Corpus Linguistics

CS140 NL Annotation for ML

January 22, 2016

James Pustejovsky

Corpus Linguistics vs. Chomsky's Cognitivism

A Brief Overview of Recent History

Slides from Anna Rumshisky

Structuralist Tradition

- Bloomfeld and others, 1940s – 1950s
  - Language can be explained in probabilistic, behaviorist terms
  - Languages are diverse systems learned from the environment
  - The aim was to describe the diversity of linguistic behavior; analyze linguistic structure in formal terms – producing formal descriptions of grammar (including phonetics, morphology, syntax, etc.)

Early NLP

- Empirical and statistical methods were popular the 1950s
- Shannon's information-theoretic approach to language
  - *All of us were convinced that speech, in English or any other language, was a Markov process. From this to the conviction that ... the set of all English sentences can be generated by a Markov source was only a small step.* (Bar-Hillel, 1975)
- Early machine translation efforts of 1950s and 1960s

Chomsky vs. Corpus Linguistics

- Popularity of empirical and statistical methods faded in the 1960s under the 'cognitive revolution'
- Chomsky's mentalistic, generative approach to language revolutionized linguistics and cognitive science in the 20th century
- Influential Events
  - Chomsky's *Syntactic Structures* (1957) and *Aspects of the Theory of Syntax* (1965)
  - Chomsky & Miller's critiques of statistical language models
  - Chomsky's critique of Skinner's *Verbal Behavior*

Chomsky

- *I had no personal interest in the experimental studies and technological advances. [...] As for machine translation and related enterprises, this seemed to me pointless, as well as probably quite hopeless. [...] I could not fail to be aware of the ferment and excitement [in the early 1950s]. But I felt myself no part of it.* (Chomsky, 1975)

1969 – Whither Speech Recognition?

General purpose speech recognition seems far away. Social-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish…

It would be too simple to say that work in speech recognition is carried out simply because one can get money for it. That is a necessary but not sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamour…

Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve "the problem." The basis for this is either individual inspiration (the "mad inventor" source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach).

The Journal of the Acoustical Society of America, June 1969

J. R. Pierce
Executive Director, Bell Laboratories

(Thanks to Julia Hirshberg, Roberto Pieraccini and Francis Ganong for some slides)

Syntactic Structures (Chomsky 1957)

From now on I will consider a language to be a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements.

The fundamental aim in the linguistic analysis of a language L is to separate the grammatical sequences which are the sentences of L from the ungrammatical sequences which are not sentences of L and to study the structure of the grammatical sequences.

On what basis do we actually go about separating grammatical sequences from ungrammatical sequences?

Syntactic Structures (Chomsky 1957)

- *First, it is obvious that the set of grammatical sentences cannot be identified with any. . . finite and somewhat accidental corpus of observed utterances. . .*
- *Second, the notion "grammatical" cannot be identified with "meaningful" or "significant" in any semantic sense.*
- *Sentences (1) and (2) are equally nonsensical, but any speaker of English will recognize that only the former is grammatical.*

  (1) Colorless green ideas sleep furiously.

  (2) Furiously sleep ideas green colorless.

Syntactic Structures (Chomsky 1957)

- *Third, the notion "grammatical in English" cannot be identified in any way with the notion "high order of statistical approximation to English." It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse.*
- *Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English. Yet (1), though nonsensical, is grammatical, while (2) is not.*

Syntactic Structures (Chomsky 1957)

- *Evidently, one's ability to produce and recognize grammatical utterances is not based on notions of statistical approximation and the like. . . I think that we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure.*

Refuting Chomsky's arguments

- *. . . in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English.*
- *(1) Colorless green ideas sleep furiously. (2) Furiously sleep ideas green colorless.*
- This claim is false—all modern statistical models of language can assign probabilities to previously-unseen utterances.
- E.g., Pereira's (2000) statistical model of newspaper text assigns (1) a probability 200,000 times greater than (2)

Refuting Chomsky's arguments

- Even having pos classes is sufficient for this
- say, over 1M word Brown Corpus Penn Treebank Tagset

  http://www.mozart-oz.org/mogul/doc/lager/brill-tagger/penn.html
- $ cat * | tr '\n' ' ' | egrep -o '\w+/jj \w+/jj \w+/nn\w* \w+/vb\w* \w+/rb'

  brilliant/jj white/jj flares/nns swayed/vbd eerily/rb little/jj green/jj biplane/nn struggled/vbd northward/rb fake/jj therapeutic/jj devices/nns figure/vb prominently/rb routine/jj vital/jj statistics/nns got/vbd nowhere/rb
- $ cat * | tr '\n' ' ' | egrep -o '\w+/rb \w+/vb\w* \w+/nn\w* \w+/jj \w+/jj'

Cognitivists vs. empiricists

- Chomsky emphasized “creativity” of language, as manifested in recursive generative rules

  S => NP VP, VP => VP NP

- Empiricists emphasize common language patterns (e.g. collocations) and predictability of language
- Warren Weaver, pioneer of MT (1949)

  about half of the letters or words we choose in writing or speaking (although we are not ordinarily aware of it) are really controlled by the statistical structure of the language.

