Part-of-Speech in Historical Corpora: Comparative Tagger Evaluation on ARCHER

University of Zurich[UZH]

Gerold Schneider and Marianne Hundt
{gschneid,m.hundt}@es.uzh.ch

ICEHL 2014, Leuven

Introduction

TAGGER ACCURACY deteriorates when applied to texts different from the training corpus, e.g. with respect to register or time period (e.g. Rayson et al. (2007)). Typically, taggers are trained on PDE news texts (e.g. the Penn Treebank). They typically reach 95-97% accuracy on PDE texts of the same register given that tokenization is perfect. If these conditions are not met, accuracy can fall to and below 90%.

In our project, we are tagging and parsing ARCHER (Biber, Finegan, and Atkinson, 1994), a historical corpus sampled from British and American texts from 1600-1999 and across several registers. Its current version (V 3.2) contains 3.2 mio words. We improve tagging accuracy by

- using a version that has been automatically mapped to PDE spelling with VARD (see Schneider, Lehmann, and Schneider (2014))
- combining several part-of-speech taggers.

Methodology

DIFFERENT TAGGERS make different mistakes, as they use different algorithms, tags and partly different training sets.
→ different perspective on same data. We use the following taggers.

- **Tree-Tagger** (Schmid, 1994): a decision-tree tagger, also reports probabilities and n-best tags

  It_PRP adds_VBZ much_JJ/RB to_TO my_PRP$ satisfaction_NN ,_, that_IN her_PRP$ Character_NNP is_VBZ agreeable_JJ to_TO your_PRP$ Fancy_NNP

- **CLAWS** (Leech, Garside, and Bryant, 1994; Garside and Smith, 1997): hybrid, combining probabilistic and rule-based. It also reports probabilities and n-best tags. We map the original CLAWS4 automatically to the Penn tagset. This offers us an additional alternative perspective, and possibly reduced errors in some areas due to the increased granularity of the original CLAWS tagset.

  It_PRP adds_VBZ much_RB/DT to_IN my_PRP$ satisfaction_NN ,_, that_IN her_PRP$ Character_NN is_VBZ agreeable_JJ to_IN your_PRP$ Fancy_NN

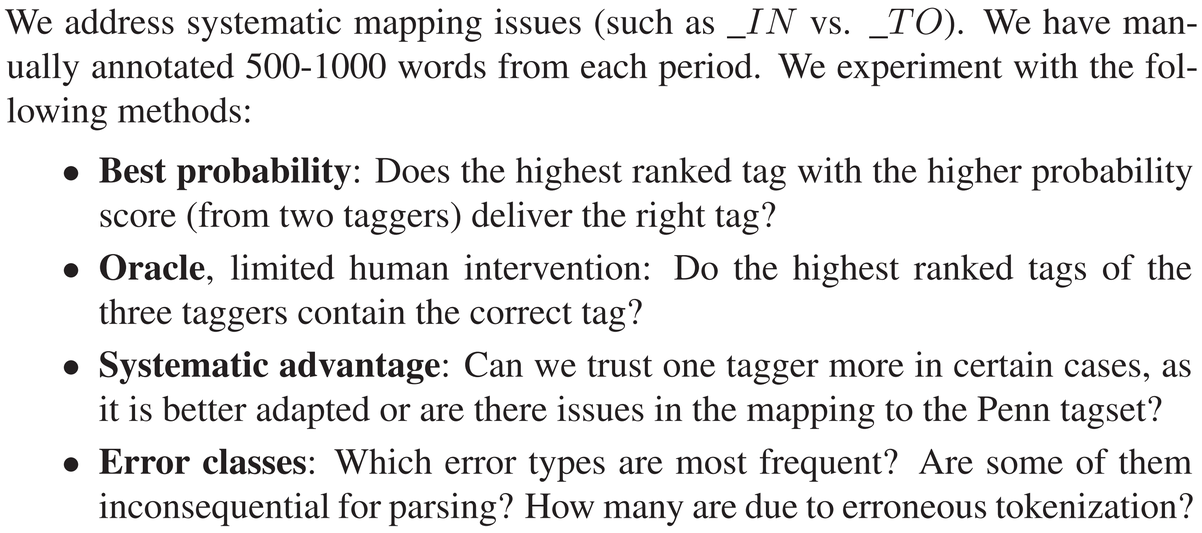
- **CandC** (Curran, Clark, and Bos, 2007; Grover, 2008): a maximum-entropy tagger, which is part of the XML pipeline LT-TTT2, which we use. As it performed slightly worse, we excluded it from most experiments.

