

Corpus Linguistics (L615)

More on annotation: reliability

Markus Dickinson

Department of Linguistics, Indiana University
Spring 2009

Annotation
scheme

Inter-annotator
reliability

Annotation error
detection

POS annotation

Syntactic annotation

Other forms of annotation

Increasing recall

Related work

References

Maintaining annotation reliability

Annotation can be difficult to annotate reliably, due to:

- ▶ the complexity of the linguistic structure
- ▶ constructions which have received little theoretical attention
- ▶ annotator fatigue and bias
- ▶ ...

How then do we ensure reliability?

The problem of errors

Linguistic searching

Annotation errors lead to low *precision* and *recall* of queries for already rare linguistic phenomena

- ▶ Meurers (2005): low precision of queries for verbal complex patterns since certain finite and non-finite verb forms are not reliably distinguished by German taggers

The problem of errors

Technology

Corpus annotation errors lead to:

- ▶ Less reliable *training* of NLP technology
 - ▶ van Halteren et al. (2001): a tagger trained on WSJ (Marcus et al. 1993) performs significantly worse than one trained on LOB (Johansson 1986)
 - ▶ Květňon and Oliva (2002) improve tagger results by cleaning the data
- ▶ Less reliable *evaluation* of NLP technology
 - ▶ van Halteren (2000): 13.6%–20.5% of the cases where the tagger disagrees with BNC annotation, the cause is an error in BNC annotation
 - ▶ Padro and Marquez (1998): because of errors in the testing data, cannot tell which of two taggers is better

This prompts researchers to develop work-around techniques to deal with noisy data (e.g., Hogan 2007)

If annotators cannot agree on how to annotate, part of the problem could be in the annotation

- ▶ Annotation scheme: could be non-well-motivated or could lack an objective right answer (e.g., information retrieval)
 - ▶ A tagset can be designed to ensure inter-annotator reliability (Voutilainen and Järvinen 1995)
- ▶ Annotation guidelines: could be underspecified, unclear, or vague
 - ▶ And it also depends upon whether the annotator is a linguistic expert or not (e.g., SUSANNE corpus annotated by an expert)

Annotation scheme

Inter-annotator reliability

Annotation error detection

- POS annotation
- Syntactic annotation
- Other forms of annotation
- Increasing recall
- Related work

References

Inter-annotator reliability

Carletta 1996

Can the annotation specified by a particular scheme be replicated by multiple annotators?

- ▶ **inter-annotator reliability**, or agreement, tests this
- ▶ n coders place linguistic units into m categories

Intuitively, we can measure something like:

- ▶ Percent of agreed-upon corpus positions
- ▶ Percent of corpus positions agreeing with an expert or a majority opinion

But we need to know how good this is, as compared to chance agreement

- ▶ Some annotation tasks are simply easier than others
- ▶ If there are 2 categories & one is chosen 95% of the time, we'd predict 90.5% agreement ($.95^2 + .05^2$)

The kappa statistic

Carletta (1999) recommends the kappa statistic to determine how different the results are from what they would be randomly

$$(1) K = \frac{P(A) - P(E)}{1 - P(E)}$$

- ▶ $P(A)$ = proportion of times these 2 annotators agree
- ▶ $P(E)$ = proportion of times we'd expect agreement by chance
 - ▶ Can be calculated differently, but based on ideas of how often each particular category is expected to be the same, e.g.:

$$(2) P(E) = p(cat_1)^2 + \dots + p(cat_n)^2$$

- ▶ We square because each annotator selects that category
- ▶ Cohen's κ adjusts this for individual annotator bias

Recommendation: $K > .8$ is good reliability and $.67 < K < .8$ allows for tentative conclusions

Other measurements

Artstein and Poesio (2008) outline a variety of measures for inter-rater reliability for computational linguistics

- ▶ It's not clear whether Carletta's K is applicable to more than 2 annotators
- ▶ Not all disagreements should be treated equally
 - ▶ e.g., for dialogue, `accept` vs. `reject` more serious than `check` vs. `info-request`

See their paper for details

Annotation error detection

To deal with errors that are already in a corpus, we can search for errors

- ▶ i.e., since I have you in my clutches, I'll tell you all about my research ...

Core idea: Instead of comparing the annotation for the same corpus position, compare the annotation for different corpus positions (cf. also Wallis 2003)

- ▶ ... but ones for which you would expect consistent annotation

The variation n -gram method

The variation n -gram method for detecting annotation errors (Dickinson and Meurers 2003a,b, 2005; Dickinson 2005):

- ▶ Finds recurring data and compares analyses in different corpus instances
- ▶ Uses shared context as a heuristic to determine when analyses should be annotated identically

Question: What corpus information can accurately distinguishes erroneous variation from legitimate ambiguity?

Corpus Linguistics

More on
annotation:
reliability

Annotation
scheme

Inter-annotator
reliability

Annotation error
detection

POS annotation

Syntactic annotation

Other forms of annotation

Increasing recall

Related work

References

Variation n-grams for POS annotation

Dickinson and Meurers (2003a)

Variation: material occurs multiple times in corpus with different annotations

Dickinson and Meurers (2003a) introduces the notions

- ▶ *variation nucleus*: recurring word with different annotation
- ▶ *variation n-gram*: variation nucleus with identical context

and provides an efficient algorithm to compute them.

