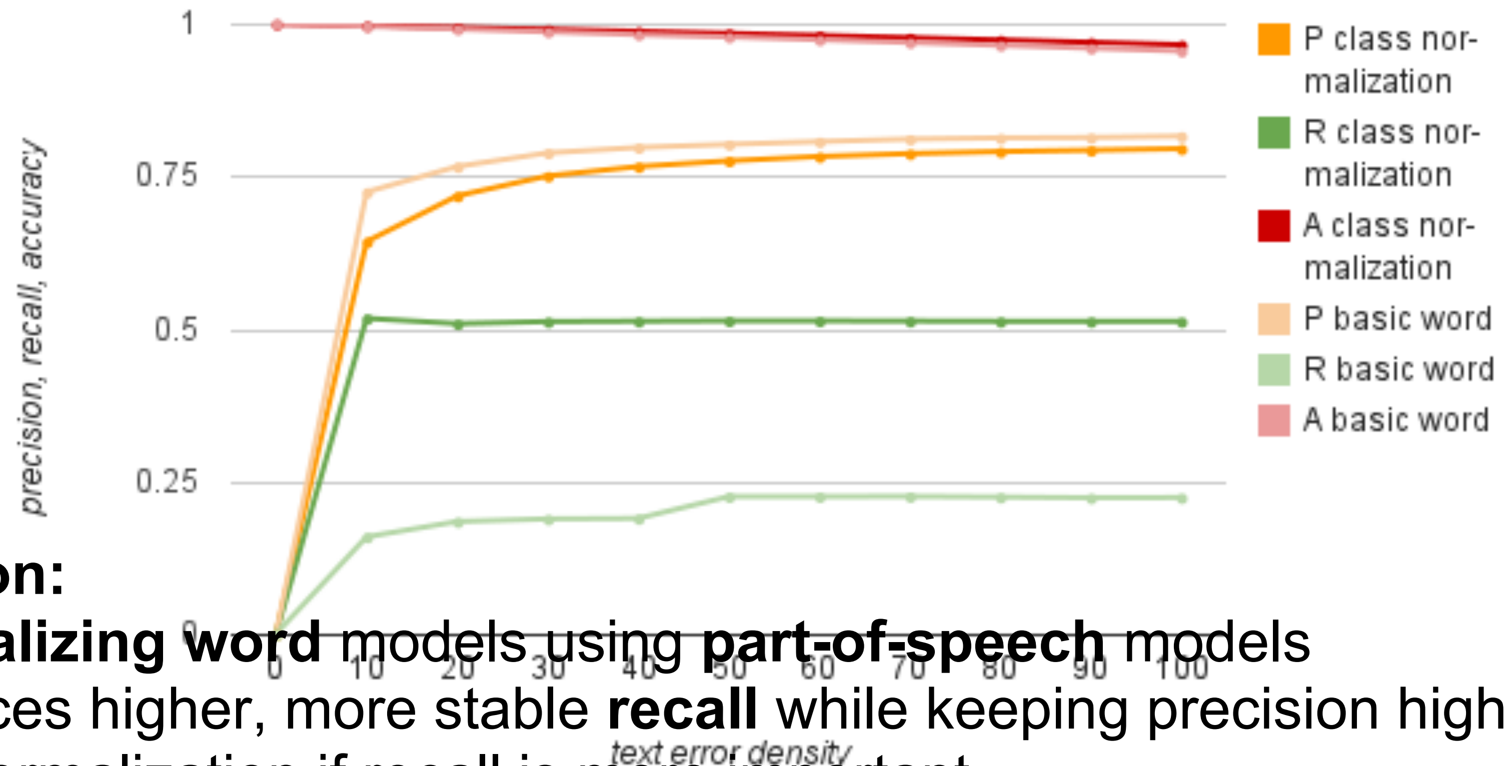
Result Summary

- Investigated impact of model combinations on detection performance
  - combination models outperform word language models
- Explored the impact of a text's error density on language model based error detection (usually not regarded)
- Investigated algorithmic detection performance when compared to humans

Thank you for listening

Future Work: Model Comparison Revised

Improvement in detection recall compared to the basic word model.



