Foundations of Natural Language Processing
Lecture 5a
Probabilities of Word Sequences

Alex Lascarides

School of informatics

Recap

- Last time, we talked about corpus data and some of the information we can get from it, like word frequencies.
- For some tasks, like sentiment analysis, word frequencies alone can work pretty well (though can certainly be improved on).
- For other tasks, we need more.
- Today: we consider **sentence probabilities**: what are they, why are they useful, and how might we compute them?

Intuitive interpretation

- "Probability of a sentence" = how likely is it to occur in natural language
  - Consider only a specific language (English)
  - Not including meta-language (e.g. linguistic discussion)

$$P(\text{the cat slept peacefully}) > P(\text{slept the peacefully cat})$$

$$P(\text{she studies morphosyntax}) > P(\text{she studies more faux syntax})$$

Language models in NLP

- It's very difficult to know the true probability of an arbitrary sequence of words.
- But we can define a **language model** that will give us good approximations.
- Like all models, language models will be good at capturing some things and less good for others.
  - We might want different models for different tasks.
  - Today, one type of language model: an **N-gram model**.

N-gram Language Model

A probability distribution over word sequences

$$w_1 \dots w_n$$

of length n

Spelling correction

Sentence probabilities help decide correct spelling.

mis-spelled text		no much effert
	↓ (Error model)	
possible outputs		no much effect so much effort no much effort not much effort ...
	↓ (Language model)	
best-guess output		not much effort

Automatic speech recognition

Sentence probabilities help decide between similar-sounding options.

speech input

↓ (Acoustic model)

possible outputs

She studies morphosyntax
She studies more faux syntax
She's studies morph or syntax
...

↓ (Language model)

best-guess output

She studies morphosyntax

Machine translation

Sentence probabilities help decide word choice and word order.

non-English input

↓ (Translation model)

possible outputs

She is going home
She is going house
She is travelling to home
To home she is going
...

↓ (Language model)

best-guess output

She is going home

LMs for prediction

- LMs can be used for **prediction** as well as correction.

- Ex: predictive text correction/completion on your mobile phone.
  - Keyboard is tiny, easy to touch a spot slightly off from the letter you meant.
  - Want to correct such errors as you go, and also provide possible completions. Predict as you are typing: ineff...

- In this case, LM may be defined over sequences of *characters* instead of (or in addition to) sequences of words.

But how to estimate these probabilities?

- We want to know the probability of word sequence $\vec{w} = w_1 \ldots w_n$ occurring in English.
- Assume we have some **training data**: large corpus of general English text.
- We can use this data to **estimate** the probability of $\vec{w}$ (even if we never see it in the corpus!)

Summary

- The probability of (arbitrary) word sequences are important for many NLP applications.

- Corpus data must inform those probabilities.

Next time: How to acquire LMs from corpus data.