Part-Of-Speech in Historical Corpora: Tagger Evaluation and Ensemble Systems on ARCHER

Gerold Schneider
University of Konstanz
and University of Zurich
gschneid@es.uzh.ch

Marianne Hundt
University of Zurich
m.hundt@es.uzh.ch

Rahel Oppliger
University of Zurich
rahel.oppliger@uzh.ch

Abstract

Tagger accuracy deteriorates when applied to texts different from the training corpus, e.g. with respect to register or time period. On historical data, accuracy can drop to and below 90%. We are tagging and parsing ARCHER, a historical corpus sampled from British and American texts from 1600-1999. We improve tagging accuracy by (1) using a version of the corpus that has been automatically mapped to PDE spelling with VARD, (2) by combining several part-of-speech taggers in an ensemble system – which improves tagging by about 1% over CLAWS and 2% over Tree-Tagger, and (3) by using a small amount of human intervention – which allows us to reach 98% accuracy from 1700 on.

1 Introduction

Part-of-speech tagging accuracy strongly deteriorates when a tagger is applied to texts which are different from the training domain. Typically, taggers are trained on present-day English (PDE) texts, specifically news texts, mostly from the Penn Treebank (Marcus et al., 1993). They then reach 95-97% accuracy on PDE texts of the same register given that tokenisation is perfect. If these conditions are not met, accuracy can drop to and below 90%. For example, Rayson et al. (2007) report that the CLAWS tagger (Garside and Smith, 1997) achieves 96 to 97% accuracy on PDE, while on Early Modern English, performance drops to 81.9% on Shakespeare texts and to 88.5% on pamphlets from the Lampeter corpus.

A major source for errors are historical spelling variants. There are two possible strategies for dealing with spelling variants: either the tagger is adapted to cope with the variant directly, or the spelling variants are normalised to PDE forms, as expected by the tagger. We have chosen the second option.

2 Data and Methods

2.1 The ARCHER Corpus

As corpus of application, we chose ARCHER (Biber et al., 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 million words. We improve tagging accuracy by using a version of the corpus that has been automatically mapped to PDE spelling with VARD, by combining several part-of-speech taggers in an ensemble system, and by using a small amount of human intervention.

2.2 Spelling Normalisation

A major source for errors are historical spelling variants. Simple variants like *call'd* for *called* typically result in wrong tagging (*call_NN d_MD*), in this case triggered by a tokenisation error, and as a consequence parsing quality is also affected.

There are two possible strategies for dealing with spelling variants. In the first strategy, the tagger is adapted to cope with the variant directly. Yang and Eisenstein (2016) present an approach using domain adaptation which has very high accuracy. They also argue that their approach circumvents the partly ill-defined task of normalisation. In the second strategy, the spelling variants are normalised to PDE forms, as expected by the tagger. We have chosen the second option. Compared to domain adapation, normalising approaches have the advantage that they allow linguists to search for all occurrences of a word form, with a single and obvious query. Spelling normalisation, according to Rayson et al. (2007), increases tagging accuracy to 85% for the Shakespeare texts, and to 89% for the Lampeter texts, when using the automatic normalisation tool VARD (Baron and Rayson, 2008). They also give an upper bound of their approach by using manual

Tree-Tagger: 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 Tagger: 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 Tagger: It_PRP adds_VBZ much_**RB** to_**TO** my_PRP$ satisfaction_NN ,_, that_IN her_PRP$ Character_**NNP** is_VBZ agreeable_JJ to_**TO** your_PRP$ Fancy_**NNP**

Table 1: Sample outputs from Tree-Tagger, CLAWS tagger and CandC (ARCHER 1671cary_d2b)
PENN tags: JJ=adjective, RB=adverb, DT=determiner, TO='to', IN=preposition, NN=common noun, NNP=proper name

normalisation: 89% for Shakespeare, and 91% for Lampeter. In other words, about half of the tagging errors could be corrected.

