UNIVERSITY of HOUSTON
DEPARTMENT OF COMPUTER SCIENCE

Corpus-based Work

Rakesh Verma

Requirements for Working with Corpora

- Historically, working with large text corpora was a major difficulty.
  - The Brown Corpus was originally compiled on an IBM 7070, which had the equivalent of 40 KB of memory.
  - Sorting the Corpus took 17 dedicated computer hours.
  - The corpus itself is 10.3 MB, filling up the IBM 7070's memory over 2500 times over.
- In the modern age of computing, this difficulty no longer exists.
  - Computers now come with memory on the order of gigabytes.
  - Data storage is now on the order of terabytes (1 TB = 1024 GB).
  - Even smartphones are capable of sorting the Brown Corpus in minutes.
- Memory and storage are the primary concerns when working with corpora.
  - Computers will only become better and more affordable in these regards.

Selecting Corpora

A corpus is a special collection of text resources that match certain criteria.

Brown Corpus was designed as a **representative sample** of American English.

A sample is representative if it coincides with the population of interest in general.

It's important that the corpus you work with is representative of the application for which it's used.

Interacting with Corpora

Text will often come in two forms.

- **Raw** text will be collected in its original form, as written or typed.
- **Markup** is a set of patterns which can be easily interpreted by a computer to carry additional information about the text's structure or format.
  - Good markup for NLP is markup that is also easily human-interpreted (ex. XML)

Manually creating markup for a raw text is effort intensive.

Creating tools to automatically produce markup has its own difficulties.

Low-level Formatting Issues

Depending on the source of the corpus, there may be various formatting and content that cannot be dealt with automatically.

Optical Character Recognition (**OCR**) is the process by which text from written sources is scanned in and digitized as electronic text data.

Diagrams, images, and tables present text in nonstandard formats, which OCR is not designed to deal with.

Resulting output files from OCR could contain garbled text from misrecognizing diagram pieces as text or reading figure captions as connected continuations of the text's paragraphs.

Before continuing, such garble must be filtered out and (optionally) parsed manually.

Letter Case

- How to handle capitalization is an open question in modern NLP.
  - Two words are identical in all but capitalization.
  - Are they the same?
- A rigid approach is to convert everything to lowercase.
  - We will have equal treatment of "The apple" at the beginning of a sentence and "the apple" in the middle of a sentence.
  - However, we will lose distinction of **proper names** (ex. "Brown" and "brown").
- A more flexible approach is to lowercase words at the start of sentences.
  - Preserves proper names in the middle of sentences.
  - Unfortunately, proper names at the beginning lose their distinction.
- Both cases fail to encapsulate capitalization as emphasis.
  - In some genres, this emphasis is VERY important to meaning.

Tokenization

The process of processing text into word tokens is called **tokenization**.

What is a word?

Kučera and Francis give the definition of a **graphic word**.

- "A string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctuation marks."
- This definition does not recognize monetary values (ex. $19.99) as words and fails to identify certain brand names as words (ex. C|Net).

One of the major indicators in English is the presence of **white space**: a gap in written text or digitally any invisible character that gives this same gap.

- Languages like Chinese and Japanese do not use spaces.

This indicator alone does not necessarily or sufficiently denote a word.

Periods

- Not all words are followed by terminating whitespace.
  - Consider the word "whitespace" in the sentence above which ends in '.'
- Punctuation marks attach to words either "before or after".
- It seems that removing all punctuation will aid in identifying word tokens.
- This introduces a new problem for periods.
  - Periods are not only used as sentence terminators.
    - "Mr. Smith is a secret agent."
    - "Apples, oranges, etc."
    - "*Wash. D.C.*"

Single Apostrophes

English has many syntactic structures that include single apostrophes.

- Contractions: I'll, isn't, there's
- Possession: Alice's dog, Bob's cat, the cat that bob own's kittens.

When we treat contractions as a single word token, we violate the English grammar rule:

- Sentence → [Noun Phrase] [Verb Phrase]

When we split on contractions, we introduce potentially wrong word tokens.

- hasn, shouldn, t, and ll

In the case of possession, 's is called a **clitic**.

- A clitic is an affix that has a syntactic role at the phrase level.
- The Penn Treebank opts for splitting tokens at 's.

Hyphenation

Do hyphenated words count as one word or two?

There are three classes of hyphen in English:

- Typographical Hyphens
  - Hyphens are often used in the middle of a word between stressed syllables if it would be cut off at the end of the page. These words should count as one word.
- Lexical Hyphens
  - Lexical hyphens are used to separate repeated vowel sequences or attach small word formatives (ex. co-operate, e-mail, non-obligatory). These are often one word.
- Grouping
  - These hyphens are used for combining a phrase into a distinct quantity, rate, or description (ex. well-adjusted, 90-cents-an-hour, 24-years-old, Mr. Goody-two-shoes).
  - These combinations carry meaning from each of their parts (ex. 24-years-old), while some are common enough to have a distinct meaning as a whole (ex. goody-two-shoes)

Homographs

Homographs are lexemes which share a written form.

- To **bow** is to bend your head or body forward as a sign of respect.
- You wear a **bow** in your hair.
- An archer fires arrows from their **bow**.

These definitions are disjoint and must be distinguished from each other by their parts-of-speech.

In the above example, bow is used as a noun twice.

- In some cases, it's insufficient to only use the part-of-speech.
- Context information must be used here to distinguish the usage.

Existing Frameworks for Working with Corpora

- The **Natural Languages Toolkit** (NLTK) is a platform for building Python programs to work with human language data.
  - It comes with interfaces for many of the corpora mentioned in past classes.
- **WordNet** is a large lexical database of English.
  - Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called **synsets**.
- **VerbNet** is the largest online verb lexicon currently available for English.
  - It is a hierarchical, domain-independent, broad-coverage verb lexicon with mappings to other lexical resources such as WordNet, FrameNet, and Xtag.
- **FrameNet** is an electronic resource based on a theory of meaning called **frame semantics**.
  - A semantic frame can be thought of as a conceptual structure describing an event, relation, or object and the participants in it.

References

Foundations of Statistical Natural Language Processing (Christopher D. Manning and Hinrich Schütze)