Contrastive viewpoints

- Chomsky (1957):

  I think that we are forced to conclude that grammar is autonomous and independent of meaning

- Corpus linguist John Sinclair (1991):

  it is folly to decouple lexis and syntax, or either of those and semantics. The realization of meaning is far more explicit than is suggested by abstract grammars. The model of a highly generalized formal syntax, with slots into which fall neat lists of words, is suitable only in rare uses and specialized texts. By far the majority of text is made of the occurrence of common words in co˜mmon patterns. Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text.

Further reading

- *Mind as machine: a history of cognitive science (2008)* by Margaret A. Boden
  - *up on GoogleBooks*

Corpus Properties

Corpora for linguistic research

- It is quite typical for researchers to use *any collection of texts* for linguistic analysis.
  - Often proceed opportunistically: whatever data comes in handy is used.
- However, the term *corpus* usually implies the following characteristics:
  - sampling/representativeness
  - finite size
  - machine-readable form
  - a standard reference
  - (time-bound)

Sampling and representativeness

- Sampling is a fundamental characteristic of any empirical work.
  - It is impossible to study every single instance of a phenomenon of interest.
  - With language, this is even more difficult: languages change continuously.
  - A corpus is a “snapshot” of the language at a specific time.
  - More on sampling in Part II

Finite size

- Usually, corpora have a fixed size.
  - E.g. BNC is 100 million words
- But not always. Some corpora keep growing over time.
  - Example: COBUILD Corpus built at Birmingham university is periodically updated.
  - Very useful for lexicographic work: if the corpus is updated regularly, it remains a good source of new words and usages.

Static and non-static

- Sample corpus:
  - a corpus which represents a sample of a language within a specific period
  - BNC is a good example of this, covers 1960-1993
- Monitor corpus:
  - a dynamic sample
  - normally covers a relatively brief span of time (i.e. decades, not centuries)
  - updated regularly to keep track of changes within the language

Time

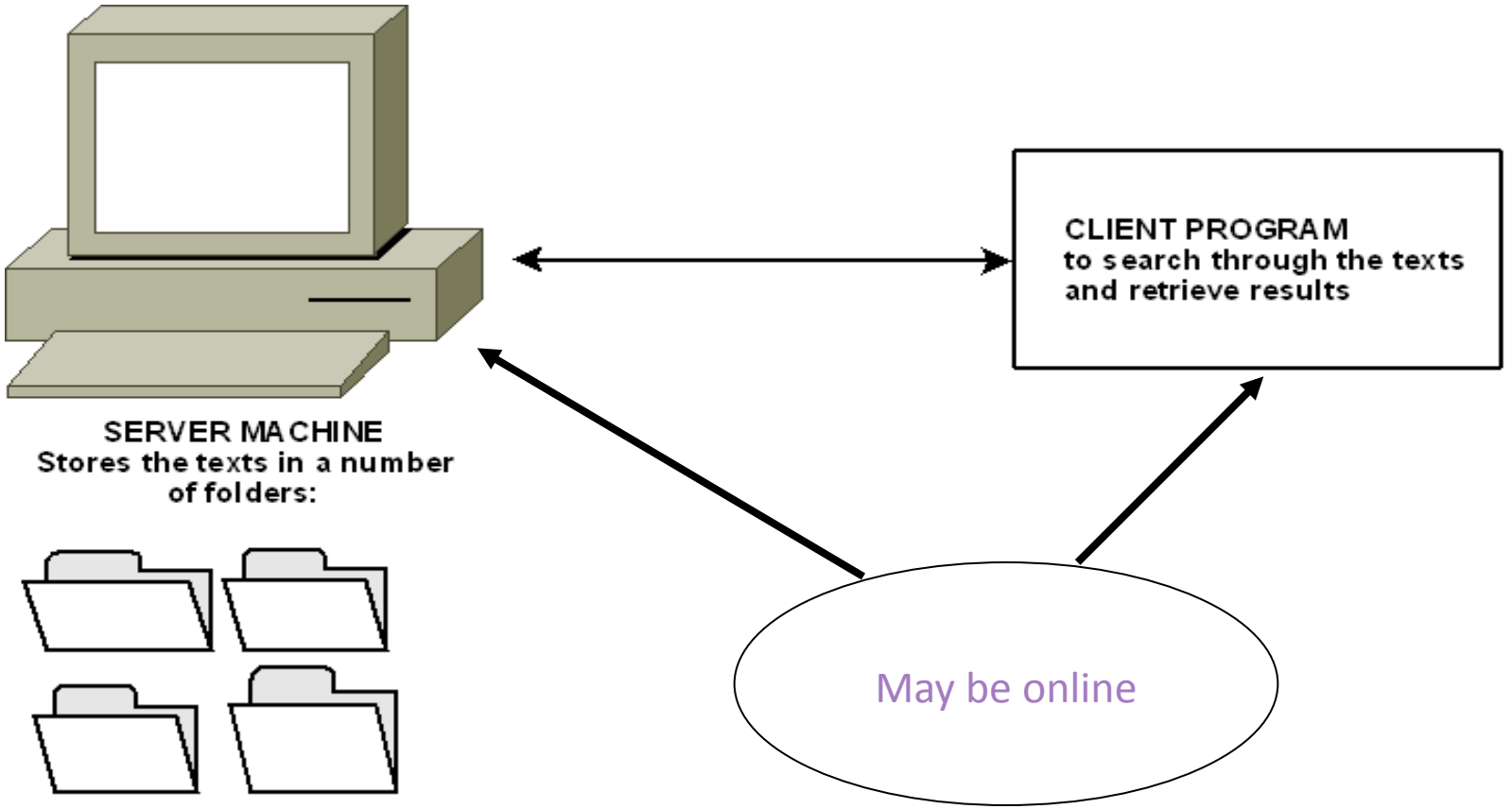
- Unless the corpus is a monitor corpus, the sampling will inevitably mean that we're restricted to a period of time.