2.3 Fully Automatic Ensemble System

Different taggers make different mistakes, as they use different algorithms, tags, and partly different training sets. They thus offer different perspectives on same data. Combinations of different systems, which are also called ensemble systems, can benefit from their mutual advantages, as long as the individual participating systems are quite accurate and diverse (Dietterich, 1997; van Halteren et al., 2001) We use the following three taggers: Tree-Tagger, CLAWS, CandC. They are presented briefly in the following.

The **Tree-Tagger** (Schmid, 1994) is a decision-tree tagger. In additon to the most likely tag, it also offers n-best tagging as an option. N-best tagging returns the *n* most likely tags, together with an estimate of the probability of each tag, given the language model.

The **CLAWS tagger** (Leech et al., 1994; Garside and Smith, 1997) is a hybrid system which combines probabilistic and rule-based approaches. Like the Tree-Tagger, it also reports n-best tags including probabilities. We map the original CLAWS5 tagset automatically to the Penn tagset. The mapping list is for example given in Wu (2010, 97). The CLAWS5 tagset comprises of 62 tags and is thus more fine-grained than the Penn Treebank tagset with its 39 tags. Mapping from CLAWS5 to the Penn tagset is mostly deterministic, and depends on the tag only. There are exceptions, though, the most notable being the fact that CLAWS5 disambiguates between *to* as infinitive particle and as preposition, while Penn gives the tag *TO* to both. We count both tags as correct in our evaluation.

The fact that CLAWS uses a different tagset offers an additional alternative perspective to us. While a larger tagset leads to a lower baseline and has the risk that the tagger needs to take potentially more difficult decisions, this potential disadvantage should disappear if a reliable mapping procedure to the more coarse-grained tagset is used. In fact, a larger tagset can also facililate the task: if particular features strongly point to a rare tag, the accuracy of recognition can in fact increase.

The **CandC tagger** (Curran et al., 2007; Grover, 2008) is a maximum-entropy tagger, as it distributed as part of the XML pipeline LT-TTT2[1].

Table 1 gives an example of the outputs by the three taggers. The differring parts are highlighted. Double-tags are given if the tagger in n-best mode outputs several tags. The tag closer to the word is the higher ranked tag. We see in this sentence that except for CLAWS, taggers tend to asssign proper name (NNP) to capitalised words. We also see that CLAWS aims to distinguish between *to* as preposition and infinitive particle.

After comparing the accuracies for each tagger in section 3.1, we show in section 3.2 that a fully automatic ensemble approach increases the accuracy. We experiment with the following methods:

Majority voting Majority voting checks if 2 of our 3 taggers agree. If they do, the majority tag is selected.

Best probability Best probability compares the probabilites that the two n-best taggers (Tree-tagger and CLAWS) return. The probabilities can be interpreted as scores, as an estimation of the tagger's confidence in its decision. The tag with the highest probability score is selected, giving precedence to the tagger which has higher confidence in its decision.

Systematic advantage It is also possible that one tagger can be trusted more, either generally or in specific cases, as it may be better adapted. For example, CLAWS correctly tags *hath* and *hast* as verb.

2.4 Semi-Automatic: Limited Human Intervention

In section 3.2 we apply methods that need limited human intervention. In these approaches, a human

[1] https://www.ltg.ed.ac.uk/software/lt-ttt2/

Tagger	Tree-Tagger	CLAWS	CandC	Best Ensemble	Error Rate Reduction	Best Oracle
16xx	87.4	87.8	82.8	88.8 (+1.0)	8.9	94.2
17xx	91.0	93.2	85.4	93.4 (+0.2)	3.0	98.2
18xx	95.2	95.0	91.8	95.6 (+0.4)	13.6	98.2
19xx	92.1	92.8	86.2	94.1 (+1.3)	22.0	98.3

Table 2: Accuracy (percent) of individual taggers and best combinations, split by century

