Complex Bigram Model for Student-Classroom Text Analysis

1. Introduction and Mathematical Foundation

This document presents a comprehensive bigram language model trained on student-classroom interaction sentences. Bigram models predict word probabilities based on the immediately preceding word, making them valuable for educational text analysis and classroom language modeling.

Mathematical Foundation: The bigram probability using Maximum Likelihood Estimation (MLE):

$$P(w_i | w_{i-1}) = Count(w_{i-1}, w_i) / Count(w_{i-1})$$

Add-One (Laplace) Smoothing: $P_{smooth}(w_i | w_{i-1}) = (Count(w_{i-1}, w_i) + 1) / (Count(w_{i-1}) + V)$

Where V is the vocabulary size.

2. Training and Test Data

Training Sentences:

1. "The student submitted homework to the teacher in the classroom"
2. "The teacher explained the lesson to all students during class"
3. "The student asked questions about the assignment after class"
4. "The teacher graded homework and returned it to students"

Test Sentence: "The student completed the assignment"

2.1 Data Preprocessing

Processed Training Sentences (with sentence boundary tokens):

1. `<s> the student submitted homework to the teacher in the classroom </s>`
2. `<s> the teacher explained the lesson to all students during class </s>`
3. `<s> the student asked questions about the assignment after class </s>`
4. `<s> the teacher graded homework and returned it to students </s>`

Processed Test Sentence: `<s> the student completed the assignment </s>`

Corpus Statistics:

- Total tokens: 54 (including boundary tokens)
- Training sentences: 4

- Average sentence length: 10.5 words
- Test sentence length: 6 words

3. Unigram Analysis

3.1 Complete Unigram Frequency Table

Word	Frequency	Probability	Rank
the	10	0.1852	1
to	4	0.0741	2
student	3	0.0556	3
teacher	3	0.0556	3
homework	2	0.0370	5
class	2	0.0370	5
students	2	0.0370	5
`<s>`	4	0.0741	2
`</s>`	4	0.0741	2
submitted	1	0.0185	10
in	1	0.0185	10
classroom	1	0.0185	10
explained	1	0.0185	10
lesson	1	0.0185	10
all	1	0.0185	10
during	1	0.0185	10
asked	1	0.0185	10
questions	1	0.0185	10
about	1	0.0185	10
assignment	1	0.0185	10
after	1	0.0185	10
graded	1	0.0185	10
and	1	0.0185	10
returned	1	0.0185	10
it	1	0.0185	10
completed	0	0.0000	-

3.2 Unigram Statistical Analysis

Total Word Count: 54 tokens **Vocabulary Size (V):** 26 unique words (including unseen "completed") **Most Frequent Words:** "the" (18.52%), "to", "<s>", "</s>" (7.41% each)

Zipf's Law Analysis:

- Rank 1: "the" - 18.52%
- Rank 2: "to", "<s>", "</s>" - 7.41% each
- Rank 3: "student", "teacher" - 5.56% each

Educational Domain Vocabulary Distribution:

- Core entities: student (5.56%), teacher (5.56%), students (3.70%)
- Academic activities: homework (3.70%), class (3.70%), lesson (1.85%)
- Actions: submitted, explained, asked, graded, returned (1.85% each)

4. Bigram Analysis

4.1 Complete Bigram Frequency Table

Bigram	Count	Raw Probability	Add-1 Smoothed
`<s> the`	4	1.0000	0.1667
the student	2	0.2000	0.0968
the teacher	2	0.2000	0.0968
the lesson	1	0.1000	0.0645
the assignment	1	0.1000	0.0645
the classroom	1	0.1000	0.0645
student submitted	1	0.3333	0.1034
student asked	1	0.3333	0.1034
student completed	0	0.0000	0.0345
submitted homework	1	1.0000	0.0370
homework to	1	0.5000	0.0741
homework and	1	0.5000	0.0741
to the	3	0.7500	0.1333
to all	1	0.2500	0.0667
teacher explained	1	0.3333	0.1034
teacher in	1	0.3333	0.1034

