Post-Processing of the Royal Society Corpus based on the Noisy Channel Model

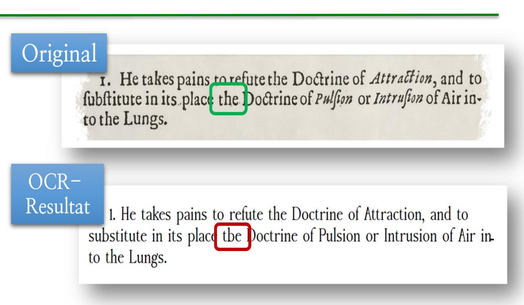
Carsten Klaus[1], Dietrich Klakow[1], Peter Fankhauser[2],

[1] Spoken Language Systems, Saarland Informatics Campus, Saarland University, Saarbrücken, Germany,
[2] Institut für Deutsche Sprache (IDS), Institute for the German Language, Mannheim, Germany
[1] {cklaus|dklakow}@lsv.uni-saarland.de, [2] fankhauser@ids-mannheim.de

Goal

This contribution presents an approach for automatic detection and correction of optical character recognition (OCR) induced misspellings in historical texts.

Problem: Due to old material the OCR procedure distributes misspellings into the digital texts. We target these misspellings in a post-correction step.



Data: Royal Society Corpus

- collection of scientific texts
- from **1665** to **1869**
- published by the *Royal Society of London*
- comprises about **10.000** documents with **35.000.000** tokens in total
- stored in ascending *corpusBuild* versions; we used **v3.7** [Kermes et al. (2016)]

Methodology: The "Noisy Channel Spell Checker"

Our tool is based on the **Noisy Channel Model** by Shannon (1948)

$$\hat{w} = \underset{c \in C}{\operatorname{argmax}}\ P(c)^{\lambda} P(w|c)$$ (weighting)

- **Given**: misspelling w
- **Wanted**: correction $\hat{w}$

Approach: Generate an appropriate set C of potential corrections and estimate the most likely candidate.

The model comprise of two components:

- (n-gram) target language model $P(c)$: How likely occurs c in its context?
- Error model $P(w|c)$: How likely is w a misspelling of c?

$$P(w|c) = \prod_i \frac{C(w_i|c_i)}{C(c_i)}$$

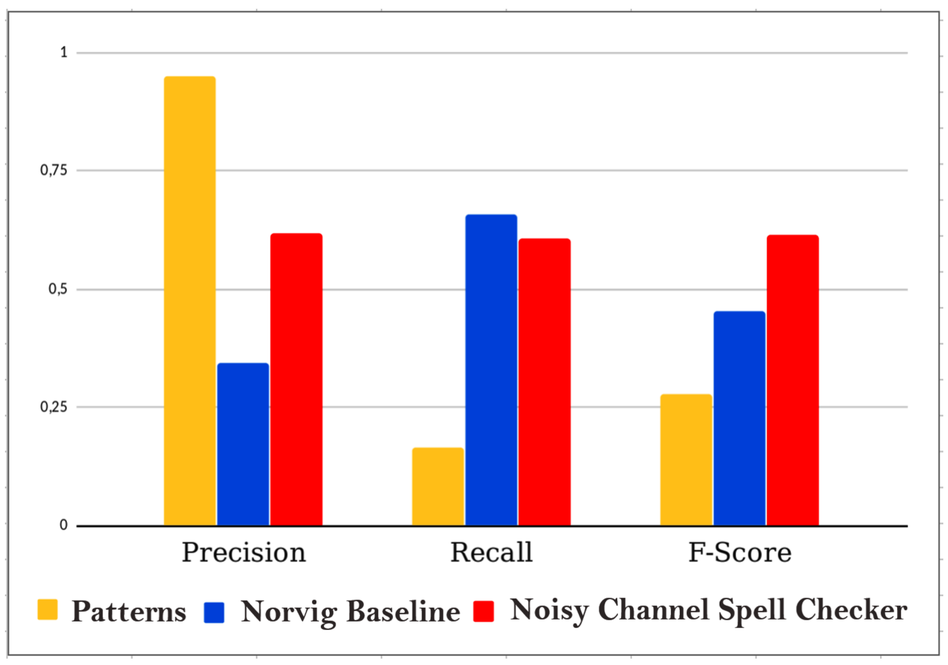
Decompose overall error probability into individual *edit* - probabilities

Special characteristic

Training of the model is completely **corpus specific** – No annotation of data required

Results and Discussion

Evaluation = Extraction of subset of documents from the corpus and manual correction to create a **ground truth**. Comparison of 3 correctors: Pattern-based (state of the art) | Peter Norvig's variant of the NCM (2009) | our Noisy Channel Spell Checker



Observations

- *Patterns* (Knappen et al.) show high Precision; typical for rule-based systems. However bad recall due to its **lack of generalization**.
- *Baseline* corrector tends to **overcorrection** – the other extreme
- Our *Noisy Channel Spell Checker* succeeds somewhere **in between**
  - Is able to **balance** Precision and Recall
  - Corrects hundreds of misspellings **properly** without overcorrecting test set too much
  - **Crucial**: choice of weighting λ

Summary

- **Conclusion**: With F1-Score of **0.61** the *Noisy Channel Spell Checker* significantly outperforms the pattern-based state of the art which only accomplishes an F1-Score of **0.28**.
- **Limitation**: High risk of overcorrection – sensitive adjustment of hyperparameters is essential.
- **Outlook**: Enhancing the denoising of the Royal Society Corpus will promote further investigation. It's conceivable to replace the current technique permanently with the approach presented here.

Scan Here

https://github.com/uds-lsv/Noisy-Channel-Spell-Checker

References

Kermes, H., S. Degaetano, A. Khamis, J. Knappen & E. Teich (2016). **The Royal Society Corpus: From Uncharted Data to Corpus**. In: Proceedings of LREC 2016. Portoroz, Slovenia, p. 1928-1931. The RSC has been made available for free download and online query from the CLARIN-D center at Saarland University under the persistent identifier http://hdl.handle.net/11858/00-246C-0000-0023-8D1C-0 , cf. also http://corpora.clarin-d.uni-saarland.de/cqpweb/. | Knappen, J., Fischer, S., Kermes, H., Teich, E. and Fankhauser, P. (2008): **The Making of the Royal Society Corpus**, in ListLang@NoDaLiDa | Shannon, C.E. (1948): **A Mathematical Theory of Communication**, in Bell System Technical Journal | UdS Fedora Commons Repository (n.d.) **The Royal Society Corpus (RSC)**, https://fedora.clarin-d.unisaarland.de/rsc/. (last accessed 29.03.2018) | Norvig, P. (2008): **Natural Language Corpus Data: Beautiful Data**. (online) http://norvig.com/ngrams/ (last accessed 08.11.2017). | Jurafsky D., Martin J. H. (2016). **Spelling Correction and the Noisy Channel** In: Speech and Language Processing, 3. Edition, S. 61-73.