- Can have interesting consequences:
  - Do you think the English language has changed since 1993? What aspects will have changed? Lexicon? Syntax?

Machine readability

- Very rare for a corpus nowadays to be in print.
- We've seen some advantages of machine-readability before.

What “machine readable” really means



Client programs for corpus search

- Tools for searching through large collections of plain text (with/out annotation). E.g.
  - WordSmith
  - MonoConc Pro
  - Very useful to build frequency lists etc…
- Corpus-specific clients E.g.:
  - SARA
    - program created for the BNC
    - sensitive to the specific annotations in the BNC
    - allows search for patterns such as DETERMINER+NOUN
- Online servers with web-based client
  - SketchEngine, etc
  - Increasingly popular

A standard reference

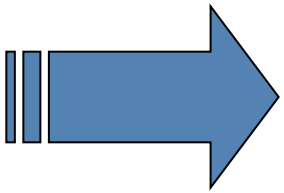
- This is not an essential aspect of a corpus, but it is useful.
- It presupposes:
  - wide availability
  - broad coverage
- If a corpus is a standard reference, then it becomes:
  - a common source of data, hence studies are replicable
  - a yardstick against which to measure other, newer corpora

Populations, samples and sampling

Samples and populations

Population

"the group (of people or things) which are of interest to the study"



Sample

a smaller, representative group selected from the population

Sampling to avoid skewness

- Remember Chomsky's criticism about the skewness of corpora:
  - any sample of the language will be biased, including some things but not others
- This is rather like sampling from the human population:
  - psychologists who select samples of people for experiments know that skewness is a risk
- A good sample should capture the variability in a population.

Prerequisites for sampling

1. definition of the boundaries of the population
   - written part of the BNC: English published within the UK between 1960 and 1993
   - Brown Corpus: written English published in the US in 1961
2. definition of the sampling unit
   - books, periodicals, radio broadcasts…
3. sampling frame = the list of sampling units
   - Brown Corpus: the list of books and periodicals in the Brown University Library and the Providence Athenaeum.
   - BNC: more sophisticated, considered who wrote what and who was the target audience

Defining the language population

1. language production
2. language reception
   - Both of these are demographically-oriented.
   - focus on characteristics of the producer or receiver
     - sex, age, social class…
   - typical of the approach in the BNC
3. language as product
   - starting point is "what's out there", irrespective of who produced it and for whom
   - typical of the approach in the Brown Corpus

Sampling in the BNC

- Population definition looked at both production and reception
- Sources for production (who publishes what?):
  - Catalogues of books published per annum
  - Lists of books in print
- Sources for reception (what is read by whom?):
  - bestseller lists & prizewinners
  - library lending statistics

Sampling techniques

- Once population is defined and sampling frame identified, actual sampling can proceed in several ways:

1. simple random sampling: identify a subset randomly from the total set of sampling units in the frame
   - ☐ may omit rare items in the population, because if X is more frequent than Y, X's chances of being selected are higher
2. stratified random sampling:
   a. split population into relatively homogeneous groups or strata
   b. sample each stratum randomly

Sampling of written text in the BNC

- After sources were selected based on production/reception criteria, they were classified on the basis of 3 main features:
  - domain ("subject")
    - imaginative, arts, belief and thought, …
  - time (when published)
    - 1960 – 1974; 1975 – 1993
  - medium
    - book, periodical, written-to-be-spoken, etc
- These then determine the strata for sampling in the BNC.

