AIA5320 Natural Language Processing
Assignment 1: Word Analogy
Report

Student ID: R13921031
Name: James Christian

October 1, 2025

1 Model Settings and Hyperparameters (5%)

1.1 Embedding Model

We used Word2Vec (Gensim implementation) with the Continuous Bag of Words (CBOW) algorithm for training custom word embeddings.

1.2 Pre-processing Steps

The following preprocessing pipeline was applied to the Wikipedia corpus:

1. **Lowercasing**: Convert all text to lowercase for consistency with evaluation requirements
2. **Non-alphabetic removal**: Remove punctuation and special characters using regex pattern `[^a-z\s]`
3. **Tokenization**: Split text by whitespace
4. **Stopword removal**: Filter common English stopwords using NLTK stopwords list
5. **Lemmatization**: Apply WordNetLemmatizer (initial approach) or preserve original forms (final approach)
6. **Length filtering**: Remove words with length ≤ 2 characters
7. **Frequency filtering**: Ignore words appearing fewer than 5 times (min_count parameter)

1.3 Hyperparameter Settings

Hyperparameter	Value
Vector size	100 dimensions
Window size	5 (context words on each side)
Minimum count	5 (minimum word frequency)
Algorithm	CBOW (sg=0)
Training epochs	5
Negative sampling	5
Workers	8 (parallel processing threads)
Random seed	42 (reproducibility)

Table 1: Word2Vec hyperparameters used for training

These hyperparameters were chosen to match the dimensionality of the pre-trained GloVe model (100 dimensions) while providing sufficient context window (5 words) and training iterations (5 epochs) for learning meaningful representations.

2 Performance with Different Sampling Ratios (10%)

We trained Word2Vec models on 5%, 10%, and 20% samples of the Wikipedia corpus (5.6M articles total) to analyze the relationship between corpus size and model performance.

2.1 Corpus Statistics

Sampling Ratio	Articles Sampled	Vocabulary Size	Training Time
5%	280,790	414,982	∼20 min
10%	563,148	647,437	∼35 min
20%	1,124,733	992,355	∼60 min

Table 2: Corpus statistics for different sampling ratios

2.2 Performance Comparison

Sampling Ratio	Vocabulary	Semantic (%)	Syntactic (%)	Overall (%)
5%	414,982	62.08	37.69	49.89
10%	647,437	64.51	41.97	53.24
20%	992,355	66.24	43.23	54.74
Pre-trained GloVe	∼400,000	65.34	61.26	63.30

Table 3: Model performance across different sampling ratios (with lemmatization)

2.3 Analysis

Vocabulary Growth:

- 5% → 10%: +56% vocabulary increase (232K words)
- 10% → 20%: +53% vocabulary increase (345K words)
- Nearly linear relationship between sampling ratio and vocabulary size

Performance Gains:

- 5% → 10%: Semantic +2.42%, Syntactic +4.28%
- 10% → 20%: Semantic +1.73%, Syntactic +1.26%
- Clear diminishing returns: doubling data from 10% to 20% yields smaller improvements

Efficiency Recommendation: The 10% sampling ratio provides optimal cost-benefit balance, achieving 97% of the 20% model's semantic performance with only 65% of the vocabulary and half the training time. For practical applications where training resources are constrained, 10% sampling is recommended unless maximum accuracy is critical.

3 Performance on Different Corpora (15%)

3.1 Results Presentation

We compared three models: Wikipedia-trained (our custom model), AG News-trained (news corpus), and pre-trained GloVe (Wikipedia + Gigaword news).

Model	Articles	Vocabulary	Semantic (%)	Syntactic (%)
Wikipedia 20%	1,124,733	992,355	66.24	43.23
AG News	120,000	23,333	8.47	8.09
Pre-trained GloVe	∼6B tokens	∼400,000	65.34	61.26

Table 4: Performance comparison across different corpora

Sub-Category	Wikipedia (%)	AG News (%)	Difference
capital-common-countries	85.18	37.55	-47.63
capital-world	80.26	10.06	-70.20
family	54.74	1.98	-52.76
gram8-plural	0.00	0.00	0.00
gram6-nationality-adjective	86.05	32.15	-53.90

Table 5: Category-specific performance comparison

3.2 Corpus Characteristics

Wikipedia Corpus:

- **Size**: 5.6M articles (1.1M sampled at 20%)
- **Topics**: Encyclopedic, broad coverage (history, science, culture, geography, biography)
- **Structure**: Long-form articles with detailed explanations
- **Vocabulary**: Comprehensive, includes technical terms and proper nouns
- **Style**: Formal, neutral, well-edited prose

AG News Corpus:

- **Size**: 120,000 news articles
- **Topics**: Current events, politics, business, sports, technology
- **Structure**: Short articles, headlines, breaking news format
- **Vocabulary**: Repetitive, focused on recent events, temporal language
- **Style**: Journalistic, time-sensitive, conversational

Key Differences:

