Tokenization, Language Modeling, and Smoothing.

In systems, I/O is limited. Domain –> I/O characteristics

Example of analogue system enforcing constraints: Telegraph

So do we enforce constraints in Natural Language, or take text from the wild and process it to fit the constraints? Same input and algo can perform differently if preprocessed differently.

Tokenization

Task of separating 'tokens' in a given sentence.

token: a single surface form word

Type: is a vocabulary word.

Core principle: Do not lose information at this phase. Input will be frozen. *Suggestion*: break everything, reconstruction is fine.

At this point in the lecture, I was told to shut down my laptop and thus have no recollection of what happened after.

Henceforth, the notes end here.