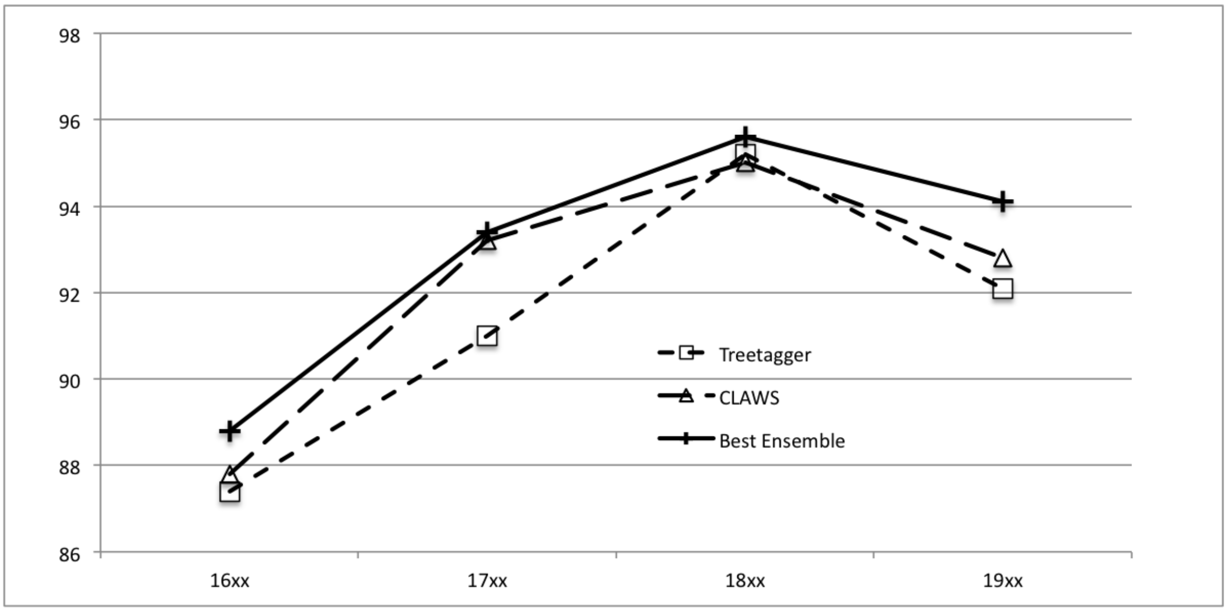

Figure 1: Tagging performance of Tree-Tagger, CLAWS, and the ensemble system

needs to chose between one of maximally three candidates. We use the following two methods. First, a tagger-internal choice: if a tagger offers several tags in n-best mode, is one of them correct? Second, the highest ranked tag suggested by each tagger is considered: if the taggers disagree, does one of them suggest the correct tag? These approaches can also be described as Oracle approaches which measure the upper bound of the taggers.

3 Results

3.1 Individual Taggers

We split the corpus into four periods – each comprising one century – and manually annotated at least 500 words from each period. The manual annotations were cross-checked by two authors and discussed until an agreement could be reached. The accuracy of each tagger is given in Table 2, columns 2 to 4. CLAWS is on average 0.78% better than Tree-Tagger. As the performance of CandC was considerably worse, we excluded it from most ensemble experiments, which we explain in the following.

3.2 Automatic Combinations

Probabilities for the most likely tags are delivered by CLAWS and Tree-Tagger in n-best tagging mode. The probabilities can be interpreted as confidence scores. If we always choose the tag whose confidence score is highest from these two taggers (Best Ensemble), we can automatically increase performance by 0.78% on average over the better performing tagger, CLAWS, as Figure 1 shows. The increase over Tree-Tagger alone is between 1 and 2 percent. The exact percentages are given in Table 2, column 4. In terms of error rate reduction, between 3% and 22% of the errors could be corrected by the best Ensemble approach, as column 5 shows.

We have also tried majority voting, but due to the relatively low performance of the CandC tagger the combined performance stays below CLAWS as single tagger.