1. **Size disparity**: Wikipedia has 10× more articles, resulting in 40× larger vocabulary
2. **Topic breadth**: News focuses on current events; Wikipedia covers universal knowledge
3. **Temporal focus**: News emphasizes recent events; Wikipedia covers all time periods
4. **Vocabulary diversity**: News reuses common words; Wikipedia has specialized terminology

3.3 Performance Analysis

Why AG News Model Fails (8% accuracy):

1. **Vocabulary Coverage (37.8% OOV rate)**: 7,382 test words missing from vocabulary
    - Family relationships (1.98%): News rarely discusses extended family
    - Morphological forms (0-6%): Insufficient examples of systematic word transformations
    - Abstract relationships: News focuses on concrete events, not linguistic patterns
2. **Domain Specificity**: News corpus learned temporal, event-specific associations

- Example: "king" → "Norodom Sihamoni" (Cambodian king frequently in news)
- Rather than general concept: "king" → "queen", "monarch", "prince"

3. **Context Differences**: Same words have different embeddings based on usage
   - "doctor" in news: associated with medical emergencies (leukemia, treatment, hospital)
   - "doctor" in Wikipedia: associated with medical professions (physician, surgeon, dentist)

Why Capitals/Nationality Perform Better (32-37%):

News articles constantly mention countries, cities, and nationalities in geopolitical coverage. These categories represent the only semantic overlap between news vocabulary and the analogy test set.

Conclusion:

Corpus choice dramatically affects word embedding quality. Wikipedia's encyclopedic breadth creates better general-purpose embeddings, while news corpora produce domain-specific representations that perform poorly on general analogy tasks. For NLP applications requiring broad language understanding, larger and more diverse corpora are essential.

4 Word Similarity Observations (10%)

We analyzed word similarity across three models to understand how training corpus affects semantic relationships.

4.1 Comparative Analysis

Word	Pre-trained GloVe	Wikipedia 20%	AG News
king	prince (0.77)	prince (0.77)	norodom (0.82)
	queen (0.75)	monarch (0.74)	sihamoni (0.79)
	son (0.70)	queen (0.73)	sihanouk (0.75)
	brother (0.70)	crown (0.67)	penh (0.74)
	monarch (0.70)	throne (0.67)	phnom (0.74)
doctor	physician (0.77)	nurse (0.71)	leukemia (0.76)
	nurse (0.75)	dentist (0.68)	treatment (0.75)
	dr. (0.72)	psychiatrist (0.64)	patient (0.75)
	doctors (0.71)	surgeon (0.64)	procedure (0.74)
	patient (0.71)	zhivago (0.63)	hospital (0.69)
woman	girl (0.85)	men (0.71)	men (0.80)
	man (0.83)	lesbian (0.63)	young (0.68)
	mother (0.83)	gender (0.63)	girl (0.67)
	boy (0.77)	female (0.61)	mother (0.64)
	she (0.76)	individual (0.61)	teenager (0.64)

Table 6: Top-5 most similar words across different models

4.2 Key Observations

1. "King" - Corpus Bias Detection:

- **GloVe/Wikipedia**: General royalty concepts (prince, queen, monarch, throne)
- **AG News**: Cambodian royalty (Norodom Sihamoni, King Sihanouk, Phnom Penh)
- **Insight**: News corpus reflects temporal events rather than general concepts, demonstrating how smaller specialized corpora create overfitted representations

2. "Doctor" - Contextual Differences:

- **GloVe**: Professional roles (physician, nurse, dr.)
- **Wikipedia**: Medical specialties (dentist, psychiatrist, surgeon) + cultural reference (Zhivago)
- **AG News**: Medical events/conditions (leukemia, treatment, procedure, hospital)
- **Insight**: Same word has different neighbors based on usage context; news emphasizes medical emergencies, Wikipedia covers professions

3. "Woman" - Semantic Quality:

- **GloVe**: Strong gender relationships (girl, man, mother) with high similarity scores (0.83-0.85)

- **Wikipedia**: Gender/social terms (men, lesbian, gender, female)
- **AG News**: Demographic descriptions (men, young, girl, teenager)
- **Insight**: GloVe captures the strongest analogical relationships; Wikipedia reflects academic/social discussions; news uses demographic framing

4. Vocabulary Richness:

- Wikipedia's 992K vocabulary captures nuanced relationships
- News's 23K vocabulary produces weaker, more generic associations
- Larger vocabularies enable more precise semantic distinctions

4.3 Conclusion

Word embeddings strongly reflect their training corpus characteristics. Wikipedia produces general-purpose embeddings with abstract relationships, while news produces domain-specific, temporally-biased embeddings. The AG News model's association of "king" with Cambodian royalty (rather than general monarchy concepts) demonstrates how smaller, specialized corpora create overfitted representations unsuitable for general NLP tasks.

