Evaluating N-gram Language Models with SRILM

Han Wu
NLP / CISC7021
FST / University of Macau
mc35374@umac.mo

1 Introduction

Natural Language Processing (NLP) is a significant domain in computer science, focusing on computer comprehension and manipulation of human language. Language models are essential in NLP, serving critical roles in various text processing tasks.

In this assignment, we explore language model construction and performance assessment using the SRILM toolkit[1], known for its language modeling capabilities. We create word-based models (1-gram to 3-gram) and character-based models (1-gram to 6-gram) and investigate how additional data can influence these models.

The report's structure: In Section 2, we delve into word-level n-gram language model construction and performance evaluation. Section 3 focuses on the creation and evaluation of character-level n-gram language models. Section 4 covers data expansion, and evaluation of language models. Moving to Section 5, we provide a summary of results and an in-depth analysis. Finally, Section 6 offers our concluding remarks.

2 Word-level N-grams

This section delves into the creation and evaluation of word-level n-gram language models, specifically 1-gram, 2-gram, and 3-gram models. Our objective is to discern the influence of different n-gram orders on language modeling while optimizing model perplexity.

2.1 Model Construction

We initially fashioned n-gram language models using the provided training data, comprising 300,000 lines of text. These models served as our initial benchmarks, affording us insights into their performance on both the training and testing sets. However, the initial perplexity values left ample room for refinement.

2.2 Data Preprocessing

To elevate the quality of our language models, we embarked on a sequence of data preprocessing maneuvers, refining the text to enhance its suitability for language modeling:

Text Purification: We transformed all text to lowercase to ensure insensitivity to case. Furthermore, the extraction of punctuation marks honed our focus on word sequences.

Stopwords[2] **Elimination**: To diminish noise and computational complexity, we excluded commonplace stopwords that wield minimal influence on language modeling.

2.3 Model Enhancement

Armed with the refined data, we re-engineered our n-gram language models, incorporating advanced methodologies to augment their efficacy:

Smoothing: Kneser-Ney smoothing, a well-regarded technique in language modeling, was applied to alleviate the issue of data sparsity. This enabled the models to ascribe non-zero probabilities to previously unseen n-grams.

Interpolation: The judicious use of interpolation facilitated the amalgamation of probabilities derived from lower-order and higher-order n-grams, striking an equilibrium between local and global contextual information.

Pruning: To optimize model efficiency, we employed pruning techniques to excise less pertinent n-grams, effectively reducing model dimensions while preserving predictive prowess.

2.4 Model Evaluation and Outcomes

Following the refinement of our n-gram models, we subjected them to renewed evaluation on both

[1] http://www.speech.sri.com/projects/srilm/

[2] https://www.kaggle.com/datasets/rowhitswami/stopwords

the training and testing datasets. The ensuing perplexity values exhibited remarkable amelioration compared to our initial models.

n-gram	train	test
1-gram	2372.35	1960.92
2-gram	150.67	379.16
3-gram	106.73	320.65

Table 1: N-gram Perplexity

n-gram	train	test
1-gram	471.87	450.52
2-gram	175.10	227.95
3-gram	82.12	195.17

Table 2: Perplexity Improvement

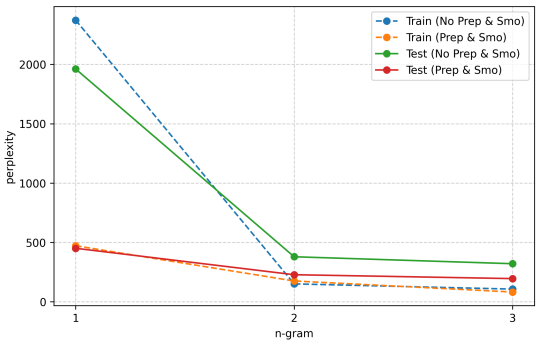

Figure 1: Word - Level Perplexity

These results underscore the pivotal role of meticulous data preprocessing and the adroit application of advanced modeling techniques in elevating the caliber of language models.

