Language Data Resources – Overview

Zdeněk Žabokrtský, Martin Popel

7th October 2020
Outline

Why data, and why is it important?

Corpora

Specialized corpora

Lexicon-like Data Resources

And many other types of language resources

Final remarks
Why data, and why is it important?
Why data?

- An elegant answer by Tony McEnery (in 2005): “Corpus data are, for many applications, the raw **fuel of NLP**, and/or the **testbed** on which an NLP application is evaluated.”
- This is important: you typically need the data not only for developing your application, but also for measuring its quality.
Why evaluation?

Why so much focus on performance evaluation right at the beginning?

- Some people from other IT fields think that presenting some weird measures is only about saturating ambitions of academic scholars...
- No, it’s not!
- So what’s so special about NLP?
- In fact we still don’t understand how language works (compared, e.g., to the level of understanding in mechanics, electromagnetic field or anorganic chemistry).
- In NLP, all our solutions are still only approximative nowadays, and often far from perfect.
- Subjective evaluation is slow and costly.
Typically, the data shows us the “ground truth” which our application tries to mimic (e.g., a possible correct translation of a sentence in language A into language B)

Whenever possible, we use a fully automatized comparison with the ground-truth data as a performance evaluation measure which tells us what works and what doesn’t!

The data gives us the “gradient” during the development of the application, both in the metaphorical sense (for more abstract design decisions) and in the literal sense (if machine learning is used, but this is almost always the case nowadays).

Sometimes, the evaluation role of the data is more crucial.
In general, when studying a specific language phenomenon or developing an end-user natural language applications, there are two basic ways to go:

- thinking about it in the context of one’s language experience, using introspection...
- or using **empirical evidence**, statistical models based on real world usage of language …

TODO: premistit do sekce o anotaci

- side remark: this includes also using brain-imaging methods or at least eye-tracking devices, but such approaches are still rare in the real NLP industry
From scepticism about linguistic data to modern data crunching?

- **1957:** Noam Chomsky’s attack: “Any natural corpus will be skewed. Some sentences won’t occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list.”

- **1988:** Frederick Jelinek: ”Every time I fire a linguist, the performance of the speech recognizer goes up” (perhaps not an exact citation)

- **200?** Eugene Charniac: “Future is in statistics.”

- **200?** Eric Brill: “More data is more important than better algorithms.”
The world of language data resources today

• Today’s language data resources map - hopelessly diverse.
• more than 1,000 submissions to every LREC (International Conference on Language Resources and Evaluation, biannual)
• data centers offering zillions of various language data packages
Why researchers develop/use so many different pieces of data?

• Is the natural language really so complex? Well, yes.
• In addition,
  • thousands of languages (plus dialects), different writing systems…
  • sometimes $O(L^2)$ when pairs of languages considered (especially in parallel data for machine translation)
  • many underlying theories
  • many end-application purposes
Let’s try to systematize the space of data resources

Classification along basic dimensions:

- corpus vs. lexicon
  - corpus - a collection of authentic utterances in a given language (texts composed of sentences composed of words)
  - lexicon in the broad sense, as a repertory of tokens’ types
- modality: spoken vs. written
  - and other, e.g. sign languages
- number of contained languages: monolingual vs. multilingual
  - a single language only vs. several languages
  - if multilingual, then possibly parallel

Why data, and why is it important?

Corpora
Specialized corpora
Lexicon-like Data Resources
And many other types of language resources
Final remarks
Classification along basic dimensions, cont.

