Optimieren von POS-Tagger-Ergebnissen

Dokumentenart:	Expose für eine Bachelorsarbeit
Autor:	Viktor Kiesel
Matrikel-Nr.:	1542010
Studiengang:	Informatik Bachelor
Betreuer:	Mathias Landhäußer, Sebastian Weigelt
Datum:	30. November 2015

1 Motivation

Der Fachbereich des maschinellen Lernens ist in der Informatik von großer Bedeutung. Das Ziel ist es neue Informationen aus bereits bekanntem Wissen zu gewinnen. Ein Algorithmus soll hierbei nicht einfach nur einen Datensatz abrufen, sondern aus den Daten Regelmäßigkeiten und Muster erlernen. Durch Verwendung dieser vorher trainierten Muster sollen neue vorher unbekannte Daten überprüft und in entsprechende Gruppen bzw. Klassen eingeteilt werden.

Klassifikationsalgorithmen, die den Wörtern in einem Satz ihre entsprechende Wortartmarkierungen (engl. *„Part of Spech“*-Tags (POS-Tags)), mit Hilfe eines vorher trainiertem Models, zuweisen, nennt man Wortartmarkierer (engl. *„Part of Speech“*-Tagger). Die Qualität dieser Markierer ist abhängig von der Qualität des trainierten Models.

Moderne Wortartmarkierer erreichen unter angepassten Bedingungen eine sehr gute Genauigkeit von rund 97%, was der Leistung eines einzelnen menschlichen Annotators entspricht oder übertrifft [MMS93, S.319].

Werden diese optimalen Bedingungen zum Negativen hin verändert, so sinkt die Genauigkeit. Ein Beispiel hierfür wäre ein Wechsel des Fachbereichs zwischen Training und Klassifikation. Es ist daher nötig ein neues Modell zu trainieren, was bei einem kleinen Datensatz suboptimale Ergebnisse liefert [Bra00, S.228]. Um eine langwierige Markierung eines großen Korpus von Hand zu vermeiden soll die alternative Methode der Kombination von Wortartmarkierern untersucht werden.

Für die Verarbeitung von natürlicher Sprache stellt die Kenntnis der Wortart eine wichtige Grundlage dar. Da Fehler sich stark durch Folgefehler auswirken können, ist es besonders wichtig sie zu vermeiden.

2 Zielsetzung

Aus verschiedenen wissenschaftlichen Arbeiten folgt, dass eine Optimierung der Wortartmarkierer durch Kombination möglich ist[BW98]. Die meist verwendete Kombinationsweise ist das einfacher Mehrheitsentscheid mit nur wenigen Wortartmarkierern. Es soll daher überprüft werden ob andere Klassifikationsalgorithmen zur Kombination genutzt werden können und ob eine größere Verbesserung als mit einem einfachen Mehrheitsentscheid möglich ist.

Aus diesem Hauptziel folgen einige Teilziele. Um überhaupt eine Kombination durchführen zu können, müssen zuallererst die unterschiedlichen Wortartmarkierer zum Markieren von Texten koordiniert werden. Die Qualität dieser Markierer muss anhand eines Goldstandards bestimmt werden.

Im Fachbereich der Wortartmarkierungen besteht der Goldstandard aus Sätzen, bei dem Wörter mit ihren korrekten Wortartmarkierungen gespeichert sind. Für diese Arbeit wurden bereitgestellte NLCI- und Parse-Korpora markiert.

Für das nächste Teilziel sollen aus Markierungen, mit unterschiedlichen Klassifikatoren, Modelle gebaut werden. Mit Hilfe dieser Modelle soll es möglich sein, weitere Texte markieren zu können.

Das letzte Teilziel ist die Analyse der Verbesserung durch Kombinationsalgorithmen.

Als nicht funktionelle Anforderung wurde, im bereits fertiggestellten Entwurf, auf eine hohe Erweiterbarkeit geachtet. Zudem sollen in dieser Arbeit keine Kombinationsalgorithmen implementiert, sondern auf das Werkzeug Weka [HFH+09] zurückgegriffen werden.