3 Character-level N-grams

In this section, our focus transitions to character-level n-gram language models, encompassing models ranging from 1-gram to 6-gram.

3.1 Data Preprocessing (Character-level)

Drawing inspiration from the effective data preprocessing methods employed in the realm of word-based models discussed in Section 2, we extended similar preprocessing steps to our character-level training data. This encompassed:

Text Refinement: We uniformly converted all characters to lowercase, thereby ensuring case insensitivity. Additionally, the removal of punctuation marks served to enhance the modeling process.

Stopword Elimination: Even at the character level, we applied the strategy of stopwords elimination to attenuate noise in the data.

3.2 Model Conception and Appraisal

To construct character-level n-gram models, we embarked upon the following endeavors:

Data Segmentation: Both the preprocessed training and validation datasets underwent segmentation into individual characters rather than words.

Application of Smoothing: Drawing from the success of the Kneser-Ney smoothing technique in our word-based models, we employed it to counteract data sparsity concerns in the character-level models.

Model Application: Subsequently, the trained character-level models were employed on the segmented training and testing datasets, enabling the computation of perplexity values for each n-gram order, ranging from 1-gram to 6-gram.

3.3 Model Evaluation and Findings

The evaluation of character-level n-gram models unveiled compelling results:

	train	test
1-gram	93.81	93.68
2-gram	27.28	27.23
3-gram	21.61	21.61
4-gram	16.20	16.21
5-gram	12.46	12.81
6-gram	5.15	5.26

Table 3: Perplexity of Character-level

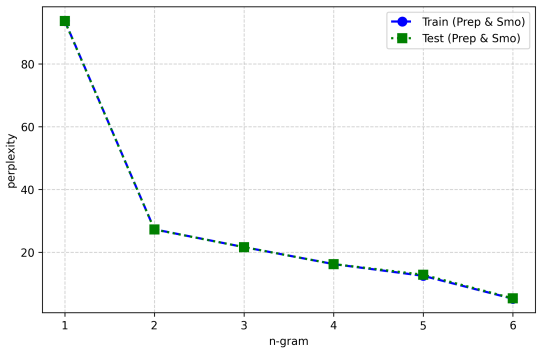

Figure 2: Character - Level Perplexity

The outcomes of this section illuminate the potential of character-level modeling in the realm of NLP tasks, corroborating their performance across diverse n-gram orders.

4 Data Expansion

In this section, we explore the influence of data expansion on language models and perplexity by incorporating the European Parliament Proceedings Parallel Corpus into our training data[3].

4.1 Data Enrichment

We enhance our training dataset with the European Parliament Proceedings Parallel Corpus. While our original training data focused on news commentary, this added corpus introduces a distinct domain - European Parliament proceedings. This domain shift may present linguistic challenges due to differing terminology and patterns.

4.2 Model Creation and Assessment

New language models are constructed using the expanded training dataset. As in previous sections, we create word-based and character-based n-gram models and assess their perplexity on both the training and validation datasets.

4.3 Model Evaluation and Outcomes

Evaluation of models trained on the expanded dataset unveils surprising results:

n-gram	train	test
1-gram	674.56	634.34
2-gram	264.58	303.63
3-gram	132.54	252.48

Table 4: Combined-Word-based Perplexity

n-gram	train	test
1-gram	122.08	120.67
2-gram	41.03	41.14
3-gram	26.98	27.18
4-gram	16.95	17.18
5-gram	10.87	11.15
6-gram	3.92	4.13

