Objective

- Train a classical n-gram language model (using KenLM)
- Score each candidate sentence (using perplexity score)
- Select the best candidate
- Analyze performance and failure cases

Dataset

- ASR Hypotheses**: 500 Ukrainian utterances with 8 candidate transcriptions each
- Training Corpora**: Three Ukrainian text corpora from UberText 2.0
    - Social: 4.5M sentences (87 MB)
    - Fiction: 7.2M sentences (398 MB)
    - News: 51.3M sentences (3.4 GB)

Methodology

1. Language Model Training

Trained n-gram models (1-gram, 2-gram, 3-gram) using KenLM on each corpus.

This guide has been an immense help in completing this task: https://apxml.com/courses/applied-speech-recognition/chapter-5-language-modeling-decoding/building-n-gram-model-kenlm

- Preprocessing**: Lowercasing, whitespace normalization
- Model Format**: Binary

2. Hypothesis Reranking

For each row:

1. Score all candidate hypotheses using the model
2. Calculate perplexity: PPL = 10^(-log_prob / word_count)
3. Select candidate with lowest perplexity

3. Evaluation

Measured exact-match accuracy: percentage of rows where the top-ranked candidate matches the reference transcription

Results

Corpus	N-gram	Accuracy	Correct
Social	1-gram	60.80%	304/500
Social	2-gram	73.60%	368/500
Social	3-gram	76.00%	380/500
Fiction	1-gram	55.40%	277/500

Corpus	N-gram	Accuracy	Correct
Fiction	2-gram	61.20%	306/500
Fiction	3-gram	63.40%	317/500
News	1-gram	60.00%	300/500
News	2-gram	79.60%	398/500
News	3-gram	80.60%	403/500

Best result: not surprising: 3-gram model trained on news corpus achieved 80.60% accuracy (403/500 correct)

Analysis

Higher-order n-grams outperform lower-order models across all corpora:

- 1-gram models score words independently, ignoring context
- 2-gram models capture local word pairs, providing significant improvement (+13-20% absolute)
- 3-gram models add minimal gains over 2-grams (+2-3% absolute)

The diminishing returns from 2-gram to 3-gram suggest that local bigram context captures most relevant information for this task

Training Corpus Size

Larger corpora generally improve performance, but domain match matters more than size:

- Fiction (398 MB) performed worst despite moderate size
- Social (87 MB) outperformed Fiction despite being smaller
- News (3.4 GB) achieved best results

Domain Matching

The ASR test data appears to be news/sports headlines (e.g., "Вайлдер — Ортіс: відео нокауту"). The News corpus achieved highest accuracy, demonstrating that domain-matched training data is more valuable than corpus size alone.

Conclusions

1. N-gram language models effectively improve ASR accuracy: Best model achieved 80.6% exact-match accuracy, a substantial improvement over random baseline (12.5% for 8 candidates)
2. Domain matching is critical: News corpus (domain-matched) outperformed larger Fiction corpus by 17 percentage points
3. Bigrams provide most value: 2-gram models offer 13-20% improvement over unigrams, while 3-grams add only 2-3%