- time axis: synchronic vs. diachronic
  - only the contemporary language (such as a few last decades) vs. a collection of texts along longer time space (such as several centuries)

- raw vs. annotated
  - if annotated, then what on which “level” (which language phenomena are captured), with which underlying theory, with what set of labels (tag set) …

- other language variables:
  - original vs. translation
  - native speaker vs. learner
  - various kinds of language disorders …

Why data, and why is it important?    Corpora  Specialized corpora  Lexicon-like Data Resources  And many other types of language resources  Final remarks
Corpora
Full Definition of CORPUS

**plural corpora** \[\text{\textcopyright} \& p(\text{e})r\text{e}\]

1. the body of a human or animal especially when dead

2. **a**: the main part or body of a bodily structure or organ <the corpus of the uterus>
   **b**: the main body or corporeal substance of a thing; *specifically*: the principal of a fund or estate as distinct from income or interest

3. **a**: all the writings or works of a particular kind or on a particular subject; *especially*: the complete works of an author
   **b**: a collection or body of knowledge or evidence; *especially*: a collection of recorded utterances used as a basis for the descriptive analysis of a language
• linguists recognized the need for unbiased empirical evidence long before modern NLP
  • excerption tickets collected systematically for Czech from 1911
Corpus size

- typically measured in tokens (words, numerals, punctuation marks ...)
- sampling is inescapable
  - an I-want-it-all corpus is far beyond our technology (even in a strictly synchronic sense)
- but still, the corpora sizes have been growing at an exponential pace for some time:
  - Brown Corpus in 1964 \(\approx 1\text{MW}\)
  - (electronic corpus of Czech texts in 1970s: 500kW)
  - British Natural Corpus in 1994 \(\approx 100\text{MW}\)
  - English Gigaword in 2004 \(\approx 1\text{GW}\)
  - Araneum Russicum Maximum in 2018 \(\approx 10\text{GW}\)

- so roughly 1 order of magnitude per decade
- but no exponential growth lasts forever - what’s the upper limit?
- probably at least 1 TW for "big" languages: web-based data - Google’s 5-gram dataset for 10 European Languages in 2009 based on \(\approx 1\text{TW}\)
Balanced corpora

• a discussion that used to be heated some decades ago
• an (IMHO elusive) goal: a balanced corpus whose proportions correspond to the real
  language usage
• criteria for choosing types of texts their relative proportion in the corpus (and eventually
  concrete texts)?
  • style, genre
  • reception vs. perception (a few influential authors vs. production of a large community)?
• actually no convincing generally valid answers for an optimal mixture …
• …but at least some strategies seem to be more reasonable than others
• the most famous example of clearly imbalanced corpora: the Wall Street Journal Corpus
  • unfortunately used as a material source for the Penn Treebank, which is undoubtedly
    among the most influential LR
  • “NLP = Wall Street Journal science”
• an example of a clearly imbalanced corpus: Wall Street Journal Corpus
  • unfortunately used as a material source for the Penn Treebank, which is undoubtedly among the most influential LR
  • an occasional criticism of NLP: “NLP = a Wall Street Journal science”
• a lesson taken
  • no broadly accepted criteria for balancing a corpus
  • thus we can hardly reach a perfectly balanced corpus
  • but we should at least avoid building a strikingly imbalanced corpus
Corpus annotation

- raw texts – difficult to exploit
- solution: gradual “information adding” (more exactly, adding the information in an explicit, machine tractable form)
- annotation = adding selected linguistic information in an explicit form to a corpus
- some examples of possible annotations:
  - morphological annotation (assign a lemma, part of speech and other morphological categories to all tokens)
  - syntactic annotation – phrase-structure or dependency syntactic trees
  - semantic annotation – dozens of theoretical approaches, no consensus so far
  - anaphora – e.g. what pronouns refer to
  - word sense disambiguated corpora – which sense is used for a given polysemous word in a given context
  - sentiment corpora – positive/negative emotions induced by some expressions in a text
Specialized corpora
Parallel corpora

• specific feature: alignment between corresponding units in two (or more) languages
  • document level alignment
  • sentence level alignment
  • word level alignment
  • (morpheme level alignment?)
• example: The Rosetta Stone
• example: CzEng - a Czech-English parallel corpus, roughly 0.5 words for each language, automatically parsed (using PDT schema) and aligned
Named entity corpora

• specific feature: instances of proper names, such as names of people, geographical names,