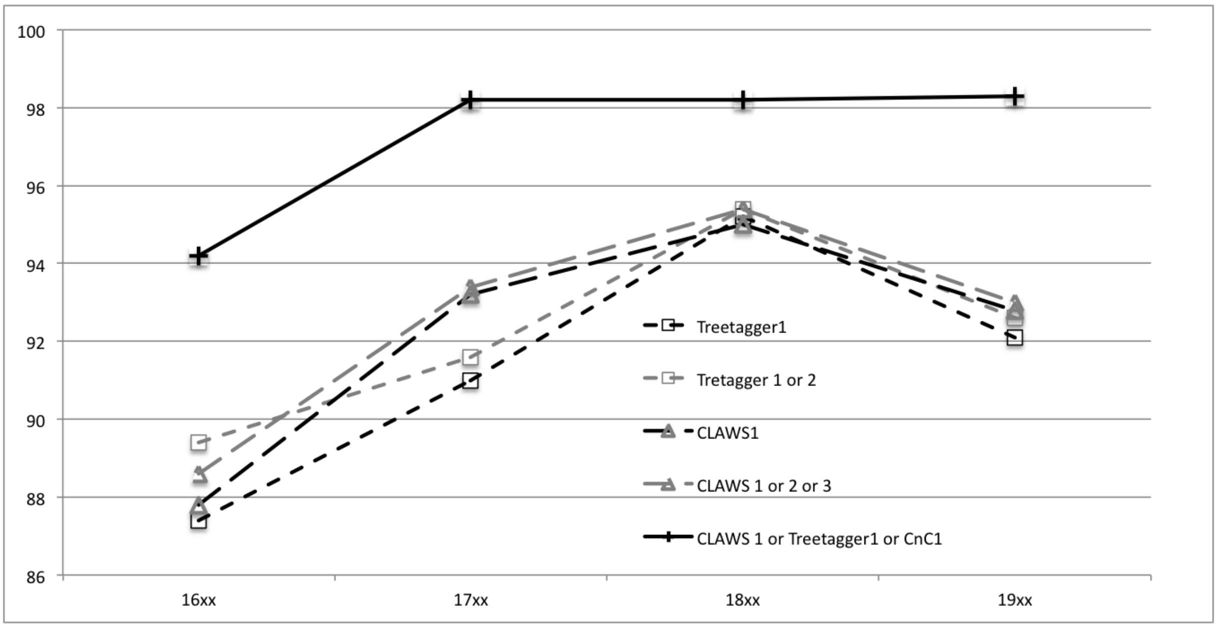

Figure 2: Performance with limited human intervention: choose one of three in ambiguous cases

3.3 Semi-Automatic Combinations

With limited human intervention, performance can be further improved if a human chooses either one of the maximally 3 most promising n-best tags from the same tagger, or the top tag from the three different taggers. Figure 2 shows the results of both approaches. Choosing between several options in n-best mode increases performance only slightly. A major reason for the modest improvement is that alternative tags are only suggested for a small minority of all word tokens: about 5% in the earliest texts, and about 2% in the 20^{th} century.

The second option – manually selecting the top tag if the taggers disagree – leads to a strong improvement, by 2-5 %, to above 98% except in the $17t^{th}$ century, as Table 2, last column, shows. Disagreement between taggers is quite frequent though: in the 20^{th} century, all three taggers suggest the same tag in 423 out of the 529 words in the evaluation set; in 106 cases (20% of all words) the user needs to select the correct tag. In terms of entropy, we can observe that on average, there are 1.31 tags to choose from per word. Split by century, there are 1.36 tags per word in the 16xx texts, 1.31 in the 17xx texts, 1.23 in the 18xx texts, and 1.33 in the 19xx texts.

The fact that the value is lowest for 18xx and not 19xx indicates that the texts from the 20^{th} century are in fact harder to tag for the tagger model than those from the 19^{th} century, which we discuss in the following.