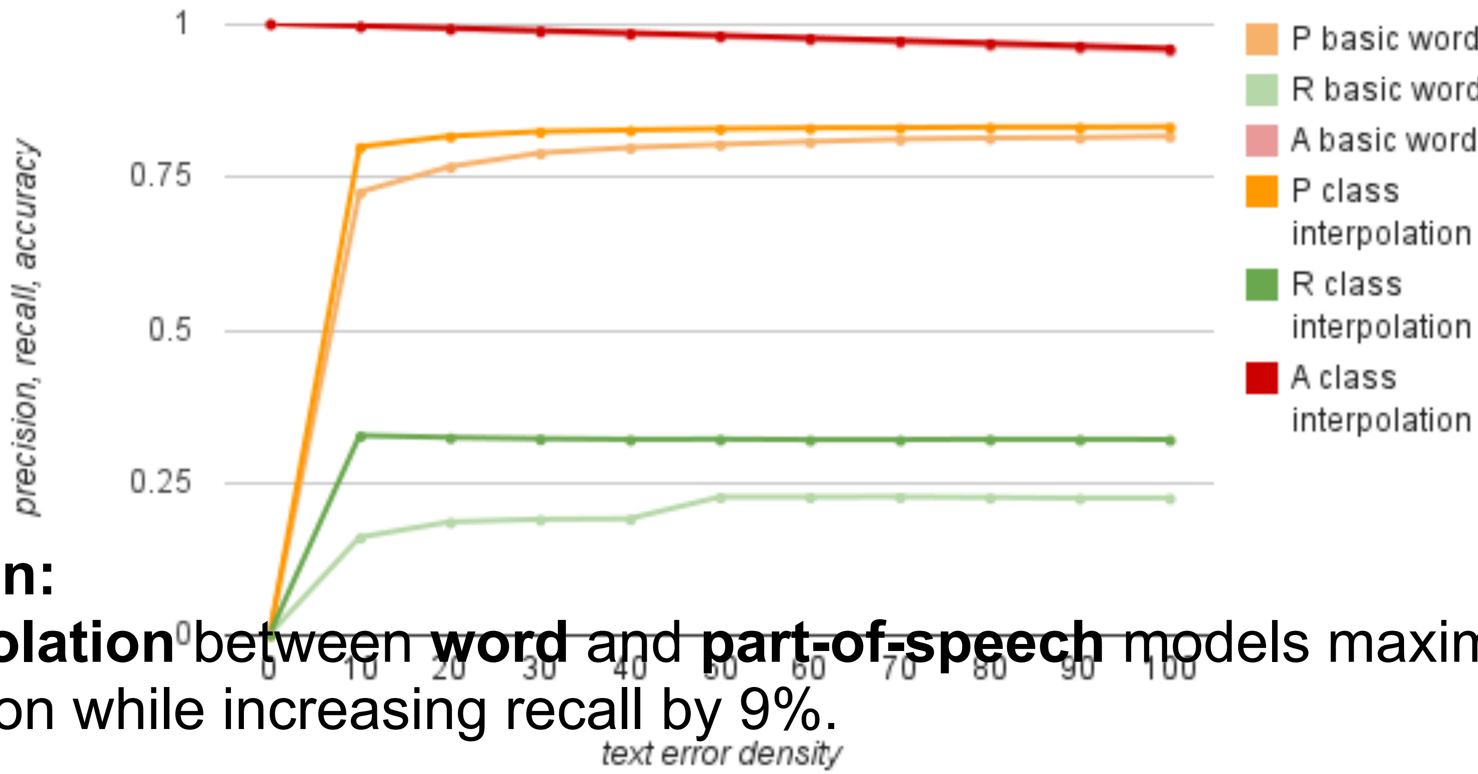
Conclusion:

- **Normalizing word** models using **part-of-speech** models produces higher, more stable **recall** while keeping precision high
- Use normalization if recall is more important

Shown model uses **normalization** to combine **word** and **part-of-speech** probabilities. Model with highest f1-score.

Evaluation Results

Improvements in error detection precision.



Conclusion:

- **Interpolation** between **word** and **part-of-speech** models maximizes precision while increasing recall by 9%.

Shown model uses linear **interpolation** to combine **word** and **part-of-speech** probabilities. Model with highest precision.