Bigram	Count	Raw Probability	Add-1 Smoothed
teacher graded	1	0.3333	0.1034
explained the	1	1.0000	0.0370
lesson to	1	1.0000	0.0370
all students	1	1.0000	0.0370
students during	1	0.5000	0.0741
students </s>	1	0.5000	0.0741
during class	1	1.0000	0.0370
class </s>	2	1.0000	0.1111
asked questions	1	1.0000	0.0370
questions about	1	1.0000	0.0370
about the	1	1.0000	0.0370
assignment after	1	1.0000	0.0370
after class	1	1.0000	0.0370
graded homework	1	1.0000	0.0370
and returned	1	1.0000	0.0370
returned it	1	1.0000	0.0370
it to	1	1.0000	0.0370
in the	1	1.0000	0.0370
classroom </s>	1	1.0000	0.0370

4.2 Bigram Probability Calculations

Raw MLE Calculations:

- P(student | the) = Count(the student) / Count(the) = 2/10 = 0.2000
- P(teacher | the) = Count(the teacher) / Count(the) = 2/10 = 0.2000
- P(completed | student) = Count(student completed) / Count(student) = 0/3 = 0.0000

Add-1 Smoothed Calculations:

- P_smooth(student | the) = (2 + 1) / (10 + 26) = 3/36 = 0.0833
- P_smooth(teacher | the) = (2 + 1) / (10 + 26) = 3/36 = 0.0833
- P_smooth(completed | student) = (0 + 1) / (3 + 26) = 1/29 = 0.0345

4.3 Educational Domain Bigram Patterns

Strong Educational Collocations:

- "the student" (20% of "the" contexts)
- "the teacher" (20% of "the" contexts)
- "to the" (75% of "to" contexts)
- "homework to/and" (educational workflow patterns)

Classroom Activity Sequences:

- student → submitted/asked (academic actions)
- teacher → explained/graded (instructional actions)
- homework → to/and (assignment handling)

5. Test Sentence Analysis and Perplexity

5.1 Bigram Probability Calculation for Test Sentence

Test Sentence: <s> the student completed the assignment </s>

Bigram Decomposition with Add-1 Smoothing:

1. P(<s> the) = (4 + 1) / (4 + 26) = 5/30 = 0.1667
2. P(the | student) = (2 + 1) / (10 + 26) = 3/36 = 0.0833
3. P(student | completed) = (0 + 1) / (3 + 26) = 1/29 = 0.0345
4. P(completed | the) = (0 + 1) / (10 + 26) = 1/36 = 0.0278
5. P(the | assignment) = (1 + 1) / (1 + 26) = 2/27 = 0.0741
6. P(assignment | </s>) = (0 + 1) / (1 + 26) = 1/27 = 0.0370

5.2 Sentence Probability and Perplexity

Sentence Probability: P(sentence) = 0.1667 × 0.0833 × 0.0345 × 0.0278 × 0.0741 × 0.0370 P(sentence) = 7.93 × 10^(-7)

Detailed Calculation:

- Step 1: 0.1667 × 0.0833 = 0.01389
- Step 2: 0.01389 × 0.0345 = 0.000479
- Step 3: 0.000479 × 0.0278 = 0.0000133
- Step 4: 0.0000133 × 0.0741 = 9.86 × 10^(-7)

- Step 5: $9.86 \times 10^{-7} \times 0.0370 = 7.93 \times 10^{-7}$

Perplexity Calculation:

- Number of words: N = 6 (including boundary tokens)
- Log probability: $\ln(7.93 \times 10^{-7}) = -14.04$
- **Perplexity = $\exp(-(-14.04)/6) = \exp(2.34) = 10.38$**

5.3 Perplexity Interpretation

Model Performance Analysis:

- Perplexity of 10.38 indicates moderate confidence
- Lower perplexity would indicate better predictive performance
- The unseen word "completed" contributes to higher perplexity
- Educational domain vocabulary coverage affects model certainty

6. Next Word Prediction Examples

6.1 Predicting After "The student"

Context Analysis: Given "the student"

From training data bigrams:

- $P_{smooth}(\text{submitted} \mid \text{student}) = (1 + 1) / (3 + 26) = 2/29 = 0.0690$
- $P_{smooth}(\text{asked} \mid \text{student}) = (1 + 1) / (3 + 26) = 2/29 = 0.0690$
- $P_{smooth}(\text{completed} \mid \text{student}) = (0 + 1) / (3 + 26) = 1/29 = 0.0345$

