ECE365: Introduction to NLP

Spring 2021

Lecture 2

[Reading J&M 2.2, 2.3, 2.4]

In this module

- Models using word-level information
- Models for sentence-level information
- Models of meaning

Levels of Language

Phonetics, Phonology	All sounds
Morphology	Word formation
Syntax	Clauses and sentences
Semantics	Meaning
Pragmatics	Language use

Words as Text Units

- Suma's cat is different from other cats.

Word form
Lemma
Word type
Word token

Words as Text Units

- How many words are there in English?

- The Second Edition of the 20-volume *Oxford English Dictionary*, published in 1989, … at the very least, a quarter of a million distinct English words, excluding inflections, and words from technical and regional vocabulary not covered by the *OED*, or words not yet added to the published dictionary, of which perhaps 20 per cent are no longer in current use. If distinct senses were counted, the total would probably approach three quarters of a million.

Words as Text Units

- Function words
- Content words

Recently, Illinois ECE alumnus Arvind Krishna (MSEE '87, PhD '91) was announced as the new CEO of IBM. Krishna previously served as the vice president for database services and has been with the company since 1990.

Words as Text Units

- Function words – express grammatical relationships
- Content words

Recently, Illinois ECE alumnus Arvind Krishna (MSEE '87, PhD '91) was announced as the new CEO of IBM. Krishna previously served as the vice president for database services and has been with the company since 1990.

Words as Text Units

- Function words – express grammatical relationships

- Content words -- words that name objects and their qualities

  Recently, Illinois ECE alumnus Arvind Krishna (MSEE '87, PhD '91) was announced as the new CEO of IBM. Krishna previously served as the vice president for database services and has been with the company since 1990.

Words as Text Units

- Stop words

  A list of most common words in the language.