We address systematic mapping issues (such as $\_IN$ vs. $\_TO$). We have manually annotated 500-1000 words from each period. We experiment with the following methods:

- **Best probability**: Does the highest ranked tag with the higher probability score (from two taggers) deliver the right tag?
- **Oracle**, limited human intervention: Do the highest ranked tags of the three taggers contain the correct tag?
- **Systematic advantage**: Can we trust one tagger more in certain cases, as it is better adapted or are there issues in the mapping to the Penn tagset?
- **Error classes**: Which error types are most frequent? Are some of them inconsequential for parsing? How many are due to erroneous tokenization?

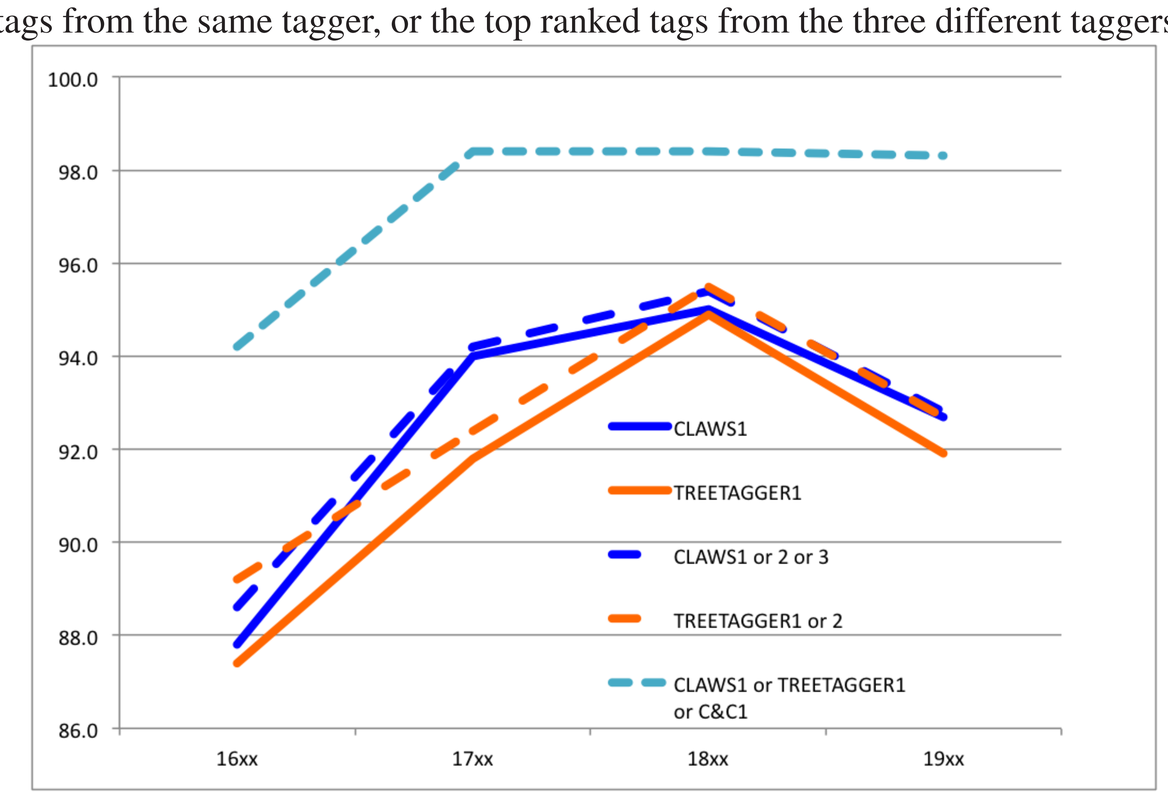
Results: Automatic Combination

PROBABILITIES FOR THE MOST LIKELY TAGS are delivered by CLAWS and Tree-Tagger. They can be interpreted as confidence scores. If we always choose the tag whose confidence score is highest from these two taggers, we can automatically boost performance by 0.79% on average.