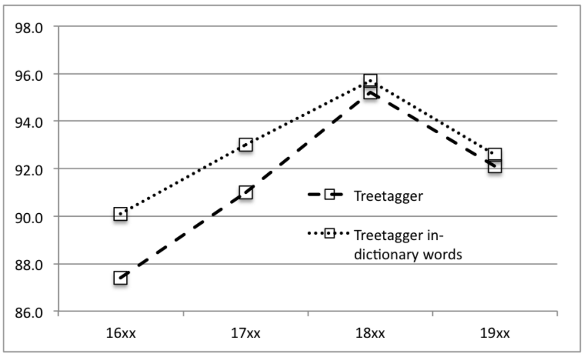

Figure 3: Influence of unknown words: tagging accuracy of the Tree-tagger on known words, and on all words

4 Discussion

4.1 Dropping Accuracy in the 20^{th} Century

One of the most surprising outcomes of our experiments was the fact that all taggers had lower accuracy on the 20^{th} century texts than on the 19^{th} century texts. One possible explanation is that this could be a random fluctuation caused by genre variation, for which we did not control. We extended our random sample and annnotated further 20^{th} century texts, but the performance did not change significantly. In future research, we will use an evaluation set that is stratified by genre. A second, more likely explanation is that some linguistic features of the 20^{th} century are harder to process. An important feature is the strong growth in vocabulary, for

Tree-Tagger Confusion	16xx	17xx	18xx	19xx	TOTAL
NN / NNP	8	14	1	6	29
VB / VBP	4	0	2	3	9
VBP / VB	3	1	1	3	8
VB / NN	1	0	4	1	6
VBD / VBN	2	1	1	2	6
JJ / VBN	0	0	3	2	5
JJ / NNP	0	2	0	3	5
VBD / NNP	0	2	0	3	5
RB / IN	3	1	0	1	5
NNS / NNP	1	4	0	0	5
RB / JJ	1	1	1	2	5
FW / NNS	4	0	0	0	4
NN / NNS	0	2	1	0	3
DT / NN	3	0	0	0	3
FW / NN	3	0	0	0	3
RB / NNP	1	0	0	2	3
VBG / NN	2	0	0	1	3
VBP / NN	2	0	0	1	3

Table 3: Most frequent tag confusions by the Tree-Tagger

CLAWS Tagger Confusion	16xx	17xx	18xx	19xx	TOTAL
NNP / NN	5	5	3	8	21
VB / VBP	8	2	4	2	16
DT / JJS	2	2	2	0	6
WP / WDT	2	0	1	2	5
CD / NN	4	0	0	0	4
DT / PRP	4	0	0	0	4
JJS / JJR	2	0	2	0	4
VBD / VBN	2	0	0	2	4
NN / VBP	1	2	1	0	4
WRB / IN	1	0	0	3	4
DT / JJS	0	2	2	0	4
NN / IN	0	2	0	2	4
FW / NN	3	0	0	0	3
FW / NNS	3	0	0	0	3
IN / RP	3	0	0	0	3
RB / IN	3	0	0	0	3
IN / RB	2	1	0	0	3
VBG / NN	2	0	0	1	3
JJ / NN	1	0	0	2	3
JJ / RB	1	2	0	0	3
NN / NNP	1	1	0	1	3
VBD / JJ	0	1	0	2	3
NNP / JJ	0	1	0	2	3

Table 4: Most frequent tag confusions by the CLAWS tagger

example the use of abbreviations. The Tree-tagger optionally marks out-of-vocabulary words. There are more out-of-vocabulary words in the 20^{th} century texts than in the 19^{th} century. Per century, the percentages of unknown words are: 5.2% in 16xx, 2.8% in 17xx, 2.2% in 18xx, and 3.0% in 19xx. While a higher amount of unknown words affects tagging accuracy, but also the accuracy of words that are known to the tagger decreases in 19xx, as Figure 3 shows. Out-of-vocabulary can thus only serve as a partial explanation.