5 Additional Insights (5%)

5.1 Capitalization Issue Discovery

During initial evaluation with the pre-trained GloVe model, we encountered a critical issue where 51% of predictions failed (9,962 out of 19,544 questions). Investigation revealed a vocabulary mismatch:

Root Cause:

- Google Analogy dataset uses proper capitalization: `Athens Greece Baghdad Iraq`
- GloVe vocabulary contains only lowercase words
- Testing confirmed: `'Athens' in model` → `False`, but `'athens' in model` → `True`

Impact:

Category	Before Fix	After Fix
capital-common-countries	0.00%	93.87%
capital-world	0.00%	88.95%
gram6-nationality-adjective	0.00%	87.87%
Failed predictions	9,962	0

Table 7: Performance before and after capitalization fix

Solution: Modified prediction code to lowercase all input words before vocabulary lookup:

```
word_a_lower = word_a.lower()
# Use lowercase versions for similarity search
```

Lesson Learned: This issue highlights the critical importance of preprocessing consistency between training and evaluation. The 51% failure rate could have been misinterpreted as poor model quality when it was actually a data format mismatch. This emphasizes that data preparation is as critical as model architecture in NLP tasks.

5.2 Lemmatization Trade-off Analysis

Our initial preprocessing included lemmatization, which converted all word forms to their base forms (e.g., "cars" $\rightarrow$ "car"). This caused complete failure on morphological analogy tasks.

Experimental Setup: We retrained the 20% Wikipedia model without lemmatization to test the hypothesis that lemmatization destroys morphological patterns.

Preprocessing	**Vocabulary**	**Semantic (%)**	**Syntactic (%)**	**Plural (%)**
With Lemmatization	992,355	66.24	43.23	0.00
Without Lemmatization	1,016,479	68.68	53.49	65.39
Improvement	+24,124	+2.44	+10.26	+65.39

Table 8: Impact of lemmatization on model performance

Analysis:
Lemmatization creates a fundamental trade-off:
Advantages:

- Reduces vocabulary size (consolidates word forms)
- Helps with data sparsity for rare words
- Useful for information retrieval where morphological variants should match

Disadvantages (demonstrated):

- **Destroys morphological information**: Cannot distinguish grammatical forms
- **Loses semantic nuance**: "brothers" (sibling relationship) vs. "brother" (single sibling)
- **Breaks analogy tasks**: Morphological analogies require preserving word forms

Example of Lemmatization Failure:

- Original analogy: `car:cars::truck:trucks`
- After lemmatization: `car:car::truck:truck` (all become singular!)
- Result: Model cannot learn the singular → plural relationship

Conclusion: The 10.26% syntactic improvement and complete recovery of plural detection (0% → 65.39%) demonstrate that preserving original word forms produces superior embeddings for analogy tasks. Lemmatization should be avoided for word embedding training intended for general NLP applications. It is appropriate only for specific use cases like search engines where morphological distinctions are unimportant.

5.3 t-SNE Visualization Analysis

We performed quantitative analysis of t-SNE visualizations for family relationship words across pre-trained GloVe and custom models.

Metric	GloVe	Custom	Interpretation
Words captured	46	42	Custom missing 4 words
Silhouette score	-0.0711	0.0547	Custom has better clustering
Spread (X-axis std)	1.12	1.06	Similar horizontal dispersion
Spread (Y-axis std)	1.11	0.78	Custom more compressed
Avg pairwise distance	2.04	1.68	Custom clusters tighter (18%)
Male-female separation	0.28	0.43	Custom has 55% stronger separation

Table 9: Quantitative comparison of t-SNE clustering patterns

Key Findings:

1. **Clustering Quality**: The silhouette score reveals that custom embeddings (0.0547) have better-defined gender clusters than GloVe (-0.0711). Positive scores indicate clearer separation between male and female family terms.
2. **Spatial Organization**: Custom model produces tighter, more organized clusters with 18% smaller average pairwise distances (1.68 vs. 2.04), suggesting more consistent semantic relationships within the family domain.
3. **Gender Separation**: The most striking difference is male-female center distance: custom (0.43) vs. GloVe (0.28), representing 55% stronger gender-based clustering. This indicates the custom model learned clearer gender distinctions from Wikipedia's consistent usage patterns in biographical articles.
4. **Vocabulary Coverage**: Custom model captures 42/46 family words (91%). Missing words likely include low-frequency pronouns filtered by min_count=5.

Conclusion:

Despite training on less diverse data (Wikipedia only vs. Wikipedia + Gigaword), the custom model produced more semantically organized family embeddings. This demonstrates that domain-appropriate training data matters more than sheer corpus size for specific semantic categories. Wikipedia's encyclopedic coverage of family relationships provided consistent, high-quality examples for learning these distinctions.

6 Summary and Recommendations

Based on our experiments, we provide the following recommendations:

1. **Preprocessing**: Avoid lemmatization for general-purpose embeddings to preserve morphological patterns (+10% syntactic accuracy, +65% plural accuracy)

2. **Corpus Selection**: Use large, diverse corpora like Wikipedia for general NLP tasks. Domain-specific corpora (news) are suitable only for specialized applications

3. **Sampling Strategy**: 10% sampling provides optimal cost-benefit ratio, achieving 97% of maximum semantic performance with half the training time

4. **Data Validation**: Always verify preprocessing consistency between training and evaluation data to avoid subtle bugs like capitalization mismatches

5. **Model Selection**: For analogy tasks, custom-trained models can match or exceed pre-trained models when trained on appropriate corpora with proper preprocessing