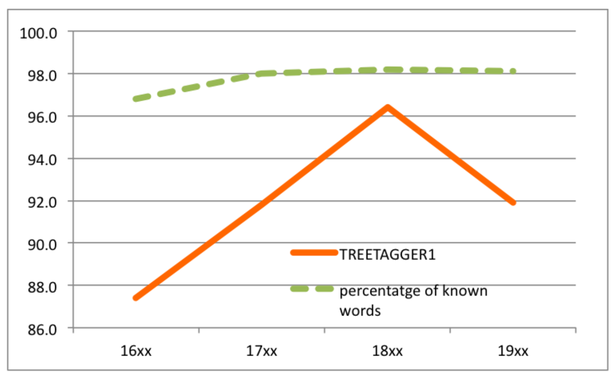


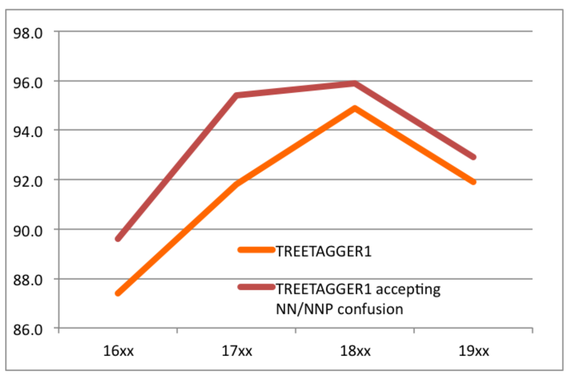
Results: Semi-Automatic Combinations

WITH LIMITED HUMAN INTERVENTION performance can be further improved if a human chooses either one of the maximally 3 promising n-best tags from the same tagger, or the top ranked tags from the three different taggers.



Major reasons for errors include: unknown words (left) and proper vs. common noun (right, $\_NNP$ vs. $\_NN$); the latter error is mostly due to different capitalisation practice in previous periods (not normalised during VARDing), e.g.:
He had been very restless all Night, his Pulse irregular, his Tongue rough and dry, with Flushings in his Cheeks. (ARCHER 1735gool_m3b)





Conclusions

- Careful mapping to PDE spelling with VARD allows one to achieve PDE accuracy levels from about 1700.
- Automatically combining 2 taggers with sufficiently different approaches improves tagging performance by 0.79% on average.
- Limited human intervention (choosing 1 of max. 3) improves tagging accuracy by additional 2-5%, reaching above 98% on texts after about 1700.
- The hybrid (partly rule-based) CLAWS tagger performs considerably better on historical texts. It possibly profits from a more fine-grained tagset (to be verified).
- Surprisingly, 19^{th} century texts can be easier to tag than PDE, partly due to correlation between unknown words and tagging performance.
- We need to annotate more text to control for register variation.

Acknowledgements

THIS PROJECT is partially supported by the Zurich Center for Linguistics (http://www.linguistik.uzh.ch). We are grateful to our student assistants Henning Beywl and Rahel Oppliger for programming and evaluation tasks, and to Michael Percillier and Christian Mair for providing the automatically VARDed version of ARCHER-3.2.

References

Biber, Douglas, Edward Finegan, and Dwight Atkinson. 1994. Archer and its challenges: Compiling and exploring a representative corpus of historical English registers. In Udo Fries, Peter Schneider, and Gunnel Tottie, editors, *Creating and using English language corpora, Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993*. Rodopi, Amsterdam, pages 1–13.

Curran, James, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic, June. Association for Computational Linguistics.

Garside, Roger and Nicholas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In Roger Garside, Geoffrey Leech, and Tony McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pages 102–121.

Grover, Claire. 2008. LT-TTT2 example pipelines documentation. Technical report, Edinburgh Language Technology Group,.

Leech, Geoffrey, Roger Garside, and Michael Bryant. 1994. CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, pages 622 – 628, Kyoto, Japan.

Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics 2007*. University of Birmingham, UK.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Schneider, Gerold, Hans Martin Lehmann, and Peter Schneider. 2014. Parsing Early Modern English corpora. *Literary and Linguistic Computing*, first published online February 6, 2014 doi:10.1093/llc/fqu001.