Table 5: Combined-Char-based Perplexity

[3]https://www.statmt.org/europarl/

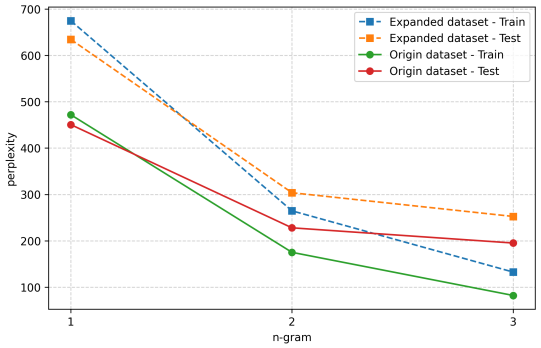

Figure 3: Expanded Dataset Perplexity-Word

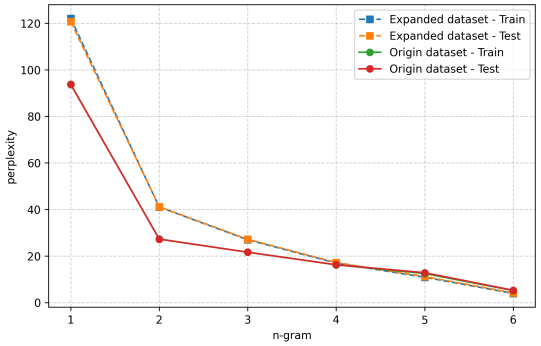

Figure 4: Expanded Dataset Perplexity-Character

Our findings suggest that incorporating data from a distinct domain, European Parliament proceedings, results in increased perplexity. This elevation indicates a notable impact of domain shift between the training and validation data on model performance.

4.4 Domain Adaptation Considerations

The observed perplexity increase prompts consideration of domain adaptation strategies when working with multilingual or multi-domain data. Future efforts may involve techniques such as fine-tuning or domain adaptation to effectively adapt models to the target domain.

5 Results, Analysis, and Conclusion

In this section, we provide a summary of our findings and offer a comprehensive analysis of the key factors that influenced the performance of our language models.

5.1 Summary of Results and Analysis

We commenced this assignment by constructing and evaluating both word-based and character-based n-gram language models, spanning different n-gram orders. Notably, the incorporation of data

preprocessing techniques, Kneser-Ney smoothing, interpolation, and pruning led to remarkable improvements in perplexity values.

Exploring character-level modeling, we created models ranging from 1-gram to 6-gram, uncovering their distinctive characteristics. Throughout these experiments, data preprocessing remained pivotal in enhancing model performance.

The introduction of additional training data from the European Parliament Proceedings Parallel Corpus presented a unique challenge. It was observed that this domain shift resulted in increased perplexity values, underscoring the significance of domain adaptation in language modeling.

Impact of N-gram Orders: Our analysis unveiled that higher-order n-gram models excel in capturing intricate language patterns but often grapple with data sparsity issues. The application of interpolation techniques successfully struck a balance between local and global contextual information.

Data Preprocessing: Across all modeling endeavors, data preprocessing consistently played a pivotal role in enhancing model performance by elevating data quality and mitigating noise.

Character vs. Word: Character-level models showcased distinct capabilities, particularly in effectively handling out-of-vocabulary words and accommodating morphologically rich languages.

Domain Shift: The unexpected increase in perplexity following the introduction of data from a different domain underscored the imperative of domain-specific modeling and the intricate challenges posed by domain shifts.

5.2 Conclusion

In conclusion, this assignment allowed us to explore various aspects of language modeling, from word-based to character-based models, and the effects of data preprocessing. We also encountered the challenge of domain shift, emphasizing the need for domain-specific modeling approaches.

As we move forward, it is essential to consider domain adaptation strategies when working with diverse text data sources. Further research and experimentation in this area can lead to more robust and domain-aware language models.

References

Andreas Stolcke. 2002. *SRILM - an extensible language modeling toolkit*. In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*.

Philipp Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*. In *MT Summit 2005*.

Stanley F. Chen and Joshua Goodman. 1999. *An empirical study of smoothing techniques for language modeling*. In *Computer Speech & Language*, volume 13, number 4, pages 359–394, Elsevier.