Sentence Generation using Fan Theories

TANVI SAHAY

University of Massachusetts Amherst

SPOILER ALERT

Goals

- Application of natural language processing techniques for sentence generation
- Performance assessment based on Domain Knowledge

Database

Fan theories!

Based on the previous events, what do fans of the show predict will happen next

- Theories are non - repetitive
- Data base has high noise
- Jargon!
- Need of Domain Knowledge

Cleganebowl!

R + L = J

Tools Analyzed

Stanford Log-linear Part-Of-Speech Tagger
Stanford Named Entity Recognizer (NER)
Stanford Deterministic Coreference Resolution System
Stanford Open Information Extrcation (OpenIE)

Models

Tokenized text + bigram LM

['jon','snow','may','fulfil','the','azor','ahai','prophecy']

OpenIE relation tuples + bigram LM

1.0	Jon Snow	be	second
0.97	Jon Snow	coming of	Azor Ahai

OpenIE Named Entity Information + bigram LM

Arya	's friend is	Gendry	🟢
His	was	temporarily freed	🔴

OpenIE relation tuples + HMM

Character-level LSTM

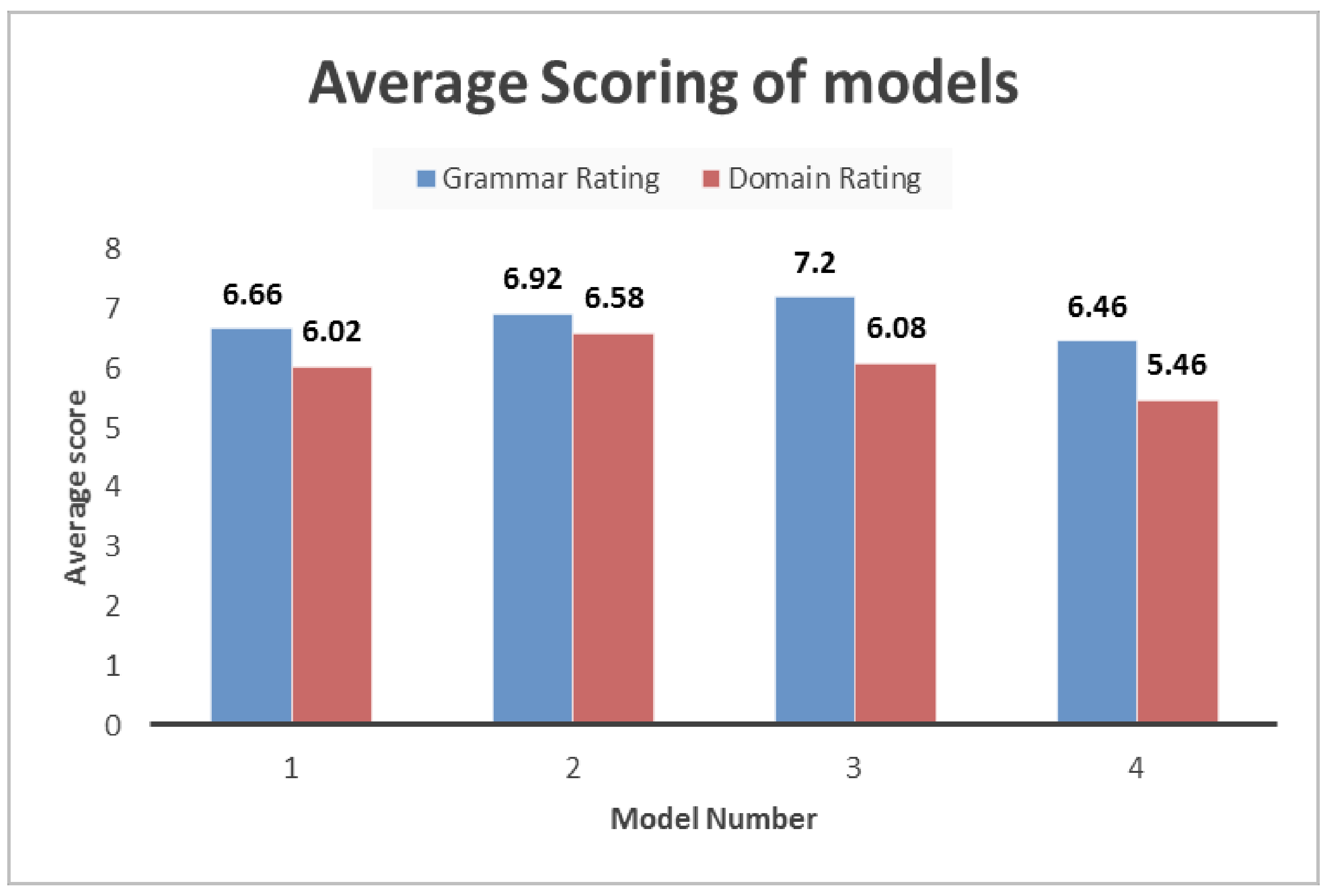
