EIN POS-TAGGER FÜR „DAS" MITTELHOCHDEUTSCHE[1]

Nora Echelmeyer (nora.echelmeyer@ilw.uni-stuttgart.de), Nils Reiter (nils.reiter@ims.uni-stuttgart.de), Sarah Schulz (sarah.schulz@ims.uni-stuttgart.de), Universität Stuttgart

Einleitung

Ein grundlegender Schritt für eine Vielzahl von Aufgaben aus dem Bereich des *Natural Language Processing* (NLP) ist das *Part of Speech* (PoS)-Tagging. Ein PoS-Tagger annotiert im Kontext eines Satzes jedes Wort mit seiner Wortart aus einer Menge an festgelegten Wortarten (Tagset).

Ein Großteil der dazu vorhandenen Arbeiten konzentriert sich auf das Englische, auch für das Neuhochdeutsche sind vergleichbar viele Daten verfügbar. Historische Sprachstufen stellen hingegen eine Herausforderung für NLP-Aufgaben wie PoS-Tagging dar, da sie keine Standardsprache kennen, sondern nur als Vielfalt dialektaler Varietäten existieren, und ihre Verschriftlichung nicht nach einheitlichen Regeln erfolgt. Dies schlägt sich in einer hohen Varianz nieder, was die Annotation einer ausreichenden Menge an Referenzdaten erschwert.

Mit diesem Beitrag möchten wir einen PoS-Tagger für das Mittelhochdeutsche vorstellen, der auf einem thematisch breiten und diachronen Korpus trainiert wurde. Als Tagset verwenden wir ein Inventar aus 17 universellen Wortart-Kategorien (*Universal Dependency*-Tagset, Nivre et al. 2016). Mit den annotierten Daten entwickeln wir ein Modell für den TreeTagger (Schmid 1995), das frei zugänglich ist.

Dabei vergleichen wir drei verschiedene Möglichkeiten, den PoS-Tagger zu trainieren. Zunächst verwenden wir ein kleines, manuell annotiertes Trainingsset, vergleichen dessen Ergebnisse dann mit einem kleinen, automatisch disambiguierten Trainingsset und schließlich mit den maximal verfügbaren Daten.

Mit dem Tagger möchten wir nicht nur eine „Marktlücke" schließen (denn bisher gibt es keinen frei verwendbaren PoS-Tagger für das Mittelhochdeutsche), sondern auch eine größtmögliche Anwendbarkeit auf mittelhochdeutsche Texte verschiedener Gattungen, Jahrhunderte und regionaler Varietäten erreichen und weiteren Arbeiten mit mittelhochdeutschen Texten den Weg ebnen.

Forschungsstand

Tagset

Als PoS-Tagset hat sich für Neuhochdeutsch das Stuttgart-Tübingen-Tagset (STTS) etabliert (Schiller et al. 1999). Um auf die Besonderheiten historischer Sprachstufen besser eingehen zu können, entwickelten Dipper et al. (2013) ein hieran angelehntes Historisches Tagset (HiTS), das aus 12 Wortklassen besteht, die sich ihrerseits in 84 Wortarten gliedern.

Mit dem Ziel eines universellen Tagsets, welches konsistente Annotation vereinfacht und sprachübergreifendes Lernen für automatische Syntaxannotationen ermöglicht, wurde im

[1] veröffentlicht in: DHd 2017. Digitale Nachhaltigkeit. Konferenzabstracts. Universität Bern. 13. bis 18. Februar 2017, S. 141–147.

Rahmen des *Universal Dependency*-Projekts (UD) ein Tagset aus 17 Tags erstellt. Dieses kann bei Bedarf um sprachspezifische Tags erweitert werden.

