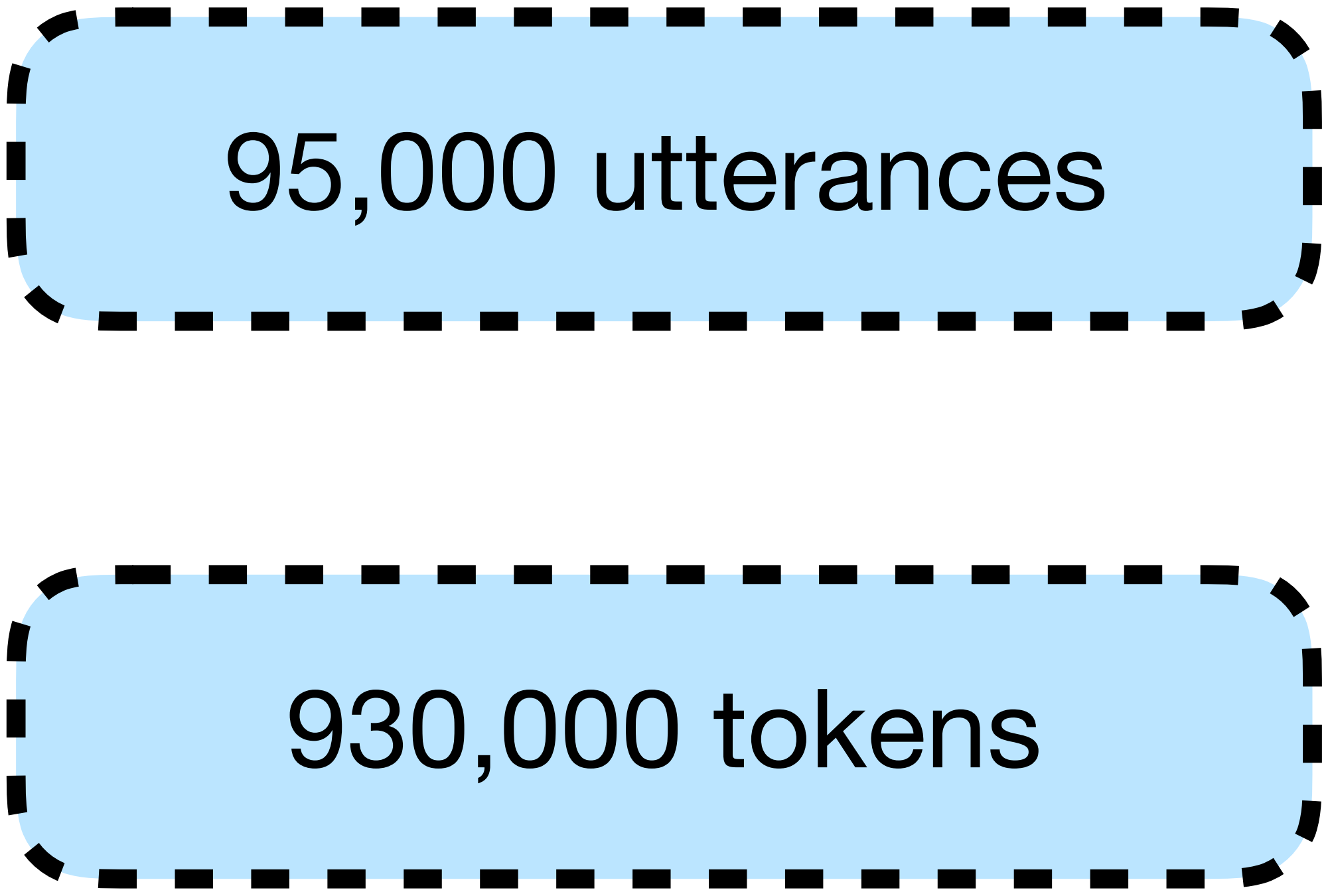
Experiments and Results

13

SEAME corpus, a conversational Mandarin-English code-switched speech corpus

Data Description

Cleaning and removing non-clean utterances


95,000 utterances
930,000 tokens

Experiments and Results

Data Description

SEAME corpus, a conversational Mandarin-English code-switched speech corpus

Cleaning and removing non-clean utterances

95,000 utterances

930,000 tokens

Perplexity Experiments

Smoothing Technique	Dev		Test	
	mixed LM	DLM	mixed LM	DLM
Good Turing	338.2978	**329.1822**	384.5164	**371.1112**
Kneser-Ney	329.6725	**324.9268**	376.0968	**369.9355**

Mixed LM : Baseline n-gram LM trained directly on entire data

Both the models are bigram LMs