OCR Nachkorrektur des Royal Society Corpus

Klaus, Carsten
s8caklau@stud.uni-saarland.de
Universität des Saarlandes, Saarbrücken, Deutschland

Fankhauser, Peter
fankhauser@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Klakow, Dietrich
dklakow@lsv.uni-saarland.de
Universität des Saarlandes, Saarbrücken, Deutschland

Einleitung

Linguistische Analysen historischer Texte stellen Forscher oftmals vor große Herausforderungen. Im Gegensatz zur Digitalisierung moderner Dokumente kann es bei jahrhundertealten Texten zu Schwierigkeiten kommen. Diese weisen oftmals eine geringere Qualität auf, sodass es beim Einlesen zu Fehlern kommt. Solche können schwerwiegende Störfaktoren für weitere Analysen sein. In diesem Beitrag beschreiben wir den **Noisy Channel Spell Checker**, ein Verfahren zur automatisierten Korrektur von Optical Character Recognition (OCR) induzierten Rechtschreibfehlern in historischen Texten, genauer dem **Royal Society Corpus**.

Beim Royal Society Corpus (RSC) handelt es sich um eine Sammlung wissenschaftlicher Texte von 1665 bis 1869, veröffentlicht im Journal *Philosophical Transactions of the Royal Society of London.* Das Korpus umfasst ungefähr 10.000 Dokumente mit insgesamt 35.000.000 Tokens. Die Texte wurden mithilfe von Optical Character Recognition digitalisiert, bedingt durch das alte Material der Dokumente wurden jedoch Worte falsch erkannt und somit Rechtschreibfehler eingestreut. Diese sollen in einer Nachkorrektur berichtigt werden. (UdS Fedora Commons o.J.)

State of the Art

Das Korpus wird einer strikten Versionskontrolle unterzogen. Fortschritte bzgl. Formatierung oder Fehlerkorrektur werden in aufsteigenden *corpusBuild* Versionen festgehalten. Derzeit wird das Royal Society Corpus durch einen **Pattern**-basierten Ansatz bereinigt (Knappen, 2017). Hierbei werden Ersetzungsregeln auf die Texte angewendet um Fehler mit ihrer richtigen Form auszutauschen, wie beispielsweise *tbe # the*. Der große Nachteil dieses Verfahrens ist jedoch, dass nur ein Bruchteil der induzierten OCR Fehler abgedeckt wird, was in einer geringen Fehlererkennung resultiert. Im Folgenden erläutern wir unseren Ansatz, welcher mit einem statistischen Lernverfahren deutlich bessere Ergebnisse erzielt.

Methodik

Der *Noisy Channel Spell Checker* basiert auf dem **Noisy Channel Model** (Shannon, 1948). Ein potentiell fehlerhaftes Wort w wird wie folgt korrigiert: Aus einer Vorauswahl an geeigneten Kandidaten c aus C wird abgeschätzt welcher am ehesten als Korrektur # in Frage kommt.

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

Das Noisy Channel Model besteht zum einen aus dem **Sprachmodell** P(c) und zum anderen dem **Fehlermodell** P(w|c). Es werden hierbei zwei intuitive Gedanken kombiniert: Das Sprachmodell schätzt die Wahrscheinlichkeit des Kandidaten in seinem Wortkontext ab. Hochfrequentierte Worte sind demnach sehr wahrscheinlich. Das Gegengewicht hierzu bildet das Fehlermodell. Diese Verteilung gibt an wie sicher w eine fehlerhafte Variante von c ist, schätzt also ab, wie wahrscheinlich einzelne Korrekturschritte von w nach c sind. # ist ein frei wählbarer Parameter, mithilfe dessen man das Sprachmodell gewichten kann. (Jurafsky 2016: 61-73)

Training des Modells

Die Besonderheit unseres Ansatzes besteht darin, dass Sprach-, sowie Fehlermodell **korpusspezifisch** trainiert werden. Es sind keine aufwändigen Trainingsdatenannotationen notwendig, denn es werden lediglich die Korpusdateien verwendet.

- Das **Sprachmodell** wurde mithilfe der aktuellsten *corpusBuild* Version des Royal Society Corpus trainiert. Diese Texte sind durch die Patterns bereits best möglich bereinigt worden. Somit wurde versucht das Rauschen innerhalb der Verteilung zu reduzieren.
- Zum Trainieren des **Fehlermodells** wurden die bereits erwähnten Patterns als Wissensbasis hinzugezogen. Die Idee war hier aus der Korrektur durch die Patterns eine Wahrscheinlichkeitsverteilung zu erzeugen, also das Fehlerverhalten im Korpus zu generalisieren. Anhand eines Beispiels lässt sich dies veranschaulichen: Gegeben die Ersetzungsregel fuch # such. Diese wird in folgende Sequenz von edit Operationen aufgebrochen: f|s + u|u + c|c + h|h. Der

Trainingsprozess erfasst nun wie oft edit Operationen angewendet wurden und leitet daraus eine Verteilung ab.