Tagging

Die besten verfügbaren PoS-Tagger erreichen auf englischsprachigen Zeitungstexten über 97% Accuracy (cf. Spoustová et al. 2009). Für deutschsprachige Zeitungstexte werden um die 95% erzielt, für Web-Texte 90–93% (Giesbrecht / Evert 2009). PoS-Tagging für das Mittelhochdeutsche ist weit weniger erforscht. Schulz / Kuhn (2016) beschreiben Ansätze zum PoS-Tagging eines spezifischen Textes. Barteld et al. (2015) trainieren einen PoS-Tagger für Mittelniederdeutsch. Dipper (2011) berichtet eine Accuracy von ca. 92% für zwei spezifische Modelle für die Dialekte Ober- und Mitteldeutsch, trainiert auf normalisierten Lemmata und mit dem STTS-Tagset. Alle genannten Modelle sind auf eine bestimmte Varietät des Mittelhochdeutschen beschränkt, zudem ist keines dieser Modelle (soweit uns bekannt) öffentlich verfügbar.

Korpora

Die wohl umfangreichsten Projekte zum Mittelhochdeutschen sind das Wörterbuchnetz[2], ein online zugänglicher Verbund aus Nachschlagewerken, das linguistisch motivierte Referenzkorpus Mittelhochdeutsch[3] (ReM, Dipper 2015) sowie die Mittelhochdeutsche Begriffsdatenbank[4] (s.u.).

Die annotierten Textkorpora (cf. Dipper 2015: 521–526) können z.T. über das Suchtool ANNIS (Zeldes et al. 2009) abgefragt werden, wobei Suchanfragen auf den Ebenen Wortform, Lemma und Morphologie möglich sind.

Mittelhochdeutsche Begriffsdatenbank (MHDBDB)

Durch eine Kooperation mit dem Projekt „Mittelhochdeutschen Begriffsdatenbank“ konnten wir für die Entwicklung unseres PoS-Taggers auf eine reiche Datensammlung zurückgreifen, bestehend aus 658 Texten mit insgesamt knapp 10 Millionen Tokens. Die Texte umfassen eine Zeitspanne von etwa vier Jahrhunderten (1100–1500), verschiedene dialektale Ausprägungen sowie nahezu alle Gattungen (von großepischen Genres wie Artusroman, Heldenepik und Antikenroman über Kleinepik hin zu Lyrik sowie diversen nicht-literarischen Texten wie Kochbüchern, Alchemistischen Schriften und Flugblättern).

Kürzel	**Name**	**Beispiel**
NOM	Nomen	acker, zît
NAM	Name	Uolrîch, Wiene, Rhîn
ADJ	Adjektiv	grôz, schoene
ADV	Adverb	schone, schnelleclîche
ART	Artikel	der, eine
DET	Determinante	ditze, mîn, ieman
POS	Possessivpronomen	mîn, dîn, unser
PRO	Pronomen	ich, ez, wir
PRP	Präposition	ûf, zuo, under
NEG	Negation	nie, âne, niht
NUM	Numeral	ein, zwô, zweinzegest

[2] Online zugänglich unter http://woerterbuchnetz.de/, letzter Zugriff 23.08.2016.
[3] Online zugänglich unter http://referenzkorpus-mhd.uni-bonn.de, letzter Zugriff 23.08.2016.
[4] Online zugänglich unter http://mhdbdb.sbg.ac.at/ (MHDBDB), letzter Zugriff 23.08.2016.

CNJ	Konjunktion	als, und, abr
GRA	Gradationspartikel	sêre, vil
IPA	Interrogativpartikel	swer, swar, wie
VRB	Verb	liuhten, varn
VEX	Hilfsverb	haben, sîn, werden
VEM	Modalverb	müezen, suln
INJ	Interjektion	ahî, owê
CPA	Komparativpartikel	als, wie
DIG	Zahl (Digit)	IX, XVII, III

Tabelle 1: Grammatische Kategorien der MHDBDB

Die Daten der MHDBDB enthalten – neben Tokenisierung und Lemmatisierung – bereits grammatische Auszeichnungen (Tabelle 1). Diese sind allerdings nicht disambiguiert, da sie den Kontext eines Wortes unberücksichtigt lassen (Typ-level-Annotationen, z.B. NOM|ADJ|ADV für *guot*). Darüber hinaus kodieren sie die morphologische Zusammensetzung von Wörtern (z.B. NOM|NEG für *unheil*), so dass es zu häufigen Mehrfachauszeichnungen kommt (z.B. *unvuoge* NOM|ADJ|ADV|NEG). Hinzu kommt, dass das Tagset nicht alle möglichen Verwendungsformen der Wörter abdeckt: So kann z.B. *daz* nicht nur Artikel oder subordinierende Konjunktion sein (Satz 1), sondern auch als Relativ- (2) oder Demonstrativpronomen (3) fungieren:

