Überblick über die Aufgabe des POS-Tagging und bestehende Lösungen
Seminararbeit zum Thema Sprachverarbeitung

Tobias Knöppler
Fachbereich Informatik, Uni Hamburg
tobias@knoeppler.net

Eingereicht am 10.08.2016

Abstract

Ich stelle in dieser Arbeit die grundlegende Aufgabe von Part-Of-Speech-Tagging und die allgemeine Herangehensweise, sowie einiger konkreter Beispiele dar. Diese Arbeit soll nicht dazu dienen, konkrete POS-Tagging-Verfahren in ihrer Tiefe zu erklären, sondern das die Problemstellung einordnen zu können.

1. Einleitung

POS-Tagging ("Part-Of-Speech-Tagging") bezeichnet das Versehen der Worte innerhalb eines Textes mit den zugehörigen Wortarten. Diese Informationen werden an die Worte in Form von sogenannten Tags angefügt, daher der Name. Als Grundlage für die Entscheidung, zu welcher Wortart ein Wort gehört, werden Wörterbücher, Syntax, der Kontext und statistische Analysen herangezogen. Ist ein Wort erst einmal mit der zugehörigen Wortart versehen, kann dies für vielfältige weitere Aufgaben im Bereich des Natural Language Processing nützlich sein - so lässt sich zum Beispiel in einigen Fällen eine semantische Mehrdeutigkeit durch Kenntnis der Wortart auflösen.

2. Verfahren

2.1. Text-Korpora

Der Schwierigkeitsgrad, ebenso wie der Nutzen von POS-Tagging wird sehr stark durch das verwendete Tag-Set bestimmt. Die bereits aus der Schule bekannten Wortarten werden hier in der Regel wesentlich genauer spezifiert, um einen größeren Nutzen aus den Tags ziehen zu können.

Hierzu sind Textsammlungen (Textcorpora) nötig, die bereits mit Tags versehen wurden. Dabei existieren sehr homogene Korpora für spezielle Zwecke, ebenso wie möglichst heterogene Korpora. Außerdem bringen die meisten Korpora ihr eigenes Tag Set mit. Insgesamt ist es also für das POS Tagging essenziell, den richtigen Text-Korpus je nach Anwendungsfall zu wählen (oder zu erzeugen).

Historisch von Bedeutung ist hier der Brown Copus (für die englische Sprache), welcher etwa 1.000.000 Worte enthält. Diese wurden im Laufe der Zeit mit Tags versehen und diese Tags wurden mit der wachsenden Genauigkeit der Tagger und viel manueller Arbeit korrigiert, sodass mittlerweile für den Korpus eine nahezu hundertprozentige Genauigkeit besteht. Mittlerweile wurde der Brown Corpus jedoch für die meisten Anwendungen von einer Vielzahl jüngerer und umfangreicherer Kopora abgelöst, da die Menge an Worten im Brown Corpus für die Trainingsmethoden vieler Tagger schlichtweg unzureichend ist.

Ein prominenter Korpus für die englische Sprache ist beispielsweise der British National Corpus, der mit etwa 100 Millionen Worten - zusammengesetzt aus 90% geschriebenen Texten und 10% transkribierten gesprochenen Texten - versucht eine möglichst große Bandbreite an modernem Englisch abzudecken. Um es dennoch zu ermöglichen, den Hintergrund der Texte beim Taggen zu berücksichtigen, sind die Texte in "domain", "time" und "medium" eingeteilt, wobei "domain" die Textgattung, "time" die Zeit zu der der Text entstanden ist und "medium" das Transportmedium (z.B. Zeitungsartikel oder buch) bezeichnet.

Ein weiterer Korpus ist die Penn Treebank (PTB). Dieser Korpus entstand, indem Textsammlungen geparsed und im Nachhinein manuell korrigiert wurden. Er ist zusammengesetzt aus verschiedenen Textquellen, darunter der gesamte Brown Corpus, eine Sammlung von Telefonattransskripten, etwa einer Million Worten aus Zeitungstexten des Wall Street Journal, et cetera. Der gesamte Korpus ist mit POS-Tags versehen, während etwa zwei Drittel zusätzlich vollständig geparsed ist (d.h. es wurde ein Syntaxbaum für jeden Satz erzeugt).