Example: 12-gram with variation nucleus *off*

(3) to ward **off** a hostile takeover attempt by two European shipping concerns

In the two occurrences of this 12-gram in the WSJ, *off* is

- ▶ once annotated as a preposition (IN), and
- ▶ once as a particle (RP).

Heuristics for disambiguation

Variation can result from:

- ▶ *ambiguity*: different possible labels occur in different corpus occurrences
- ▶ *error*: labeling of a string is inconsistent across comparable occurrences

Non-fringe heuristic to detect annotation errors:

- ▶ Nuclei found at fringe of n -gram more likely to be genuine ambiguities (Dickinson 2005)
 - ▶ Natural languages favor the use of local dependencies over non-local ones

Heuristic is independent of a specific corpus, annotation scheme, or language, and receives support from:

- ▶ human category acquisition (cf. Mintz 2003)
- ▶ grammar induction (cf. Klein and Manning 2002)

Error detection for syntactic annotation

Dickinson and Meurers (2003b)

For syntactic annotation, decompose variation nucleus detection into series of runs for all relevant string lengths

- ▶ one-to-one mapping: string → syntactic category label (or special label NIL=non-constituent)
- ▶ perform runs for strings from length 1 to longest constituent in corpus

After removing nuclei of a trace of a long distance dependency, this method gives 75.86% precision for the 3,619 shortest non-fringe variation nuclei found in the WSJ corpus.

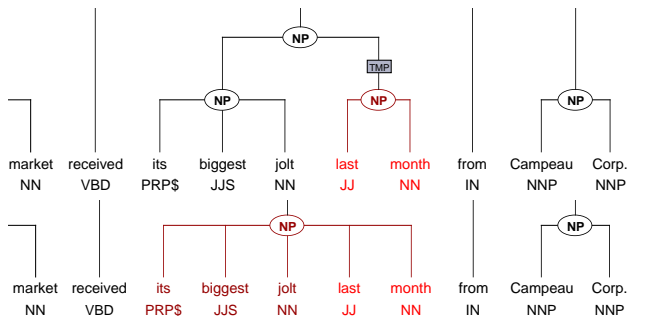
Examples from the WSJ corpus

- Variation between two syntactic category labels:

(4) maturity **next Tuesday**

labeled as **NP** twice
PP once

- Variation between constituent and non-constituent (NIL):



Other forms of annotation

- ▶ Dependency annotation: the method can be extended to working with dependency pairs (Boyd et al. 2008)
 - ▶ Relies on the method for discontinuous constituents (Dickinson and Meurers 2005)
- ▶ Semantic role annotation: the method works fairly well for semantic role labeling errors (Dickinson and Lee 2008)
 - ▶ Incidentally, Dickinson and Lee (2008) shows that building (semantic) annotation on top of erroneous (POS/syntactic) annotation leads to problems

Increasing recall

You can increase the recall of the method (i.e., detect more errors) by:

- ▶ Generalizing the context to, e.g., POS tags (Dickinson 2005)
- ▶ More accurately identifying which parts of the context are or are not relevant
 - ▶ e.g., for semantic annotation, context is important around a verbal argument, but not as much around the verb itself (Dickinson and Lee 2008)
- ▶ Generalizing the definition of a nucleus
 - ▶ For syntactic annotation, can use POS nuclei (Boyd et al. 2007)
 - ▶ For POS annotation, can use ambiguity class nuclei (Dickinson 2008b)

Related work

General error detection methods

- ▶ Eskin (2000) uses anomaly detection, flagging 7055 anomalies for the Penn Treebank, about 44% of which are errors
- ▶ Květón and Oliva (2002) search for invalid bigrams of POS tags
- ▶ Hirakawa et al. (2000) & Müller and Ule (2002) use POS annotation as input for syntactic processing and identify sentences with unexpected results

Related work

Approaches based on inconsistencies

- ▶ Look for mismatches between layers of related annotation
 - ▶ Babko-Malaya et al. (2006) find discrepancies between syntactic and semantic layers of annotation
- ▶ Look for inconsistencies in the “grammar” extracted from a corpus
 - ▶ e.g., compare edit distance between rules (Dickinson 2008a; Dickinson and Foster 2009)
 - ▶ Or, build a treebank in parallel with a grammar (e.g., Bond et al. 2004; Oepen et al. 2004)

References

- Artstein, Ron and Massimo Poesio (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4), 555–596.
- Babko-Malaya, Olga, Ann Bies, Ann Taylor, Szuting Yi, Martha Palmer, Mitch Marcus, Seth Kulick and Libin Shen (2006). Issues in Synchronizing the English Treebank and PropBank. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*. Sydney, pp. 70–77.
- Bond, Francis, Sanae Fujita et al. (2004). The Hinoki Treebank: Toward Text Understanding. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC-04)*. Geneva, pp. 7–10.
- Boyd, Adriane, Markus Dickinson and Detmar Meurers (2007). Increasing the Recall of Corpus Annotation Error Detection. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT 2007)*. Bergen, Norway, pp. 19–30.
- Boyd, Adriane, Markus Dickinson and Detmar Meurers (2008). On Detecting Errors in Dependency Treebanks. *Research on Language and Computation* 6(2), 113–137.
- Dickinson, Markus (2005). Error detection and correction in annotated corpora. Ph.D. thesis, The Ohio State University.
- Dickinson, Markus (2008a). Ad Hoc Treebank Structures. In *The 46th Annual Meeting of the Association for Computational Linguistics (ACL) with the Human Language Technology Conference (HLT) (ACL-08)*. Columbus, OH, pp. 362–370.

