Chapter NLP:II

II. Corpus Linguistics

Corpus Linguistics

Annotations

- An annotation marks a text or span of text as representing meta-information of a specific type.
- An annotation can also be used to specify relations between other annotations.
- The types are specified by an annotation scheme.

Topic: ”Google revenues“ **Genre:** ”News article“

Corpus Linguistics

Ground Truth vs. Automatic Annotation

Manual annotation

- The annotations of a text corpus are usually created manually.
- To assess the quality of manual annotations, inter-annotator agreement is computed based on texts annotated multiple times.

  Standard chance-corrected measures: Cohen's κ, Fleiss' κ, Krippendorff's α, ...

Ground-truth annotations

- Manual annotations are assumed to be correct; named ground truth.
- NLP approaches often learn from ground-truth annotations.

Automatic annotation

- Technically, many NLP algorithms can be seen as just adding annotations of certain types to a processed text.
- The automatic process usually aims to mimic the manual process.

Corpus Linguistics

Three Ways of Obtaining Ground-Truth Annotations

Expert annotation

- Experts (for the task, for linguistics, . . .) manually annotate each corpus text.
- Usually better results than with “laymen”, but often time-consuming and cost-intensive.

Crowd-based annotation

- Instead of experts, crowdsourcing is used to create manual annotation.
- Common platforms: `mturk.com`, `upwork.com`, . . .
- Access to many lay annotators (cheap) or semi-experts (not too cheap).
- Distant coordination overhead; results for complex tasks unreliable.

Distant supervision

- Annotations are (semi-) automatically derived from existing metadata.
- Examples: Sentiment from user ratings, entity relations from databases
- Enables large corpora, but annotations may be noisy.

Corpus Linguistics

Example: ArguAna TripAdvisor Corpus [Wachsmuth et al., 2014]

Compilation

- 2100 manually annotated hotel reviews, 300 each out of 7 locations.
- 420 each with overall user rating 1–5.
- At least 10, but as few as possible hotels per location.
- Additional 196,865 not manually-annotated reviews.

title: *great* **location**, *bad* **service**

sentiment score: 2 of 5

body: *stayed at the darling harbour holiday inn.* *The* **location** *was great, right there at China town, restaurants everywhere, the monorail station is also nearby.* *Paddy's market is like 2 mins walk.* **Rooms** *were however very small.* *We were given the 1st floor* **rooms**, *and we were right under the monorail track,* *however* **noise** *was not a problem.* **Service** *is terrible.* **Staffs** *at the* **front desk** *were impatient,* *I made an enquiry about* **internet access** *from the* **room** *and the person on the phone was rude and unhelpful.* *Very shocking and unpleasant encounter.*

Annotation

- Manual annotations. Clause-level sentiment polarity, hotel aspects.
- Distant supervision. Review-level sentiment scores from overall ratings (analog for other user ratings).

Corpus Linguistics

On Representativeness

- *"extent to which a sample includes the full range of variability in a population"* [Biber 1993]

  Here: Sample is our corpus, population is all of the language variety.

- *"A corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety."* [Leech 1991]

  Question: If we find certain features in the corpus, are we likely to find the same features in further data of that type?

- But—what is representative to the users of language?

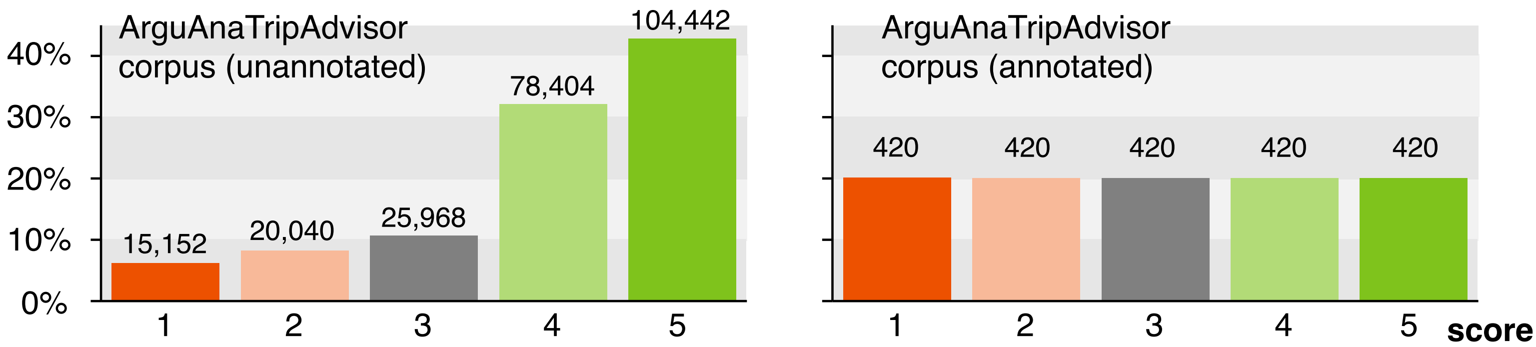
  "According to claims, the most likely document that an ordinary English citizen will cast his or her eyes over is The Sun newspaper" [Sinclair 2005]

  Keyword: reception versus production

- Corpus representativeness is important for generalization, since the corpus governs what can be learned about a given domain.

Corpus Linguistics

Representative Data versus Balanced Data



- A corpus is representative for some output information type C, if it includes the full range of variability of texts with respect to C.
- The distribution of texts over the values of C should be representative for the real distribution.
- Balance with respect to a feature means that no value/level of the feature dominates; equally distributed with respect to a feature (e.g. genre, category of linguistic phenomena).
- A balanced distribution, where all values are evenly represented, may be favorable (particularly for machine learning).