Für die deutsche Sprache gibt es zum Beispiel den TIGER Corpus, der aus etwa 900.000 Worten aus schriftlichen Nachrichtentexten der Frankfurter Rundschau aufgebaut ist (8).

2.2. Entscheidungsgrundlagen

Oft lässt sich die Wortart direkt von einem Wort ableiten - ohne weitere Informationen zu berücksichtigen. Das ist dann der Fall, wenn das Wort in einem Wörterbuch enthalten (also bereits bekannt) ist und es nur eine einzige Wortart haben kann (d.h. keine Mehrdeutigkeit der Wortart besteht). Leider ist in den meisten kein ausreichend vollständiges Wörterbuch verfügbar. Die eigentliche Herausforderung bilden deshalb die übrigen Worte, also unbekannte Worte und solche, die mehreren Wortarten zugehörig sein können. Im deutschen ist die erste Kategorie besonders groß, da es eine unendliche Zahl valider, zusammengesetzter Worte gibt, was ein vollständiges Wörterbuch unmöglich macht. In der englischen Sprache spielen zusammengesetzte Worte zwar eine kleinere Rolle, aber dafür existieren zahlreiche Worte, die mehreren Wortarten angehören können, insbesondere, da Substantive sich nicht durch Großschreibung von anderen Wortarten abheben.

Um diese Fälle, in denen sich die Wortart auf triviale Weise entscheiden lässt, dennoch zu disambiguieren, bieten sich mehrere Ansätze.

Die naivste Methode ist, einfach jedem Wort die Wortart zuzuweisen, in der es statistisch am häufigsten auftritt. Das nützt zwar im Blick auf unbekannte Worte nichts, aber für mehrdeutige Wortarten verbessert diese Technik das Resultat deutlich. Zudem ist leicht einzusehen, dass auf diese Weise noch immer keine zufriedenstellende Genauigkeit erreicht werden kann; da gerade im Englischen viele Worte in mehr als einer Wortart jeweils häufig auftreten (z.b. fish).

Um unbekannte Worte klassifizieren zu können, können gewisse Eigenschaften der Worte als betrachtet werden, in dem Versuch, aus diesen eine Wahrscheinlichkeit für unterschiedliche Wortarten abzuleiten, beispielsweise die Anzahl von Ziffern oder Großbuchstaben im Wort oder ob es einen Bindestrich enthält. Mit handgeschriebenen Regeln oder Maschinenlernverfahren kann dies in die Entscheidung für eine Wortart miteinfließen.

Der nächste Schritt ist, den syntaktischen Kontext des Wortes mit in die Klassifizierung einzubeziehen. Ein simples Beispiel hierfür ist, dass ein Substantiv nach einem Artikel wesentlich wahrscheinlicher ist, als ein Verb. Die Betrachtung eines größeren Kontextes um das Wort herum führt jedoch in der Regel zu proportional größerem Rechenaufwand.

Dasselbe gilt in noch größerem Maße für das Einbeziehen des semantischen Kontextes. Dies erlaubt zwar einige Fälle zu entscheiden, in denen die Wortart rein syntaktisch nicht auflösbar wäre, aber der Aufwand hierfür ist so gewaltig, dass dies bei den meisten aktuellen POS-Taggern keine Rolle spielt.

Die genannten Merkmale bilden zusammen die Eigenschaften auf deren Grundlage die meisten POS-Tagger trainiert werden.

2.3. Genauigkeit

Die Genauigkeit eines POS-Taggers wird in der Regel in % der richtig eingeordneten Worte angegeben. Dieser Wert ist jedoch mit Vorsicht zu genießen, da erstens nicht immer klar ist, ob diese Trefferquote sich auf den bekannte oder unbekannte Worte bezieht und zweitens der Nutzen eines POS-Taggers für viele Anwendungen noch mehr von der Genauigkeit abhängt, mit der er ganze Sätze vollständig tagged.

