Introduction to Statistical NLP

(Diana Inkpen, 2021-2004

Based on Manning&Schutze Ch1)

Rational versus Empiricist Approaches to Language (I)

- ***Question:*** What prior knowledge should be built into our models of NLP?
- ***Rationalist Answer:*** A significant part of the knowledge in the human mind is not derived by the senses but is ***fixed*** in advance, presumably by genetic inheritance (Chomsky: ***poverty of the stimulus***).
- ***Empiricist Answer:*** The brain is able to perform ***association***, ***pattern recognition***, and ***generalization*** and, thus, the structures of Natural Language can be ***learned***.

Rational versus Empiricist Approaches to Language (II)

- ***Chomskyan/generative linguists*** seek to describe the language module of the human mind (the ***I-language***) for which data such as text (the ***E-language***) provide only indirect evidence, which can be supplemented by native speakers intuitions.
- Empiricists approaches are interested in describing the E-language as it actually occurs.
- Chomskyans make a distinction between ***linguistic competence*** and ***linguistic performance***. They believe that linguistic competence can be described in isolation while Empiricists reject this notion.

Today's Approach to NLP

- From ~1970-1989, people were concerned with the science of the mind and built small (toy) systems that attempted to behave intelligently.
- Recently, there has been more interest on engineering practical solutions using automatic learning (knowledge induction via machine learning including deep learning).
- While Chomskyans tend to concentrate on categorical judgements about very rare types of sentences, statistical NLP practitioners concentrate on common types of sentences.

Why is NLP Difficult?

- NLP is difficult because Natural Language is highly ambiguous.
- Example: "*The company is training workers*" has 2 or more ***parse trees*** (i.e., syntactic analyses).
- "*List the sales of the products produced in 1973 with the products produced in 1972*" has 455 parses.
- Therefore, a practical NLP system must be good at making disambiguation decisions of word sense, word category, syntactic structure, and semantic scope.

Methods that don't work well

- Maximizing coverage while minimizing ambiguity is inconsistent with symbolic NLP.
- Furthermore, hand-coded syntactic constraints and preference rules are time consuming to build, do not scale up well and are brittle in the face of the extensive use of metaphor in language.
- ***Example:*** if we code

  animate being --> **swallow** --> physical object

  I swallowed his story, hook, line, and sinker.

  The supernova swallowed the planet.

What Statistical NLP can do for us

- Disambiguation strategies that rely on hand-coding produce a knowledge acquisition bottleneck and perform poorly on naturally occurring text.
- A Statistical NLP approach seeks to solve these problems by automatically learning lexical and structural preferences from corpora. In particular, Statistical NLP recognizes that there is a lot of information in the relationships between words.
- The use of statistics offers a good solution to the ambiguity problem: statistical models are robust, generalize well, and behave gracefully in the presence of errors and new data.

Corpora

- Brown Corpus – 1 million words
- British National Corpus – 100 mil. words
- American National Corpus – 10 mil. words -> 100
- Penn TreeBank - parsed WSJ text
- Canadian Hansard – parallel corpus (bilingual)
- English Gigaword Corpus
- Wikipedia dumps

Dictionaries:

- Longman Dictionary of Contemporary English
- WordNet (hierarchy of synsets)
- Wiktionary

Things that can be done with Text Corpora (I) Word Counts

- ***Word Counts to find out***:
  - What are the most common words in the text.
  - How many words are in the text (word tokens and word types).
  - What the average frequency of each word in the text is.
- ***Limitation of word counts:*** Most words appear very infrequently and it is hard to predict much about the behavior of words that do not occur often in a corpus. ==> Zipf's Law.

Things that can be done with Text Corpora (II)
Zipf's Law

- If we count up how often each word type of a language occurs in a large corpus and then list the words in order of their frequency of occurrence, we can explore the relationship between the frequency of a word, ***f,*** and its position in the list, known as its rank, ***r***.
- Zipf's Law says that: $f \propto 1/r$
- Significance of Zipf's Law: For most words, our data about their use will be exceedingly sparse. Only for a few words will we have a lot of examples.

Things that can be done with Text Corpora (III) Collocations

- A ***collocation*** is any turn of phrase or accepted usage where somehow the whole is perceived as having an existence beyond the sum of its parts (e.g., disk drive, make up, bacon and eggs).
- Collocations are important for machine translation.
- Collocations can be extracted from a text (example, the most common ***bigrams*** can be extracted). However, since these bigrams are often insignificant (e.g., “at the”, “of a”), they can be ***filtered***.

Things that can be done with Text Corpora (IV) Concordances

- Finding ***concordances*** corresponds to finding the different contexts in which a given word occurs.
- One can use a ***Key Word In Context (KWIC)*** concordancing program.
- Concordances are useful both for building dictionaries for learners of foreign languages and for guiding statistical parsers.