• example: Czech Named Entity Corpus - two-level hierarchy of 46 named entity types, 35k NE instances in 9k sentences


ČINNOST POBOČKY EVROPSKÉ BANKY PRO OBNOVU A ROZVOJ ( BERD ) v Praze slavnostním přestřižením stuhy včera zahájil president BERD Jacques Attali.
Coreference corpora

• specific feature: capturing relations between expressions that refer to the same entity of the real world

(credit: Shumin Wu and Nicolas Nicolov)

• example: Prague Dependency Treebanks (around 40k coreference links in Czech texts)
 Sentiment corpora

- specific feature: capture the attitude (in the sense of emotional polarity) of a speaker with respect to some topic/expression
- simply said: “is this good or is it bad?”
- obviously over-simplified, but highly demanded e.g. by the marketing industry

(credit: SemEval 2014 documentation)

- example: MPQA Corpus
Highly multi-lingual corpora

• specific feature: as many languages as possible
• examples:
  • W2C - at least 1MW for more than 100 languages
  • The Bible Corpus - translations of the Bible into 900 languages
  • The OPUS Corpus - the open parallel corpus – 60 languages
Lexicon-like Data Resources
Inflectional lexicons

- specific feature: capturing the relation between a lemma and inflected word forms, ideally in both directions
- example: MorfFlex CZ, around 120M word forms associated with 1M lemmas
• specific feature: capturing the relation between a base word and a derived word (typically by prefixing and/or suffixing)

• example: DeriNet, 1M lemmas, 700k derivation links
• specific feature: capturing semantic relations between words, such as synonymy and antonymy

• example:

Main Entry: great
Part of Speech: adjective
Definition: excellent, skillful
Synonyms: able, absolute, aces, adept, admirable, adroit, awesome, bad*, best, brutal, cold*, complete, consummate, crack*, downright, dynamite, egregious, exceptional, expert, fab, fantastic, fine, first-class*, first-rate, good, heavy*, hellacious, marvelous, masterly, number one, out of sight, out of this world, out-and-out, perfect, positive, proficient, super-duper, surpassing, terrific, total, tough, transcendent, tremendous, unmitigated, unqualified, utter, wonderful
Antonyms: ignorant, menial, poor, stupid, unskilled, weak

* = informal/non-formal usage
Wordnets

- specific feature: hyponymy (hyperonymy) forest composed of synsets (sets of synonymous words)
- example: Princeton Wordnet
• specific feature: wordnets of several languages interconnected through English as the hub language

**Architecture of the EuroWordNet Data Base**

(credit: intuit.ru)
Valency lexicons

- specific feature: capturing combinatory potential of a word (most frequently of a verb) with other sentence elements

- example: VALLEX - Valency Lexicon of Czech Verbs

  odpovídat\textsuperscript{odf}, odpovědět\textsuperscript{odf}

  1. odpětíř, děvát odpověď
     
     **Frame:**
     
     - \textit{ACT}\textsuperscript{obl}
     - \textit{ADDR}\textsuperscript{obl}
     - \textit{PAT}\textsuperscript{opt}
     - \textit{EFF}\textsuperscript{obl}
     - \textit{MANN}\textsuperscript{top}
     - \textit{MEAN}\textsuperscript{top}
     
     **Example:**
     
     \textit{inf:} odpovídám mu na jeho dotaz pravdu / činím / smířím / že .. \textit{pf:} odpověděm mu na jeho dotaz pravdu / činím / smířím / že ..