Top Predictions:

1. "submitted" - 0.0690 (23.8%)
2. "asked" - 0.0690 (23.8%)
3. "completed" - 0.0345 (11.9%)
4. Other words - 0.0345 each

6.2 Predicting After "The teacher"

Context Analysis: Given "the teacher"

From training data:

- $P_{smooth}(\text{explained} \mid \text{teacher}) = (1 + 1) / (3 + 26) = 2/29 = 0.0690$

- P_smooth(graded | teacher) = (1 + 1) / (3 + 26) = 2/29 = 0.0690
- P_smooth(in | teacher) = (1 + 1) / (3 + 26) = 2/29 = 0.0690

Top Predictions:

1. "explained" - 0.0690 (instructional activity)
2. "graded" - 0.0690 (assessment activity)
3. "in" - 0.0690 (location context)

6.3 Educational Context Predictions

Predicting After "homework":

- P_smooth(to | homework) = 0.0741
- P_smooth(and | homework) = 0.0741

Predicting After "students":

- P_smooth(during | students) = 0.0741
- P_smooth(`</s>` | students) = 0.0741

7. Model Evaluation and Educational Domain Analysis

7.1 Coverage and Performance Statistics

Vocabulary Coverage:

- Training vocabulary: 25 unique words
- Test vocabulary: 26 words (including "completed")
- Out-of-vocabulary rate: 3.8% (1/26)
- Bigram coverage: 34 unique bigrams in training

Smoothing Effectiveness:

- Handles 100% of unseen bigrams
- Provides uniform probability floor for OOV words
- Maintains probability mass distribution across vocabulary

7.2 Educational Domain Insights

Academic Role Modeling:

- "Student" actions: submitted, asked, completed

- "Teacher" actions: explained, graded, returned
- Clear role-based activity patterns captured

Classroom Interaction Patterns:

- Homework workflow: student → submitted → homework → to → teacher
- Question pattern: student → asked → questions → about → assignment
- Grading cycle: teacher → graded → homework → and → returned

Temporal and Spatial Context:

- Time markers: during, after
- Location: classroom, class
- Activity sequencing: assignment → after → class

7.3 Educational Applications

Potential Use Cases:

1. **Automated Essay Scoring:** Predict academic writing patterns
2. **Classroom Chatbots:** Generate contextually appropriate responses
3. **Educational Content Generation:** Create classroom scenario text
4. **Student Writing Analysis:** Identify typical academic language patterns

Model Strengths:

- Captures educational terminology effectively
- Models student-teacher interactions accurately
- Handles academic workflow sequences
- Provides reasonable predictions for classroom contexts

Limitations:

- Limited training data leads to data sparsity
- High smoothing impact due to small corpus
- Cannot capture long-range educational dependencies
- Vocabulary limited to basic classroom scenarios

8. Conclusion and Future Enhancements

8.1 Key Findings

The bigram model successfully captures fundamental patterns in student-classroom interactions with a perplexity of 10.38. The model demonstrates effective learning of:

- **Educational roles** (student vs teacher activities)
- **Academic workflows** (homework submission and grading cycles)
- **Classroom interactions** (questions, explanations, assignments)

8.2 Mathematical Summary

Core Model Statistics:

- Training corpus: 54 tokens, 25 unique words
- Bigram types: 34 unique bigrams
- Test sentence probability: 7.93×10^{-7}
- Model perplexity: 10.38
- Smoothing parameter: $\alpha = 1$ (add-one Laplace)

Performance Metrics:

- Vocabulary coverage: 96.2%
- Unseen bigram handling: 100% via smoothing
- Educational domain accuracy: High for basic classroom scenarios

8.3 Future Improvements

Model Enhancements:

1. **Larger Training Corpus:** More diverse educational scenarios
2. **Advanced Smoothing:** Kneser-Ney or interpolated smoothing
3. **Higher-order N-grams:** Trigrams for better context modeling
4. **Domain Adaptation:** Specialized educational vocabulary weighting

Educational Applications:

- Integration with learning management systems
- Adaptive educational content generation
- Student writing assessment tools

- Intelligent tutoring system language components

The model provides a solid foundation for educational text analysis and demonstrates the effectiveness of n-gram models in capturing domain-specific language patterns in academic settings.