(1) Daz edel kint hât mir verjehen, daz ez in troume sî geschehen.
(2) Wie staete ist ein dünnez eis daz ougestheize sunnen hât?
(3) Daz sage ich iu vür ungelogen.

Trainings- und Testdaten

Obige Beobachtungen zeigen exemplarisch, dass die grammatischen Auszeichnungen der MHDBDB einer Überarbeitung bedürfen, um für die Entwicklung eines PoS-Taggers nutzbar zu sein. Dazu annotieren wir ein Teilkorpus, das für den mittelhochdeutschen Wortart-Tagger als Trainings- und Testdatei dient und für eine automatische Re-Annotation der restlichen MHDBDB-Daten herangezogen wird. Um Anschlussuntersuchungen sowie sprachübergreifende Betrachtungen zu ermöglichen, greifen wir für unsere Annotationen auf die universellen Kategorien aus dem UD-Tagset zurück (Tabelle 2).

Tag		**Anmerkungen**	**Beispiele**
ADJ	adjective	vorangestellt, nachgestellt, Partizipien	der ritter **guot**; daz **elder** kint; **roemisch** lant; der **ander** man
ADP	adposition	Prä-, Post- und Zirkumposition	**mit** dem swerte; **gein** Nantes; **âne** ir schulde
ADV	adverb	auch adverbial gebrauchte Adjektive und relativischer Gebrauch	der ritter **lîdenlîche** leit; **so** sprach der künec; rehte liebe im **nie** geschach; hôret, **swie** er ze strîte quam
AUX	auxiliary verb	Hilfs- und Modalverben	ich **muoz** ir dienen; die **sint** enterbet; ir **habet** ez von mir gehôrt
CONJ	coordinating conjunction	nebenordnend	ritter **unde** diep; zwei teil **oder** mêr; **denne** ich welle jehen
DET	determiner	Artikel (bestimmt und unbestimmt); attribuierende Demonstrativ-, Possessiv- und Relativpronomen	**ein** maere; **der** ritter guot; **diz** bîspel; **dirre** âventiure; **ir** triuwe; **mîn** bruoder; **dehein** man; in **welhem** lande
INTJ	interjection		ouwê; ach!
NOUN	noun	auch substantivierte Adjektive, Verben, Numeralia	diu **vrouwe**; **duc** Orilus; **rîche** und **arme**; die **drî**; daz **singen**
NUM	numeral	nur Kardinalzahlen	die **drî** ritter

PART	particle	Negationspartikel; abgetrennter Verbzusatz; zu (mit Inf.); Vergleichs-partikel; Abtönungspartikel	daz **en**weiz ich **niht**; lâzet allez trûren **abe**; daz ist swere **ze** halten; snêwîz **als** ein harm
PRON	pronoun	Personal-, Relativ-, Reflexivpronomen; substituierende Possessiv-, Indefinit-, Demonstrativ-, Interrogativpronomen	**er** lac tôt; Isenhârten, **der** den lîp verlôs; die kuenen heten **sich** berâten; der **sîne** sprach; da was **nieman**; **allez**, **daz** ich habe; **man** saget; **diz** was dô getân; **swaz** er gebôt
PROPN	proper noun	Eigennamen (auch mehrteilig)	Parzivâl; Orilus de Lalander; Nantes
SPUNCT	punctuation	Satz-beendende Zeichen	. : ! ?
PUNCT	punctuation	alle sonstigen Satzzeichen	, ; < > „ / ( ) usw.
SCONJ	subordinating conjunction	unterordnend	Er sagete **daz** Isenhart küneclîch bestatet wart; **sît** er an mir ist sus verzagt; **ob** mich gelücke wil bewarn
SYM	symbol		
VERB	verb	alle Vollverben	Er **lac** tôt; wir suln kurzwîl **phlegen**; er **hat** ein grôz her; ir reht was **vernomen**
X	other		

Tabelle 2: *Universal Dependency* (UD)-Tagset. Zum besseren Verständnis wurde das Tagset mit Beispielen und Anmerkungen versehen. Das Tag SYM wurde nicht benötigt; hingegen wurde das Tag SPUNCT hinzugefügt, um Satz-beendende Satzzeichen von anderen Satzzeichen zu unterscheiden.

