Detecting Errors in Corpus Annotation

Variation Detection in Spoken Language Treebanks
(Dickinson & Meurers 2005a)

Detmar Meurers
University of Tübingen

CLARA Thematic Training Course on Methods and Technologies
for Consolidating and Harmonising Treebank Annotation
UFAL, Charles University, Prague
December 13–16, 2010

The challenges of spoken language

- Transcribed spoken language corpora differ from written language corpora in a variety of ways, including:
  - Repetitions, false starts, and other speech errors
  - Typically shorter sentences
  - Punctuation inserted into a transcription
- Not much systematic work on syntactic annotation error analysis for spoken language corpora.

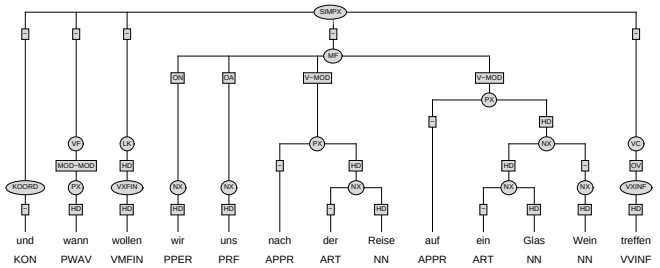
Error detection for spoken language corpora

- What is involved in applying the variation *n*-gram error detection method to a spoken language corpus?
- What insights can be gained for the annotation scheme and the method?
- For our case study, we used
  - 24,901 dialog turns (248,922 tokens) of the German Verbmobil treebank (Hinrichs et al. 2000),
  - focusing on the syntactic annotation.

The Verbmobil corpus

- Domain-specific: transcripts of appointment negotiation, travel planning, hotel reservation, and personal computer maintenance scenarios
- Annotation consists of tree structures with node and edge labels (Stegmann et al. 2000)
  - tree structure encodes:
    - topological field structure at top-level
    - syntactic categories
  - node labels encode:
    - sentence level: turn type
    - field level: topological field names
    - phrase level: syntactic categories
    - lexical level: STTS POS (Schiller, Teufel & Thielen 1995)
  - edge labels on phrase level encode:
    - grammatical functions

A simple example from the Verbmobil corpus



Two interesting aspects of the corpus

- **Repetition**: dialogues on a specific topic tend to include the same contents
  - → One encounters the same strings again and again in a corpus.
  - For example, one finds 35 instances of

    (1) , *guten Tag* , *Frau*
    , good day , Mrs.

    33 times as DM/NX and twice as NIL.
- **Hesitations and false starts**: identical words appear next to each other.

  (2) *und und Auto*
  and and car

  → Surrounding context is not informative in such cases.

Applying the variation *n*-gram approach

- We used the variation n-gram algorithm developed for discontinuous syntactic annotation (Dickinson & Meurers 2005b).
- Dialog turn boundaries are used as borders for *n*-gram expansion.
- 1426 nonfringe variation nuclei are detected
  - largest size: 14 words
- compare to 500 nuclei detected for TIGER treebank, a corpus of written text three times as large.

The nature of dialog turns

- Investigated the effect of stopping the *n*-gram search at dialog turn boundaries
  - Allowed *n*-grams to go beyond a dialog turn
  - Obtained 1720 shortest variation nuclei
  - Gain of 20% over the case where variation detection is limited to a single sentence

⇒ Repeated segments frequently go beyond one dialog turn.

The nature of punctuation

- Investigated the role of punctuation, inserted into transcribed speech of the corpus
  - Removed all punctuation from the corpus and reran the error detection code (ignoring dialog turn boundaries)
  - Obtained 1056 shortest variation nuclei
  - Loss of almost 40% of detected cases

⇒ Punctuation inserted in speech corpora provides useful context for detecting variation *n*-grams.

Punctuation ambiguity

- Punctuation symbols are not always reliable indicators of context identity.
- Commas after enumerated list element (NX underlined):

  (3) *das wäre Donnerstag , Freitag , Samstag .*
  that would be Thursday , Friday , Saturday .

- Commas used in date expressions (NX underlined):

  (4) *ab achten Mai , Freitag , den achten Mai , hätte*
  from eighth May , Friday , the eighth May , would've
  ich für vier Tage Zeit
  I for four days time

⇒ Attractive to distinguish different uses of punctuation.

Problems disambiguating locally

- In specific cases, local context is not sufficient.
- Example: *fahren* (drive) in variation 4-gram (5)

  (5) a. *wir wollten nach Hannover fahren .* (VXINF)
  we wanted to Hannover drive .

  b. *daß wir am Mittwoch und Donnerstag*
  that we on Wednesday and Thursday
  nach Hannover fahren . (VXFIN)
  to Hannover drive .

⇒ A more sophisticated notion of context for such cases?

Distinguishing levels of annotation

- The Verbmobil annotation employs different kinds of non-terminal categories:
  - sentence level: turn type
  - field level: topological field names
- Danger of comparing "apples with oranges"
- Problem surfaces frequently for unary projections, e.g,
  - NX noun phrase (e.g., as part of the Mittelfeld)
  - NF/NX for extraposed noun phrase (NF = Nachfeld)

⇒ Identify these as different representation levels, which need to be kept distinct for variation analysis

- Topological field labels also inherently non-endocentric:
  - "The C-position only occurs in verb-final clauses", but whether a clause is verb-final or not is a property of the sentence, not of the C field itself.

Summary

- The variation *n*-gram approach can be applied to spoken language corpora to detect annotation errors.
- Repetitions are prevalent in domain-specific speech, which makes method well-suited for detecting errors in such corpora.
- The role of segmentation, inserted punctuation, and the nature of repetition requires special attention.

References

Dickinson, M. & W. D. Meurers (2005a). Detecting Annotation Errors in Spoken Language Corpora. In *The Special Session on treebanks for spoken language and discourse at NODALIDA-05*. Joensuu, Finland. URL http://ling.osu.edu/~dickinso/papers/dickinson-meurers-nodalida05.html.

Dickinson, M. & W. D. Meurers (2005b). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. pp. 322–329. http://www.aclweb.org/anthology/P/P05/P05-1040.

Hinrichs, E., J. Bartels, Y. Kawata, V. Kordoni & H. Telljohann (2000). The Tübingen Treebanks for Spoken German, English, and Japanese. In W. Wahlster (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Berlin: Springer, Artificial Intelligence, pp. 552–576.

Schiller, A., S. Teufel & C. Thielen (1995). *Guidlines für das Taggen deutscher Textcorpora mit STTS*. Tech. rep., IMS-CL, Univ. Stuttgart and SfS, Univ. Tübingen. http://www.cogsci.ed.ac.uk/~simone/stts_guide.ps.gz.

Stegmann, R., H. Telljohann & E. W. Hinrichs (2000). *Stylebook for the German Treebank in VERBMOBIL*. Verbmobil-Report 239, Universität Tübingen, Tübingen, Germany. http://verbmobil.dfki.de/cgi-bin/verbmobil/htbin/decode.cgi/share/VM-depot/FTP-SERVER/vm-reports/report-239-00.ps.

Thielen, C. & A. Schiller (1996). Ein kleines und erweitertes Tagset fürs Deutsche. In H. Feldweg & E. W. Hinrichs (eds.), *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, Tübingen: Max Niemeyer Verlag, vol. 73 of *Lexicographica: Series maior*, pp. 215–226.