Resultate und Diskussion

Als Testmenge haben wir 26 Dokumente aus dem Korpus extrahiert. Diese wurden eigens korrigiert um einen Gold Standard zu erhalten. Als Evaluationsmetriken wählten wir *Precision* (Anteil der validen Korrekturen), *Recall* (Abdeckung der einzelnen Fehler) und daraus den *F1-Score* (harmonisches Mittel aus Pre. und Rec.). Um die Ergebnisse unserer Arbeit zu vergleichen, haben wir zwei weitere Methoden auf die Testdaten angewendet. Dies waren zum einen die **Pattern s** und zum anderen nutzten wir als Referenzkorrektur für das Noisy Channel Model eine Implementierung von **Peter Norvig** (Norvig, 2009). Die Ergebnisse sind in Abbildung 1 aufgetragen.

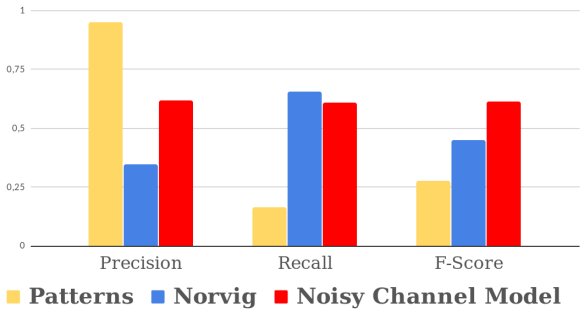

Abbildung 1: Resultate einzelner Korrekturmethoden angewendet auf den Testdatensatz

Man kann erkennen, dass die Pattern korrektur (gelb) die beste Precision erziel t . Dies ist ein typisches Verhalten regelbasierte r Systeme. Im Gegensatz dazu decken die beiden anderen Verfahren eine größere Menge an Fehlern ab, dies wird am höheren Recall deutlich. Besonders Norvigs Variante (blau) ist hier führend, jedoch tendiert diese auch zur Überkorrektur von richtig erfassten Wörtern. Wir waren bestrebt, dass unser Spell Checker (rot) dies weitestgehend vermeidet, indem es Precision und Recall möglichst balanciert. Es werden also viele OCR Rechtschreibfehler korrigiert und gleichzeitig wird die Rate an Falsch Positiven gering gehalten . Hierbei w ar das Optimieren der Gewicht u ng # des Language Models ein essentieller Bestandteil der Arbeit, sodass unser Modell schlussendlich einen F-Score von 0.61 2 erzielte. Bei der Überlegung unseren Ansatz auf andere historische, unaufbereitete Texte anzuwenden empfiehlt es sich das Fehlerverhalten in diesen Texten bestmöglich zu generalisieren . Deshalb sollte bereits eine Wissensbasis in Form von Ersetzungspatterns vorliegen um das Error Model korpusspezifisch zu trainieren, das heißt genauso wie in diesem Beitrag beschrieben.

Zusammenfassung

Im Vergleich zur derzeitigen pattern-basierten Methode verbesserte der *Noisy Channel Spell Checker* die Korrekturqualität um mehr als das Doppelte. Es werden nun Fehler berichtigt, die die Patterns nicht einmal als solche erkennen. Die Hauptmotivation zum Aufbau des Royal Society Corpus sind Untersuchungen der diachronischen Entwicklung von wissenschaftlichem Englisch (UdS Fedora Commons o.J.). Die Bereinigung der Texte macht es möglich, dass diese Analysen in Zukunft weitaus genauer und verlässlicher werden.

Bibliographie

Jurafsky Daniel / Martin James H. (2016): *"Spelling Correction and the Noisy Channel"* In: Speech and Language Processing, 3. Edition, S. 61-73.

Kermes, Hannah / Degaetano-Ortlieb, Stefania / Khamis, Ashraf / Knappen, Jörg / Teich, Elke (2016): *"The Royal Society Corpus: From Uncharted Data to Corpus"*, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA).

Knappen, Jörg / Fischer, Stefan / Kermes, Hannah / Teich, Elke / Fankhauser, Peter (2017): *"The Making of the Royal Society Corpus"*, in ListLang@NoDaLiDa.

Norvig, Peter (2008): *"Natural Language Corpus Data: Beautiful Data"*. [online] http://norvig.com/ngrams/ [letzter Zugriff 08. November 2017].

Shannon, Claude E. (1948): *"A Mathematical Theory of Communication"*, in Bell System Technical Journal.

UdS Fedora Commons Repository (o.J.): *"The Royal Society Corpus (RSC)"*, https://fedora.clarin-d.uni-saarland.de/rsc/ . [letzter Zugriff 29. März 2018].