Manuelle PoS-Annotationen

Das manuell annotierte Teilkorpus besteht aus 20.000 Tokens. Ein Teil der Daten (1.500 Tokens) wurde doppelt annotiert, um das Inter-Annotator-Agreement zu bestimmen (Cohen‘s kappa: 0.88; Cohen 1960). Um der Heterogenität der Sprache gerecht zu werden, enthält das Teilkorpus zufällig ausgewählte Abschnitte aus verschiedenen Textsorten des Gesamtkorpus.

Durch die Annotation aller Wörter im Kontext eines Satzes wurden Ambiguitäten aufgehoben. Zur Bestimmung der Wortart kann der Substitutionstest herangezogen werden, bei dem ein Wort durch ein Wort der gleichen Kategorie ersetzt wird. So wird *schoene* in *daz schoene wîp* durch ein anderes Adjektiv (z.B. *daz minnicliche wîp*), in *die schoene saz bî ime* hingegen durch ein Nomen ersetzt (z.B. *die vrouwe saz bî ime*).

Als Schwierigkeiten bei der Annotation stellten sich u.a. die Trennschärfe von DET und ADJ heraus (insb. für Wörter wie „viele“/„alle“) oder die Annotation von noch nicht lexikalisierten bzw. grammatikalisierten Formen (z.B. das mittelhochdeutsche *sît daz*, bei dem die Bestandteile ADP und PRON noch identifizierbar sind, wohingegen neuhochdeutsch „seitdem“ eine SCONJ ist).

Ein weiterer Sonderfall des Mittelhochdeutschen besteht in der (weitgehend unsystematischen) Verwendung klitischer Formen, z.B. der Verschmelzung von Negationspartikel und Verb (*enmac*), der Kontraktion mehrerer Pronomen (*siz = sie+ez*), von Pronomen und Adposition (*zem = ze+im*) o.Ä. In solchen Fällen werden alle miteinander verschmolzenen Wörter annotiert, wobei ein + die Verschmelzung der Wörter anzeigt (*zem* ADP+PRON). Das UD-Tagset muss für das Mittelhochdeutsche also um „kombinierte Tags“ (in unseren Daten finden sich 23 verschiedene Kombinationen) erweitert werden.

Automatische Disambiguierung des Gesamtkorpus

Das annotierte Teilkorpus dient neben seiner direkten Verwendung als Trainings- und Testkorpus (Modell 1) auch der automatischen Disambiguierung des Gesamtkorpus. Hierfür verwenden wir einen sequenziellen Tagger (Conditional Random Fields), der auf dem manuell annotierten Subkorpus trainiert wurde. Dieser lernt anhand der Annotationen und wortbasierten

Eigenschaften, die ambigen Annotationen auf ihre disambiguierten Entsprechungen (UD-Tagset) abzubilden. Da sich in den Daten *auch* nicht-ambige Wörter befinden, lernt der Tagger an vielen Stellen 1-zu-1-Abbildungen, die als Anker fungieren können.

Die Disambiguierung des Gesamtkorpus erreicht eine Accuracy von 86,9%. Die auf diese Weise disambiguierten Daten kommen für die Modelle 2 und 3 als Trainingsdaten zum Einsatz.