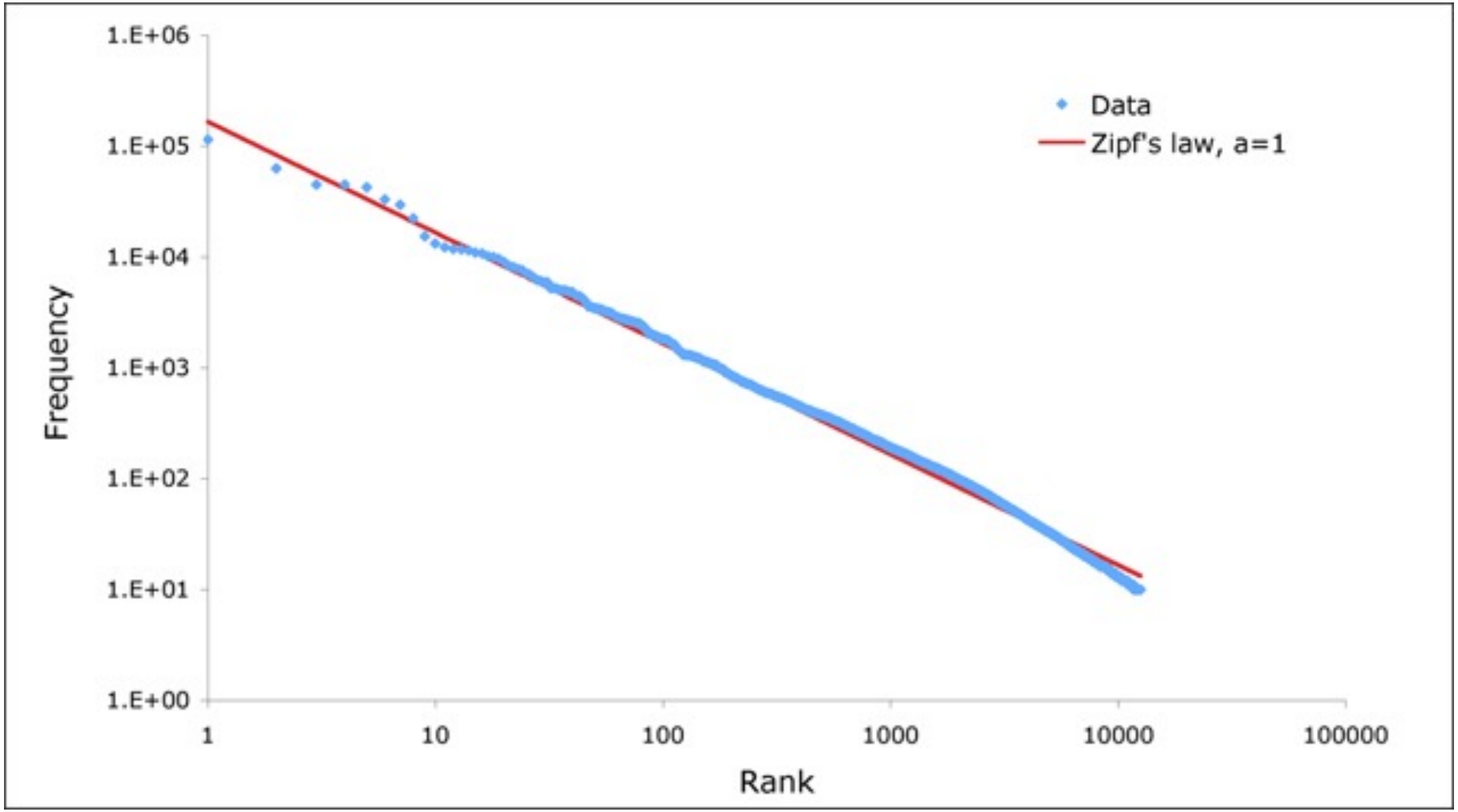
Words are Sparse

Sparse data due to Zipf's Law
frequency inversely proportional to rank

- To illustrate, let's look at the frequencies of different words in a large text corpus

- Assume "word" is a string of letters separated by spaces

Zipf's Law


1.E+06
1.E+05
1.E+04
1.E+03
1.E+02
1.E+01
1.E+00
Frequency
1
10
100
1000
10000
100000
Rank
Data
Zipf's law, a=1

Words are Sparse

Most frequent words in the English Europarl corpus (out of 24m word tokens)

any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

Words are Sparse

But also, out of 93,638 distinct words (word types), 36,231

Examples:

- cornflakes, mathematicians, fuzziness, jumbling
- pseudo-rapporteur, lobby-ridden, perfunctorily,
- Lycketoft, UNCITRAL, H-0695

Words are Sparse

- Regardless of how large our corpus is, there will be a lot of

- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen

Tokenization

- Tokenization is splitting running text into units
  - Friends, Romans, Countrymen, lend me your ears;

Tokenization in Other Languages

- Chinese
  - Word segmentation algorithm

- 莎拉波娃现在居住在美国东南部的佛罗里达。
- 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
- Sharapova now lives in US southeastern Florida

Tokenization Method

- Byte-Pair Encoding
  - Subword level
  - Better handle unseen words
  - Begins with all characters and sequentially merges characters

low

few

lower

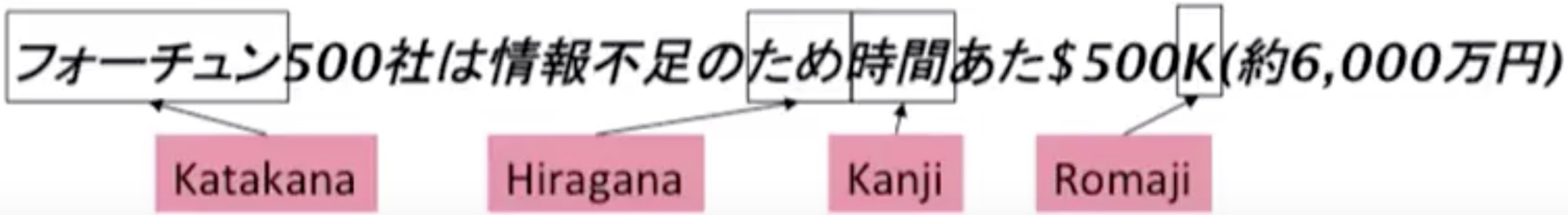
lowest

mow – {m, ow}

{m, l, o, w, e, r, s, t, f, ow}

Tokenization in Other Languages

- Japanese



Normalization

- Define equivalence classes of terms =

putting words/tokens in a standard format, choosing a single normal form for words with multiple forms

  - U.S.A = USA

  - Index and query terms need to have same form

Case Folding

- Application dependent
  - IR - Reduce to lower case, except when middle of sentence
  - Sentiment analysis, machine translation – retain case
    - US vs us

What about *organize*, *organizes*, and *organizing*?

Lemmatization

- Find the dictionary headword form
  - Reduce inflections to the base form
  - Am, are, is → be
  - Car, cars, car's, cars' → car

  the girl's cars are of different colors →

Stemming

- Reduce terms to their stems

  - word =
  stem (core meaning bearing unit) + affix (attached to stem for grammatical function)

  - Language dependent
  - chopping affix
    - Automate, automatic, automation → automat

    - Porter's Stemming algorithm for English

Summary