  2. inf: reagovat\textsuperscript{pf}, reagovat
     
     **Frame:**
     
     - \textit{ACT}\textsuperscript{obl}
     - \textit{PAT}\textsuperscript{opt}
     - \textit{EFF}\textsuperscript{obl}
     
     **Example:**
     
     \textit{inf:} pokouším odpovídat na chlad zarudnutým, gruzínští milicionií neodpovídali střelbu (SYN) \textit{pf:} 
     
     
     vyjádřit odpovědění střelbu (SYN); na výzvu dobře odpověděl zmenou vlastního politického chování (SYN)

  3. lmit odpovědět\textsuperscript{mif}, mit odpovědnost
     
     **Frame:**
     
     - \textit{ACT}\textsuperscript{obl}
     - \textit{ADDR}\textsuperscript{obl}
     - \textit{PAT}\textsuperscript{opt}
     - \textit{MEAN}\textsuperscript{top}
     
     **Example:**
     
     odpovídá za své děti; odpovídá za ztrátu svým majetkem

  4. lmit odpovídat\textsuperscript{mif}, být ve shodě / v souhlasu; korespondovat
     
     **Frame:**
     
     - \textit{ACT}\textsuperscript{obl}
     - \textit{PAT}\textsuperscript{obl}
     - \textit{REG}\textsuperscript{top}
     
     **Example:**
     
     řešení odpovídá svými vlastnostmi požadavkům
And many other types of language resources
Speech corpora

- specific feature: recordings of authentic speech, typically with manual transcriptions
- for training Automatic Speech Recognition systems
- example: The Switchboard-1 Telephone Speech Corpus, 2,400 telephone conversations, manual transcriptions
Datasets primarily unintended as corpora

- Web as a corpus
- Wikipedia as a corpus
- Enron corpus - 600,000 emails generated by 158 employees of the Enron Corporation
“Metainformation” about languages

• example: The World Atlas of Language Structures (WALS)
  • http://wals.info/
  • specific feature: various language properties (related e.g. to word order, morphology, syntax) captured for hundreds of languages
Final remarks
“Let’s consider any language, for example English...”
“Our approach is theory neutral and language independent...”

• Be very careful when you hear (or say) that some language data resource (or an annotation scheme, or a probabilistic model, or a technological standard...) is
  • theory neutral,
    • If fact we cannot “measure” language structures per se, and thus we always rely on some assumptions/abstractions/conventions etc.
  • or language independent.
    • In fact it is impossible for a human (a linguist or an NLP developer) to take into account all the diversity in morphology/syntax/semantics in all languages.
    • Keep in mind: even the seemingly harmless assumption in such “language independent” approaches that “a sentence is a sequence of words” doesn’t make sense in some languages.

Why data, and why is it important?  Corpora  Specialized corpora  Lexicon-like Data Resources  And many other types of language resources  Final remarks
Current trends in language resources ...

trends in the last decade

- sure, as big data as possible, as everywhere else, but apart from that also:
- multi-linguality
- under-resourced languages
- social media analysis
- discourse, dialog and interactivity
- treebanking
- evaluation methodologies
Before you start compiling your own dataset, have a glimpse into huge existing catalogues such as

- LRE Map https://lremap.elra.info/
- Linguistic Data Consortium https://www.ldc.upenn.edu/
- LINDAT/CLARIN Repository https://lindat.mff.cuni.cz/repository/

Academic data centers and commercial vendors of language data resources
• as in any other human activity, competing with others is an important source of motivation of NLP researchers

• a typical scenario

  1. an NLP task is selected (such as machine translation for a pair of languages) and defined in detail, including a performance measure specification

  2. training data are published so that participants can start developing their solutions and optimize them with respect to the data

  3. final evaluation dataset is published only for very limited time (such as previously unseen texts to be translated)

  4. participants submit their systems’ outputs for the final dataset, their solutions are evaluated and compared
Examples:

- Workshop in Machine Translation’s shared task - a yearly competition in Machine Translation
- the CoNLL series, selected tasks:
  - 2019 – Cross-Framework Meaning Representation Parsing
  - 2018 – Universal Morphological Reinflection
  - 2017 – Multilingual Parsing from Raw Text to Universal Dependencies
  - 2014 – Grammatical Error Correction
  - 2006 – Multi-Lingual Dependency Parsing
- the SemEval series, selected tasks:
  - 2018 – Affect and Creative Language in Tweets
  - 2016 – Textual Similarity and Question Answering
  - 2015 – Word Sense Disambiguation and Induction