Experiment und Evaluation

Um die Schwierigkeit der Aufgabe und die Tagging-Qualität einschätzen zu können, vergleichen wir drei verschiedene Modelle:

- Baseline: Anwendung des neuhochdeutschen TreeTagger-Modells auf den Testdaten.
- Modell 1: Der TreeTagger wird nur mit den manuell annotierten Daten trainiert, die Evaluation erfolgt als 5-fache Kreuzvalidierung (Cross-Validation), so dass in jedem Durchgang 16k Tokens als Trainingsdaten zur Verfügung stehen. Vorteil: Qualitativ hochwertige Trainingsdaten, Nachteil: Geringe Datenmenge.
- Modell 2: Der TreeTagger wird auf zufällig ausgewählten Sätzen trainiert, die zusammen etwa 16k automatisch disambiguierte Tokens umfassen. Die Trainingsmenge ist damit gleich groß wie für Modell 1 und erlaubt, die Auswirkungen der nicht-perfekten Disambiguierung abzuschätzen.
- Modell 3: Der TreeTagger wird mit allen automatisch disambiguierten Daten aus der MHDBDB trainiert (9,9M Tokens).

Modell	**Precision**	**Recall**	**F-Score**	**Accuracy**
Baseline	40,3	35,4	33,1	45,4
Modell 1 (kleines Trainingsset, manuell annotiert)	86,0	80,3	82,2	87,0
Modell 2 (kleines Trainingsset, autom. disambiguiert)	84,8	68,8	72,3	84,7
Modell 3 (großes Trainingsset, autom. disambiguiert)	91,2	79,6	82,9	90,9

Tabelle 3: Ergebnisse des PoS-Taggings mit verschiedenen Modellen.5 Alle Modelle wurden auf den gleichen Daten evaluiert, für Modell 1 kam Cross-Validation zum Einsatz. Der Precision, Recall und F-Score ermöglichen eine tiefere Einsicht in die Performanz unter Berücksichtigung aller Wortartenklassen, während Accuracy die Gesamtperformanz sichtbar macht und als Vergleichswert zu State-of-the-Art-Ergebnissen dient.

Die Ergebnisse der unterschiedlichen Modelle sind in Tabelle 3 zusammengefasst. Zunächst zeigt sich erwartungsgemäß, dass die Baseline keine zufriedenstellenden Ergebnisse liefert. Modell 1 erreicht eine Accuracy von 87%, Modell 2 gut 2 Prozentpunkte weniger. Angesichts der Tatsache, dass die Trainingsdaten automatisch disambiguiert wurden, ist das nur ein geringer Verlust. Die Performanz steigt deutlich, wenn das große Datenset zum Training herangezogen wird (Modell 3). Gegenüber Modell 1 erreichen wir eine Verbesserung von ca. 3 Prozentpunkten Accuracy und damit insgesamt fast 91%. Eine Kombination der Modelle 1 und 3 erzielte keine Verbesserungen gegenüber Modell 3.

Eine Inspektion der von Modell 3 produzierten Annotationen ergibt, dass ein Großteil der Fehler (53%) auf die kombinierten Tags entfallen, die überwiegend als Pronomen oder Verben

[5] Für die Evaluation wurden die kombinierten Tags zu einer Klasse zusammengefasst.

getaggt werden. Die nächsthäufigsten Fehlerklassen sind Numeralia und Partikeln. Die meisten Inhaltswörter (Nomen, Verben) werden korrekt erkannt (> 90%).

Fazit

Mit unserem Beitrag stellen wir einen nahezu universellen PoS-Tagger für das Mittelhochdeutsche vor, der auf Daten trainiert wurde, die dialektal, zeitlich sowie genremäßig variantenreich sind. Damit gehen wir davon aus, dass der Tagger auf ebensolchen Daten Ergebnisse erzielt, die ihn für darauf aufbauende Forschungen einsetzbar machen.

Daneben haben wir gezeigt, dass der Vorverarbeitungsschritt der Disambiguierung keineswegs perfekt funktionieren muss, um mit den Daten weiterzuarbeiten. Die Accuracy von ca. 87% für die Disambiguierung führt zwar bei gleicher Datenmenge zu einem Verlust an Tagging-Performanz, durch die größere, automatisch vorverarbeitete Datenmenge wird dieser aber mehr als aufgefangen.