We also noted that the 20^{th} century texts contain considerably more features which are particularly frequent in social media, for example telegram style and spoken features like contractions. Some do not contain apostrophes (e.g. *youre* instead of *you're*), which almost inevitably lead to tagging errors. Another feature are compressed and complex NPs. Two examples of sentences containing these features, and the relevant tags assigned by the CLAWS tagger are given in (1) and (2).

> *(1) Saturday 10 24 - A._NN Boiled_JJ sap_VBP this P.M. are having another good run of sap .* (ARCHER 1920rich_y7a_s193)

> *(2) Specify Regal_JJ Mk V for 1960 Reliant_JJ 's Silver Jubilee year .* (ARCHER 1960illn_a8b_s102)

4.2 Error Analysis

Error classes We have conducted an error analysis, to check which types of tagging error are particularly frequent, to find out if causes can be isolated, and if some tagging errors are more serious than others.

The most frequent types of confusion of the Tree-tagger, i.e. all errors that occur at least 3 times, are given in Table 3. The equivalent figures for the CLAWS tagger can be seen in Table 4. The most prominent cause of error is different capitalisation practice in previous periods. It needs to be pointed out that capitalisation is not normalised by VARD. An example of a sentence in which nouns are generally capitalised is given in (3).

> *(3) He had been very restless all Night, his Pulse irregular, his Tongue rough and dry, with Flushings in his Cheeks.* (ARCHER 1735gool_m3b_s59)

While the Tree-Tagger tends to assign proper noun (*NNP*) to capitalized common nouns (*NN*) too often, the CLAWS tagger shows the opposite trend to overgeneralise *NN* to too many *NNP*s. An example is given in (4), the words in bold are incorrectly assigned common noun tags by CLAWS.

> *(4) Recently Whiting developed the* **Bus** *and* **Car Washer** *, shown above , which shampoos a bus from end to end in only 45 seconds ...* (ARCHER 1942news_a7a_s132)

The second most frequent error is a confusion between infinite verb and inflected verb in the present. Due to the considerably freer word order in the earlier texts, material intervening between the auxiliary verb and the main verb frequently leads to situations in which the tagger's observation window is too small. An example which includes two errors of this type is given in (5), where the tagger assigned non-third person present tense (*VBP*) instead of nonfinite form (*VB*).

> *(5) ... whereas quite contrary they will without the least opposition* **permit** *the first , but with the greatest difficulty* **admit** *of the last .* (ARCHER 1665head_f2b_s24)

The confusions involving the tag *FW* (foreign word) involve French and Latin expressions, which are more frequent in earlier texts. It is difficult to see further clear trends in the tagger confusion tables. The larger amount of unknown words in earlier texts and in the 20^{th} century typically leads to unspecific, context-dependent errors. Most of the remaining errors are too sparse in our small evaluation set to show clear trends or a sigificant decrease in PDE.

4.3 Underspecifiying Nouns

The most frequent tagger confusion, the one between common noun and proper name, is due to the fact that the distinction between common noun and proper name is particularly hard to make, because often the majortiy of nouns are capitalised in the earlier texts (see example 3), and it can also be argued that it is possibly inconsequential for the subsequent step of syntactic parsing.

We have therefore considered an evaluation variant in which the distinction between common noun (*NN(S)*) and proper name (*NNP(S)*) is not made.