Results

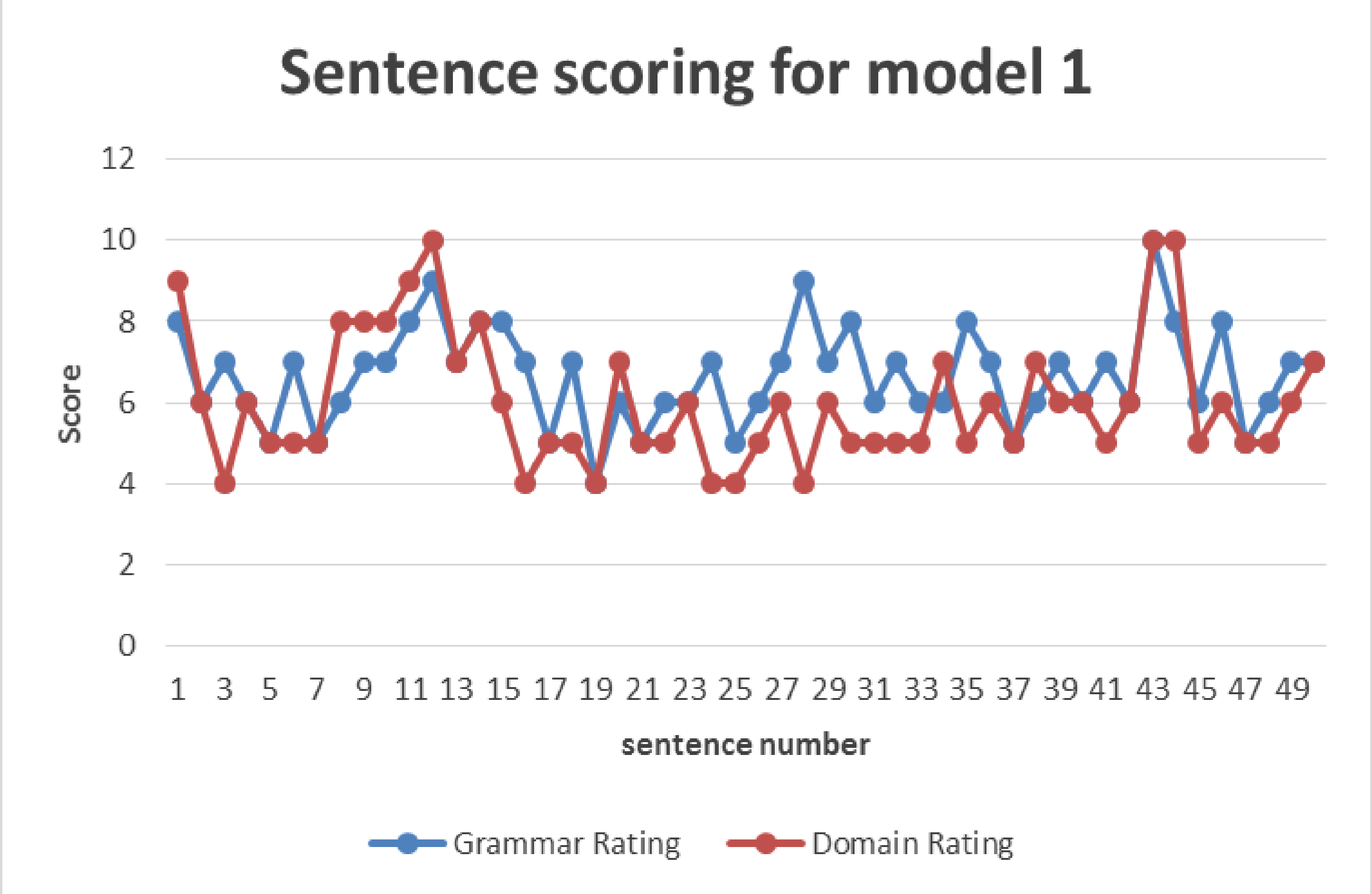
10 sentences generated for most famous characters and rated based on both grammatical correctness and domain relevance

Top 5 most talked about characters (NER):
Jon Arya Cersei Dany Ned

Example Sentences:

- **arya get married .**
- **arya 's joins up with band .**
- **arya arrived back trying .**
- **arya winterfell**
- **(arya stark's kill list. in season 3) arya (the the the**





Conclusions

- Model 3 performs best, owing to the training data containing minimum noise and only relevant sentences.
- The Stanford POS Tagger, NER and Coreference resolution systems perform well despite the noisy data.
- In general, most models are able to map grammatical correctness better than domain relevance.
- Noise removal improves the quality of sentences being generated but to a limited extent only.
- Despite being powerful generative models character LSTMs fail on small and noisy databases.

Scope of Improvement

Further noise removal using the resolved co-references

Word-level LSTM