Es bietet sich an, einen Tagger, der nur die Häufigkeitsverteilung eines Wortes auf die Wortklassen betrachtet (unigram-most-likely), als Grundlage für die Untergrenze der angestrebten Genauigkeit zu wählen (erstmals vorgeschlagen von Gale u. a. (1992)).

Die Genauigkeit dieses Verfahrens wurde (nach Charniak (1993)) bei 90-91% korrekt getaggter Worte ermittelt.

Als Obergrenze der erreichbaren Genauigkeit wird üblicherweise der Gold Standard verwendet, in diesem Fall die Übereinstimmung menschlicher Experten. Diese beläuft sich auf etwa 96-97% (nach Marcus u. a. (1993)). (Voutilainen, 1995) hat jedoch herausgefunden, dass 100% Übereinstimmung erreichbar sind, wenn die Experten sich absprechen. Daraus lässt sich ableiten, dass die Abweichungen nicht durch Mehrdeutigkeiten in der Sprache, sondern vielmehr durch Fehler der Experten zustande gekommen waren. Dies ist wichtig, da einige der heutigen POS-Tagger Genauigkeiten von 97% erreichen, womit der Gold Standard ansonsten bereits erreicht wäre.

2.4. Ansätze

Es existieren zahllose Ansätze und Implementationen für POS-Tagger, sodass es schwierig ist, im Zuge dieses Artikels einen oder zumindest einige wenige herauszuheben, zumal selbst noch dann bei weitem zu viele Verfahren übrig bleiben, wenn man nur diejenigen mit der höchsten Genauigkeit betrachtet. Deshalb ist die folgende Auswahl von POS-Taggern einigermaßen willkürlich zusammengestellt, mit dem Ziel einen groben Einblick zu bieten. Auf die genaue Funktionsweise einzugehen würde den Rahmen dieser Arbeit jedoch sprengen, deshalb begnüge ich mich mit einigen wichtigen Eigenschaften, die der allgemeinen Einordnung der Verfahren dienen sollen. Die Daten zur Genauigkeit der Implementationen sind auf dem Teil der Penn Treebank, welcher auf Texten des *Wall Street Journal* beruht, gemessen und der ACL-Wiki[1] entnommen.

[1] https://www.aclweb.org/aclwiki/index.php?title=POS*Tagging*(State_of_the_art)#WSJ

2.4.1. Hidden Markov Model (HMM)

Einige der frühen POS-Tagger basieren auf Hidden Markov Modellen. Bei diesem Verfahren wird jeweils ein nach Anzahl Worten begrenzter Kontext vor und/oder nach dem zu klassifizierenden Wort betrachtet und auf dieser Grundlage eine statistische Wahrscheinlichkeit für die jeweiligen Wortarten ermittelt. Verwendet man für die Länge dieses Kontexts 1 (nur das betrachtete Wort), so erhält man eben das Modell, welches die Untergrenze der erstrebten Genauigkeit bildet (vlgl. 2.3). Das Betrachten sehr großer Kontexte hingegen scheitert am Rechenaufwand, welcher exponentiell mit der Größe des Kontexts wächst. Ein POS-Tagger, der auf einem Hidden Markov Modell beruht, ist beispielsweise der TnT-Tagger (Brants, 2000), der eine Genauigkeit von 96,46% erreicht.

