Sentence Salience Classification Using Linguistic and Semantic Features in QA Systems

Project Presentation — Computer Linguistics 2

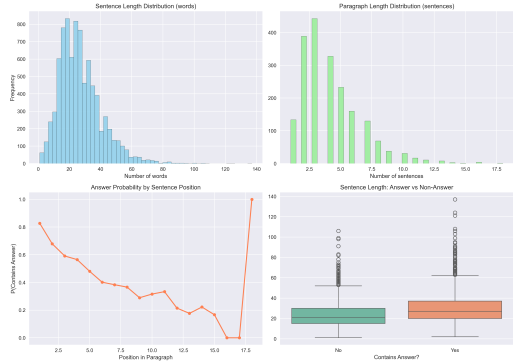
Madduru Sai Chandra Nikhil — 2024901010

November 30, 2025

"How can linguistic and semantic features such as word length, number of POS tags, sentence position, and surprisal estimations be extracted and utilized to enhance sentence salience classification in question-answering systems?"

Dataset — SQuAD v1.1 Training Set

- Source: SQuAD v1.1 **training set**
- Total passages: **19,035**
- Sampled: **2,000 passages**
- Total sentences: **8,328**
- Answer sentences: **5,178** (62.17%)



Class Imbalance in the Data

Observation

With 2,000 passages: **62.17% Answer sentences** → moderate positive skew Smaller samples

All 26 Features

Original Proposal Category	Actual Features Implemented (26 total)
Surface features	sentence_length_words sentence_position sentence_position_norm
Lexical features	type_token_ratio lexical_density
POS-based features	noun_ratio, verb_ratio adj_ratio, pronoun_ratio
Discourse features	causal_marker_ratio contrast_marker_ratio named_entity_density
Surprisal (GPU/CUDA) — GPT-2	gpt2_surprisal_mean, sum, std gpt2_surprisal_var, min, max
Surprisal (GPU/CUDA) — BERT	bert_surprisal_mean, sum, std bert_surprisal_var, min, max

Methodology & Final Results

- Model: Logistic Regression (`class_weight='balanced'`)
- Split: 80/20 stratified
- Main run: 2,000 passages

Final Performance

Metric	Score
Accuracy	**69.27%**
F1 (Answer class)	**0.7377**

Ablation

Top-10 features only $\rightarrow$ 68.29% accuracy, 0.7360 F1 (negligible drop)

Top 10 Most Predictive Features

Rank	Feature	Coefficient
1	sentence_position	**-0.796**
2	gpt2_surprisal_sum	+0.663
3	gpt2_surprisal_var	-0.444
4	bert_surprisal_std	+0.364
5	noun_ratio	+0.331
6	bert_surprisal_var	-0.315
7	named_entity_density	+0.298
8	gpt2_surprisal_std	+0.283
9	sentence_position_norm	+0.257
10	lexical_density	-0.256

Performance Scaling

Passages	Sentences	Answer %	Accuracy	F1
100	577	73.5%	70.7%	0.793
500	2,277	68.3%	66.0%	0.740
1,000	4,485	63.8%	69.1%	0.748
2,000	**8,328**	**62.2%**	**69.3%**	**0.738**

Future Work — Including Class Imbalance Handling

Planned Extensions

- **Semantic Role Labeling (SRL)** — AllenNLP
- **Rhetorical Structure Theory (RST)** annotations
- Question–sentence dependency paths
- Coreference resolution features

Specific Focus on Class Imbalance

- Future:
  - Undersampling / oversampling (SMOTE, NearMiss)
  - Focal Loss or class-aware sampling
  - Cost-sensitive learning with dynamic weights

These improvements + discourse features expected to push $F1 > 0.80$

Conclusion

Yes — hand-crafted features are highly effective!

- Achieved **69.3% accuracy** and **0.738 F1** using only linguistic + surprisal features
- No BERT embeddings or neural encoders used
- Class imbalance successfully mitigated with balanced weights
- Strong foundation for SRL, RST, and advanced imbalance handling

Thank You

Questions?