Tagger Combination	16xx	17xx	18xx	19xx
Tree-Tagger	87.4	91.0	95.2	92.1
Tree-Tagger NN/NNP	89.0	93.8	95.4	93.6
CLAWS	87.8	93.2	95.0	92.8
CLAWS NN/NNP	88.8	94.4	95.4	94.5
Best Ensemble	88.8	93.4	95.6	94.1
Best Ensemble NN/NNP	89.6	94.4	96.0	95.8
Oracle	94.2	98.2	98.2	98.3
Oracle NN/NNP	94.4	98.4	98.4	98.7

Table 5: Accuracy of taggers if common noun (*NN*) and proper name (*NNP*) are not distinguished

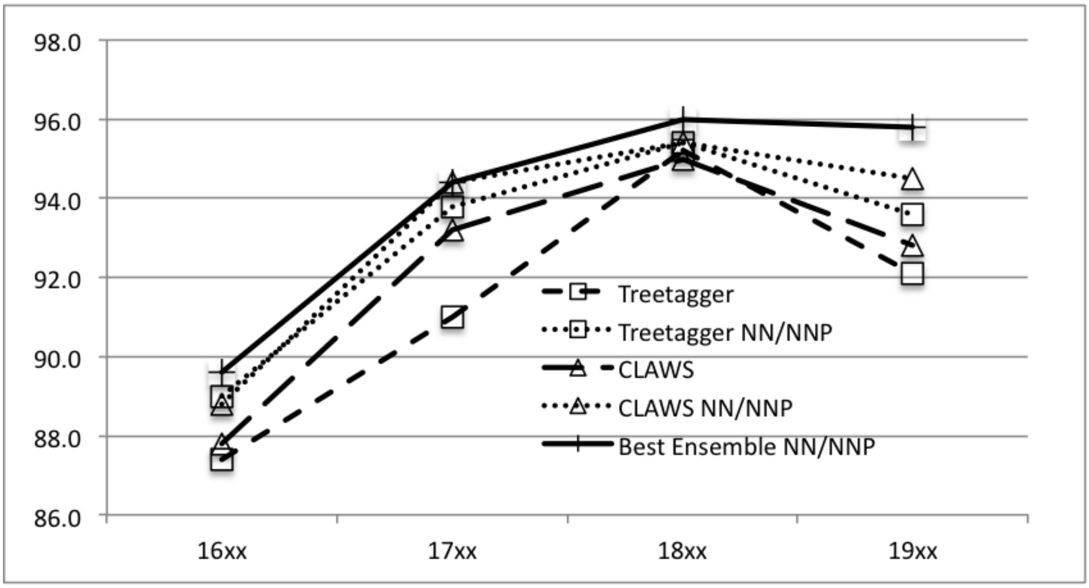

Figure 4: Proper vs. common nouns: Accuracy of Tree-Tagger, CLAWS tagger and Best Ensemble if common noun (*NN*) and proper name (*NNP*) are not distinguished

The accuracies are given in Table 5; Figure 4 contains a visualisation of the accuracy of the individual taggers, with and without the distinction, and the best ensemble, without the distinction between *NN* and *NNP*. We can also see that in this setting, where the tendency of CLAWS to overassign common noun tags to capitalised words is discounted, we reach the same level of accuracy for 19th and 20th century texts.

5 Related Approaches

First, we summarize approaches to present-day language data. Ensemble systems for POS tagging systems have been used by several authors. For example, van Halteren et al. (2001) use an ensemble system to tag two PDE English corpora, the Penn Treebank (Marcus et al., 1993) and LOB (Johansson, 1986). They combine four probilistic taggers with significantly different algorithms (HMM, memory-based, transformation rules, and maximum entropy), reporting that error rates could be reduced by 11% (Penn) to 24% (LOB). On Penn, the single best tagger reached 96.9% accuracy, the best combination increased to 97.2%. On LOB, the single best tagger reached 97.6% accuracy, the best combination increased to 98.1%. Loftsson (2008) combines a rule-based and two probabilistic systems for tagging Icelandic, a morphologically rich langauge in which data sparseness is particularly acute. The combined system, using a simple voting scheme, increases tagging accuracy by almost 1.5% over the best single tagger. In particluar, the improvement is much larger when including the rule-based tagger rather than using three probabilistic taggers, as the comparable approach of Helgadóttir (2004) did, which indicates that the different perspective which the rule-based tagger offers – like CLAWS has done in our approach – is particularly beneficial.