2.4.2. Neuronale Netze

Zahlreiche POS-Tagger - darunter einige mit Genauigkeiten über 97% -beruhen auf verschiedenen Arten neuronaler Netze. Dabei sind diverse Varianten eines einfachen Perzeptrons ebenso vertreten, wie komplexere Netze; etwa Long-Short-Term-Memory-Netzwerke. Allen gemeinsam ist das grundlegende Verfahren: Die Wort wird mit bestimmten, ausgesuchten Eigenschaften kodiert und das neuronale Netz kombiniert die (in der Regel zahlreichen) Eigenschaften der einzelnen Worte solcherart zu Metaeigenschaften, dass diese möglichst viel über die Wortart des Wortes aussagen. Bevor das Neuronale Netz dies effizient tun kann, muss es mit bekannten Worten und ihren dazugehörigen Wortarten trainiert werden. Damit ist dies ein Fall von überwachtem Lernen. Wie bei allen Maschinenlern-Algorithmen ist bei diesem Verfahren eine Sorgfältige Auswahl der verwendeten Eigenschaften (Features) der Worte und ein gutes Preprocessing derselben essentiell. In einigen Fällen stellt das Neuronale Netz gar selbst nur einen weiteren Preprocessing-Schritt dar, der die Dimensionalität der Wort-Feature-Vektoren verringert (indem ein kleines Perceptronen-Layer auf ein großes folgt) und dabei möglichst viel vom Informationsgehalt der Daten bewahrt. Danach folgt dann der eigentliche Klassifizierer. Die besten Ergebnisse der auf Neuronalen Netzen basierenden POS-Tagger (namentlich das Bidirectional LSTM-CRF Model von Huang u. a. (2015)) erreichen eine Genauigkeit von 97,55% und gehören damit auch insgesamt zu den besten verfügbaren POS-Taggern.

2.4.3. Nearest Neighbour Algorithmus

Dieser Algorithmus beruht auf der simplen Idee, einem unbekannten Wort (das wieder als Vektor von Eigenschaften/Features angegeben ist) die Wortart des ähnlichsten, bereits bekannten Wortes (des "nächsten Nachbarn) zuzuweisen. In einer etwas verbesserten Form des Verfahrens werden aus den zur Klassifizierung verwendeten, bekannten Worten diejenigen aussortiert, die für die Einordnung keines der Worte in der Trainingsmenge relevant waren. In einem weiteren Optimierungsschritt, welcher in unüberwachtem Lernen besteht, werden aus einer unklassifizierten Wortmenge all jene Worte zum reduzierten Klassifizierer hinzugefügt, die mit diesem signifikant schlechter als mit dem ursprünglichen Nearest Neighbour Klassifizierer zugeordnet werden können. In dieser Form erreicht der Tagger eine Genauigkeit von 97,50%. Er wurde von Søgaard (2011) entworfen.

3. Fazit

Aufgrund der erreichten Trefferquoten aktueller POS-Tagger wird POS-Tagging oft als gelöstes Problem betrachtet. Dagegen ist jedoch einzuwenden, das

1. eine Trefferquote von 97,50% noch immer bedeutet, dass beispielsweise ein Satz mit 12 Worten nur zu 73,79% vollständig richtig disambiguiert wird (der tatsächliche Wert dürfte noch niedriger sein, da es in einem langen Satz schwieriger ist, einzelne Worte zu klassifizieren, als in einem kurzen, da der Kontext komplexer wird) und
2. es viele Gebiete gibt, in denen die Datenlage (d.h. die Verfügbarkeit guter Textkorpora) den Einsatz vieler POS-Tagger nur begrenzt zulässt oder zumindest ihre Genauigkeit schmälert. Dazu zählen sowohl viele Sprachen, zu denen es wenige Daten gibt, als auch speziellere Textsorten, wie

   - gesprochene Sprache
   - Chatnachrichten
   - etc.

   Für diese Textsorten werden mitunter deutlich schlechtere Ergebnisse erziehlt.

Aus diesem Grund halte ich es für voreilig, POS-Tagging als vollends gelöstes Problem zu betrachten, auch wenn die aktuellen Resultate bereits viele Zwecken genügen.

References

Thorsten Brants. "TnT – Statistical Part-of-Speech Tagging". 2000. http://www.coli.uni-saarland.de/~thorsten/tnt/.

Eugene Charniak. "Statistical Language Learning". 1993.

William Gale, Kenneth Ward Church, und Yarowsky David. "Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs". 1992.

Zhiheng Huang, Wei Xu, und Kai Yu. “Bidirectional LSTM-CRF Models for Sequence Tagging”. 2015.

Mitchell P. Marcus, Beatrice Santorini, und Mary Ann Marcinkiewicz. “Building a Large Annotated Corpus of English: The Penn Treebank”. 1993.

Anders Søgaard. “Semi-Supervised Condensed Nearest Neighbor for Part-of-Speech Tagging”. 2011.

Atro Voutilainen. 1995.

http://www.ims.uni-stuttgart.de/.