Sampling of spoken discourse in the BNC

- The features defining the sampling frame differ for spoken language:
  - demographic component
    - informal conversation recorded by 124 volunteers
    - selected by age, sex, social class, geographical region
  - context-governed component
    - more formal encounters
    - meetings, lectures, etc

Balance and representativeness

Balance and representativeness

- Balance:
  - refers to the range of types of text in the corpus
  - e.g. the BNC's construction was based on an *a priori* classification of texts by *domain*, *time* and *medium*
- Representativeness:
  - refers to the extent to which the corpus contains the full range of variation in the language.
- Representativeness depends on balance as a prerequisite.

When is a corpus representative?

- Biber (1993):
  - "Representativeness refers to the extent to which a sample includes the full range of variability in a population".
- What variability?
  - variability of text types (different genres, different registers)
  - variability of linguistic phenomena (lexical, syntactic)
    - Not all linguistic features are distributed in the same way

Variability in distributions

- Active, declarative clauses are probably more frequent overall than passives.
  - But passives become very frequent in certain types of text (e.g. academic discourse).
- Certain word orders are "marked", hence probably less frequent than the unmarked.
  - *cf.* SVO vs other orders in Maltese

Variability in distributions

- Some words may be completely absent in everyday usage, but highly frequent in specialised registers.
  - neutrino, morpheme, palato-alveolar…
- The same is true of word senses:
  - *qoxra* (MT) = *shell* – probably the most frequent sense
  - *qoxra* can also mean “seafaring vessel” (*qoxra tal-baħar*)
    - more likely to be used in this sense in the fishing/ sailing register

The need for a priori criteria

- Problem:
  - before we begin to sample for representativeness, we need a notion of what the range of variability is.

- Therefore some criteria need to be defined a priori.

Linguistic variability and text type

- It is likely that genre or register or text type is a determining factor of linguistic variability.
- All the foregoing examples were made with reference to text type.
- Two plausible views:
  1. sample based on text type to capture linguistic variability (as in the BNC)
  2. sample based on a predefined model of what linguistic variability there is

External (situational) criteria

- Define sampling frames by the social and communicative contexts in which a particular sample of text/speech is produced.
- Biber (1993) suggested external criteria should determine the sampling frame to ensure representativeness.
- Under this view, texts are selected to cover a predefined range of uses/purposes/contexts. This is the BNC approach.

External criteria

- Sampling based on situational criteria would proceed as follows:
  1. identify the range of types / genres/ registers
  2. identify the units within each type
     - NB: The size of each category will reflect how widespread or common the type is
  3. sample from the units within each type

Internal (“linguistic”) criteria

- Define sampling frames on the basis of linguistic features (e.g. lexico-grammatical) that distinguish texts.
- Example: “to be representative our corpus should contain the majority of (word) types in the language, as defined in some standard dictionary”

Potential problems with internal criteria

- Internal criteria risk becoming circular:
  - you need a good linguistic resource (such as a corpus) to study the distribution of relevant features
  - but you're need the features to design the corpus!

Balance between text types

- We've noted that representativeness depends on balance:
  - language variation is captured in the sample if it comes from the same sources that determine the variation
- But balance is very difficult to assess.
  - Depends on an agreed-upon definition of what the range of text types is.

The notion of "domain" in the BNC

- imaginative (21.91%)
- arts (8.08%)
- belief and thought (3.40%)
- commerce/finance (7.93%)
- leisure (11.13)
- natural/pure science (4.18%)
- applied science (8.21%)
- social science (14.80%)
- world affairs (18.39%)
- unclassified (1.93%)

- Why represent commerce/ finance separately?
- Why is commerce/finance more represented than arts?
- Why not have a separate category for "poetry"?

The notion of "medium" in the written BNC

- book (55.58%)
- periodical (31.08%)
- misc. published (4.38%)
- misc. unpublished (4%)
- to-be-spoken (1.52%)
- unclassified (0.4%)

- Why more books than periodicals? Aren't periodicals more numerous?
- Why not more "unpublished"? Most written discourse remains unpublished.

Summary

- Sampling (in general)
  - inclusion of a subset of the relevant units in a population, to ensure representativeness of relevant features
- Balance
  - ensuring that the range of types of text is represented correctly in the sample
- Representativeness
  - ensuring that interesting variation of linguistic features is captured

Summary

- To achieve representativeness, we need to ensure balance.
- Balance is usually achieved through external criteria.
  - These are used to determine the sampling frame.