For tagging historical data, we have mentioned in the Methods section that Rayson et al. (2007) also used the normalisation tool VARD, but a single tagger, they report that the normalised text leads to only about half as many tagger errors as the original text. In their experiments on Early Modern German texts, Scheible et al. (2011) measured improved tagging for 47% of the normalised words are tagged better, against a loss of correct results in 3% (and 50% which stay correct or incorrect). Schneider et al. (2014), again on English texts, report that on the subsequent level of syntactic parsing, 32% of the measured syntactic dependencies improve, 2% worsen, and 65% remain unaffected. Bollmann (2013) describes a similar approach using fully automatically normalised German data.

Approaches using domain adaptation exist for English, for example Yang and Eisenstein (2016). Kroch et al. (2004) train a tagger on the historical word forms directly, Dipper (2010) uses the same approach for Middle High German. These appraoches have the advantage that they reduce the risk of error accumulation, which is typical for pipeline systems, and the disadvantage that they are particularly susceptible to sparse data problems.

To our knowledge, there are only very few approaches using ensemble systems on historical data, which has motivated our current research.

6 Conclusions and Outlook

We have demonstrated that for the task of POS tagging of historical English, a careful mapping to PDE spelling with a normalisation tool such as VARD allows one to achieve almost PDE accuracy levels from about 1700 on. We have shown that automatically combining two taggers with sufficiently different approaches improves tagging performance by 0.78% on average. Levels stay slightly below state-of-the art results, as they assume perfect tokenisation, which is unrealistic for real-world texts.

Limited human intervention (choosing one of maximally three alternatives) improves tagging accuracy by an additional 2-5%, thus 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. Surprisingly, 19^{th} century texts can be easier to tag than PDE, which is due partly to more out-of-vocabulary words, partly to "social media" style, partly to complex nouns and abbreviations, and partly to the fact that CLAWS assign common noun tags to proper names too often.

In future research, we want to use more taggers, re-train taggers including more manually annotated historical texts, annotate a larger gold standard, and control for register variation. We are currently testing alternative spelling normalisation tools. We also want to test if the advantage of the CLAWS tagger can be related to its potential to profit from a more fine-grained tagset.

References

Alistair Baron and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham. Aston University.

Douglas Biber, 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*, pages 1–13. Rodopi, Amsterdam.

Marcel Bollmann. 2013. POS tagging for historical texts with sparse training data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability in Discourse*, pages 11–18, Sofia, Bulgaria.

James Curran, 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.

Thomas G. Dietterich. 1997. Machine learning research: Four current directions. *AI Magazine*, 18(4):97–136.

Stefanie Dipper. 2010. POS-tagging of historical language data: First experiments. In *Proceedings of KONVENS*.

Roger Garside 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*, pages 102–121. Longman, London.

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

Sigrún Helgadóttir. 2004. Testing data-driven learning algorithms for PoS tagging of icelandic. In Henrik Holmboe, editor, *Nordisk Sprogteknologi 2004*, pages 257–265, Copenhagen. Museum Tusculanums Forlag.

Stig Johansson. 1986. *The Tagged LOB Corpus: User's Manual*. Norwegian Computing Centre for the Humanities, Bergen, Norway.

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. 2004. *Penn-Helsinki parsed corpus of Early Modern English*. Department of Linguistics, University of Pennsylvania, CD-ROM, first edition, release 3 edition.

Geoffrey Leech, 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.

Hrafn Loftsson. 2008. Tagging icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1).

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

Paul Rayson, 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.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' pos-tagger on early modern german text. In *Proceedings of the ACL-HLT 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*, Portland, Oregon.

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

Gerold Schneider, 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.

Hans van Halteren, Walter Daelemans, and Jakub Zavrel. 2001. Improving accuracy in word class tagging through the combination of machine learning systems. *Computational Linguistics*, 27(2).

Shaoqun Wu. 2010. *Supporting Collocation Learning*. Ph.D. thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical english. In *Proceeding of NAACL*.