3 Analyse und Evaluation

In der Vorbereitungszeit wurden 69 Wortartmarkierer recherchiert, wovon 16 genutzt werden. Ali und Pazzani[AP96] zeigten, dass eine Kombination von Klassifikatoren dann sinnvoll ist, wenn diese von einander unabhängige Fehler machen. Deswegen wurde bei der Auswahl darauf geachtet, dass die Wortartmarkierer mit möglichst unterschiedlichen Algorithmen markieren. Die andere Voraussetzung bei der Auswahl war, dass die ausgewählten Markierer die Penn-Wortartmarkierungsmenge (engl. *„Penn“*-Tagset) nutzen.

Die Evaluation der Arbeit findet durch einen Vergleich der Genauigkeiten der ausgewählten Wortartmarkierern und den Optimierungen durch Kombinationsalgorithmen statt. Als zu betrachtende Metrik wird die Wortgenauigkeiten der einzelnen Wortartmarkierer auf dem ausgewählten Goldstandard ausgewählt.

Um eine Verbesserung durch die Kombination der unterschiedlichen Wortartmarkierer feststellen zu können, muss ein Vergleichswert festgelegt wer-

den. Da der hier verwendete Goldstandard nicht mit dem *„Wall Street Journal“*-Korpus (WSJ-Korpus) übereinstimmt, wird jeder Wortartmarkierer zunächst einzeln geprüft und seine Genauigkeit auf dem verwendeten Goldstandard festgestellt. Wurde eine Genauigkeit auf dem WSJ-Korpus angegeben, so wird diese Genauigkeit mit dem Ergebnis dieser Arbeit verglichen und die Differenz festgestellt.

Neben den Wortartmarkierern sollen auch ausgewählte Kombinationsalgorithmen untereinander verglichen werden. Die Evaluation eines Algorithmus findet über Kreuzvalidierung statt. Hierbei wird ein Korpus zufällig in Trainingsbereich und Testbereich eingeteilt und darauf ein Modell trainiert und getestet. Dies wird wiederholt und die Durchschnittsgenauigkeit als Endresultat festgestellt. Mit Hilfe der Endresultate können die einzelnen Klassifikationsalgorithmen untereinander quantitativ verglichen werden.

Es soll weiterhin festgestellt werden, ob es Klassifikationsalgorithmen gibt, die eine Verbesserung zum besten Wortartmarkierer erlauben. Von den Klassifikatoren, die diese Schranke überschreiten sollen die besten festgestellt werden.

Literatur

[AP96] Ali, Kamal M. ; Pazzani, Michael J.: Error reduction through learning multiple descriptions. In: *Machine Learning* 24 (1996), September, Nr. 3, 173–202. `http://dx.doi.org/10.1007/BF00058611`. – DOI 10.1007/BF00058611. – ISSN 0885–6125, 1573–0565

[Bra00] Brants, Thorsten: TnT: A Statistical Part-of-speech Tagger. In: *Proceedings of the Sixth Conference on Applied Natural Language Processing*. Stroudsburg, PA, USA : Association for Computational Linguistics, 2000 (ANLC '00), 224–231

[BW98] Brill, Eric ; Wu, Jun: Classifier Combination for Improved Lexical Disambiguation. In: *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*. Stroudsburg, PA, USA : Association for Computational Linguistics, 1998 (COLING '98), 191–195

[HFH+09] Hall, Mark ; Frank, Eibe ; Holmes, Geoffrey ; Pfahringer, Bernhard ; Reutemann, Peter ; Witten, Ian H.: The WEKA Data Mining Software: An Update. In: *SIGKDD Explor. Newsl.* 11 (2009), November, Nr. 1, 10–18. `http://dx.doi.org/10.1145/1656274.1656278`. – DOI 10.1145/1656274.1656278. – ISSN 1931–0145

[MMS93] Marcus, Mitchell P. ; Marcinkiewicz, Mary A. ; Santorini, Beatrice: Building a Large Annotated Corpus of English: The Penn Treebank. In: *Comput. Linguist.* 19 (1993), Juni, Nr. 2, 313–330. http://dl.acm.org/citation.cfm?id=972470.972475. – ISSN 0891–2017