Um die Nutzung unserer Forschungsergebnisse nachhaltig zu ermöglichen, stellen wir das Modell sowohl auf der TreeTagger-Webseite[6] als auch über eine Webanwendung[7] zur Verfügung. Des Weiteren ist das Modell als Ressource ins Clarin-D-Repositorium[8] aufgenommen, wodurch die Metadaten sowie die Links zum Modell permanent auffindbar bleiben.

Literatur

Barteld, Fabian / Schröder, Ingrid / Zinsmeister, Heike (2015): "Unsupervised regularization of historical texts for POS tagging", in: Proceedings of the 4th Workshop on Corpus-based Research in the Humanities (CRH) 3–12 www.slm.uni-hamburg.de/germanistik/personen/zinsmeister/downloads/barteld-etai-2015.pdf [letzter Zugriff 23.08.2016].

Cohen, Jacob (1960): "A coefficient of agreement for nominal scales", in: Educational and Psychological Measurement 20: 37–46.

Dipper, Stefanie (2015): "Annotierte Korpora für die Historische Syntaxforschung. Anwendungsbeispiele anhand des Referenzkorpus Mittelhochdeutsch", in: Zeitschrift für Germanistische Linguistik 43: 516–563.

Dipper, Stefanie (2011): "Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison", in: Journal for Language Technology and Computational Linguistics 26: 25–37 (= Proceedings of the TLT-Workshop on Annotation of Corpora for Research in the Humanities 2012) www.jlcl.org/2011_Heft2/2.pdf [letzter Zugriff 23.08.2016].

Dipper, Stefanie / Donhauser, Karin / Klein, Thomas / Linde, Sonja / Müller, Stefan / Wegera, Klaus-Peter (2013): "HiTS: ein Tagset für historische Sprachstufen des Deutschen", in: Journal for Language Technology and Computational Linguistics 28: 85–137 www.jlcl.org/2013_Heft1/5Dipper.pdf [letzter Zugriff 23.08.2016].

Giesbrecht, Eugenie / Evert, Stefan (2009): "Is Part-of-speech tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus", in: Proceedings of the 5th Web as Corpus Workshop (WAC5) www.stefan-evert.de/PUB/GiesbrechtEvert2009_Tagging.pdf [letzter Zugriff 23.08.2016].

[6] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

[7] www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/PoS_Tag_MHG.html.

[8] Die Ressource kann im IMS-Repository gefunden werden: http://clarin04.ims.uni-stuttgart.de/fedora/objects/clarind-ims:92/datastreams/CMDI/content.

Mittelhochdeutsche Begriffsdatenbank (MHDBDB). Universität Salzburg. Koordination: Margarete Springeth. Technische Leitung: Nikolaus Morocutti/Daniel Schlager. 1992-2016 http://mhdbdb.sbg.ac.at/ [letzter Zugriff 23.08.2016].

Nivre, Joakim / de Marneffe, Marie-Catherine / Ginter, Filip / Goldberg, Yoav / Hajič, Jan / Manning, Christopher D. / Mc Donald, Ryan / Petrov, Slav / Pyysalo, Sampo / Silveira, Natalia / Tsarfaty, Reut / Zeman, Daniel (2016): "Universal Dependencies v1: A Multilingual Treebank Collection", in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) 1659–1666 www.petrovi.de/data/lrec16.pdf [letzter Zugriff 23.08.2016].

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): "Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)", Universität Stuttgart / Tübingen www.sfs.uni-tuebingen.de/resources/stts-1999.pdf [letzter Zugriff 23.08.2016].

Schmid, Helmut (1995): "Improvements in Part-of-Speech Tagging with an Application to German", in: Proceedings of the ACL SIGDAT-Workshop 47–50 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf [letzter Zugriff 23.08.2016].

Schulz, Sarah / Kuhn, Jonas (2016): "Learning from Within? Comparing PoS Tagging Approaches for Historical Text", in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) 4316–4322 www.lrec-conf.org/proceedings/lrec2016/pdf/1237_Paper.pdf [letzter Zugriff 23.08.2016].