- Dickinson, Markus (2008b). Representations for category disambiguation. In *The 22nd International Conference on Computational Linguistics (COLING-08)*. Manchester, pp. 201–208.
- Dickinson, Markus and Jennifer Foster (2009). Similarity Rules! Exploring Methods for Ad-Hoc Rule Detection. In *Proceedings of the Seventh Workshop on Treebanks and Linguistic Theories (TLT-7)*. Groningen, The Netherlands.
- Dickinson, Markus and Chong Min Lee (2008). Detecting Errors in Semantic Annotation. In *Proceedings of LREC 2008*. Marrakech, Morocco.
- Dickinson, Markus and W. Detmar Meurers (2003a). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of EACL-03*. Budapest, pp. 107–114.
- Dickinson, Markus and W. Detmar Meurers (2003b). Detecting Inconsistencies in Treebanks. In *Proceedings of TLT-03*. Växjö, Sweden, pp. 45–56.
- Dickinson, Markus and W. Detmar Meurers (2005). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of ACL-05*. pp. 322–329.
- Eskin, Eleazar (2000). Automatic Corpus Correction with Anomaly Detection. In *Proceedings of NAACL-00*. Seattle, Washington.
<http://www.cs.columbia.edu/~eeskin/papers/treebank-anomaly-naacl00.ps>.
- Hirakawa, Hideki, Kenji Ono and Yumiko Yoshimura (2000). Automatic Refinement of a POS Tagger Using a Reliable Parser and Plain Text Corpora. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*. ICCL, Saarbrücken, Germany.
- Hogan, Deirdre (2007). Coordinate Noun Phrase Disambiguation in a Generative Parsing Model. In *Proceedings of ACL-07*. Prague, Czech Republic, pp. 680–687.

- Johansson, Stig (1986). *The Tagged LOB Corpus: Users' Manual*. Norwegian Computing Centre for the Humanities, Bergen.
- Klein, Dan and Christopher D. Manning (2002). A Generative Constituent-Context Model for Improved Grammar Induction. In *Proceedings of ACL-02*. Philadelphia, PA.
- Květoň, Pavel and Karel Oliva (2002). Achieving an Almost Correct PoS-Tagged Corpus. In Petr Sojka, Ivan Kopeček and Karel Pala (eds.), *Proceedings of TSD-02*. Heidelberg: Springer, no. 2448 in Lecture Notes in Artificial Intelligence (LNAI), pp. 19–26.
- Marcus, M., Beatrice Santorini and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.
<ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz>.
- Meurers, Walt Detmar (2005). On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua* 115(11), 1619–1639. <http://ling.osu.edu/~dm/papers/meurers-03.html>.
- Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90, 91–117.
- Müller, Frank H. and Tylman Ule (2002). Annotating topological fields and chunks – and revising POS tags at the same time. In *Proceedings of COLING*. <http://ling.osu.edu/~dm/02/spring/795K/mueller-ule.ps>.
- Oepen, Stephan, Dan Flickinger and Francis Bond (2004). Towards holistic grammar engineering and testing—grafting treebank maintenance into the grammar revision cycle. In *Beyond Shallow Analyses—Formalisms and Statistical Modelling for Deep Analysis (Workshop at The First International Joint Conference on Natural Language Processing (IJCNLP-04))*. Hainan, China.

- Padro, Lluís and Lluís Marquez (1998). On the Evaluation and Comparison of Taggers: the Effect of Noise in Testing Corpora. In *Proceedings of ACL/COLING-98*. San Francisco, California, pp. 997–1002.
- van Halteren, Hans (2000). The Detection of Inconsistency in Manually Tagged Text. In Anne Abeillé, Thorsten Brants and Hans Uszkoreit (eds.), *Proceedings of LINC-00*. Luxembourg. Workshop information at <http://www.coli.uni-sb.de/linc2000/>.
- van Halteren, Hans, Walter Daelemans and Jakub Zavrel (2001). Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational Linguistics* 27(2), 199–229.
- Voutilainen, Aaro and Timo Järvinen (1995). Specifying a shallow grammatical representation for parsing purposes. In *Proceedings of the 7th Conference of the EACL*. Dublin, Ireland. <http://aclweb.org/anthology/E95-1029>.
- Wallis, Sean (2003). Completing Parsed Corpora. In Anne Abeillé (ed.), *Treebanks: Building and using syntactically annotated corpora*, Dordrecht: Kluwer, pp. 51–71.