

# Part-of-Speech-Tagging am Beispiel der deutschen Sprache

Stefan Bienk, Sebastian Schmidt

**Part-of-Speech-Tagging ist eine grundlegende Problemstellung der automatischen Sprachverarbeitung. Anhand von Beispielen aus der deutschen Sprache werden die wesentlichen Herausforderungen, die es bei der Entwicklung von Tagging-Algorithmen zu überwinden gilt, vorgestellt sowie ein Überblick der verschiedenen Lösungsvorschläge aus der Literatur gegeben. Darüber hinaus wird beschrieben, wie zwei dieser Ansätze am Erlanger KI-Lehrstuhl in einem zum Download angebotenen Prototypen kombiniert wurden, dessen hohe Performanz schließlich durch die Ergebnisse einer experimentellen Evaluierung belegt wird.**

## 1 Einleitung und Überblick

Das Problem des „Part-of-Speech Tagging“, kurz POS-Tagging, entspringt der KI-Teildisziplin der *Automatischen Verarbeitung geschriebener Sprache* und bezeichnet das Problem, die Wörter eines natürlichsprachlichen Textes in syntaktische Kategorien (Hauptwort, Adjektiv, Präposition, etc.) einzuordnen. Die Herausforderung besteht dabei in der Tatsache, dass ein und dasselbe Wort (reduziert auf seine Repräsentation als Zeichenkette) in Abhängigkeit des umgebenden Satzkontextes möglicherweise in verschiedene Kategorien einzuordnen ist. Der Detaillierungsgrad des festgelegten Tagsets (Menge aller möglichen Kategorien) beeinflusst dabei die Häufigkeit und den Grad von Mehrdeutigkeiten. Sinnvoll erscheint nur die Definition von Tagsets, welche in syntaktisch korrekten Sätzen die eindeutige Klassifizierung jedes Wortes erlauben. Ein linguistisch fundiertes Tagset für die deutsche Sprache liegt z.B. mit dem Stuttgart-Tübingen Tagset (STTS, vgl. [6]) vor. Das STTS umfasst insgesamt 54 Wortkategorien, die als Blätter eines Baums betrachtet werden können, welcher die (rekursive) Verfeinerung der folgenden 11 Hauptkategorien abbildet:

Nomina (N)	Adverbien (ADV)
Verben (V)	Konjunktionen (KO)
Artikel (ART)	Adpositionen (AP)
Adjektive (ADJ)	Interjektionen (ITJ)
Pronomina (P)	Partikeln (PTK)
Kardinalzahlen (CARD)	

Die für das Verständnis der folgenden Beispiele wichtige Kategorie der Verben zerfällt beispielsweise in die Unterklassen Vollverben (VV), Modalverben (VM) und Hilfsverben (VA). Die Klasse der Vollverben wird schließlich u.a. durch die Blattkategorien VVFIN (Finite Vollverbformen), VVINFIN (Vollverbinfinitive) und VVPP (Vollverbpartizipien) verfeinert. Analog dazu existieren Blattkategorien VMFIN, VMINFIN, VMPP sowie VAFIN, VAIFIN und VAPP.

Die folgenden zwei Beispielsätze zeigen, wie zwei verschiedene Kontexte die unterschiedliche Klassifizierung des Wortes „gehen“ im Sinne des STTS bedingen:

Vorzeitig **gehen** \VVINF will er sicherlich nicht.  
Jetzt **gehen** \VFIN wir.

Im ersten Beispiel ist das Wort „gehen“ als Vollverbinfinitiv auszuzeichnen, im zweiten Fall als finite Vollverbform.

Die durch die Annotation eines Textes mit POS-Tags gewonnene Metainformation ist für die weitergehende syntaktische Analyse bzw. bereits inhaltliche Verarbeitung von natürlicher Sprache von Bedeutung; der nächste Abschnitt belegt dies anhand von Anwendungsbeispielen für POS-Tagger. Darauf folgt ein Überblick über in der Literatur publizierte Ansätze zum automatischen POS-Tagging. Im Hauptteil des Artikels wird beschrieben, wie am Erlanger KI-Lehrstuhl zwei dieser Ansätze zu einer im Hinblick auf die Fehlerquote hochperformanten Lösung kombiniert wurden.

## 2 Anwendungen

Kritiker könnten einräumen, dass der Output eines POS-Tagger als „flache“ Folge von Wort/Tag-Paaren lediglich eine Traversierung der beiden untersten Ebenen eines Syntaxbaums für den Eingabesatz darstellt, der von einem syntaktischen Parser berechnet werden kann. Die Betrachtung des Tagging-Problems wäre somit nicht von praktischem Nutzen.

In Wahrheit ist mit diesem Einwand aber bereits ein Anwendungsfall für POS-Tagger identifiziert: Diese werden nämlich als Vorverarbeitungsstufe für syntaktische Parser eingesetzt. Der Parser muss in diesem Fall nur noch Wortkategorien als Abstraktion von konkreten Wörtern „kennen“, was zu einer Verringerung seiner Komplexität führt.

Darüber hinaus gibt es eine bedeutende Anzahl von Sprachverarbeitungsanwendungen, wo die Verwendung eines POS-Tagger *anstatt* eines Parsers von Vorteil oder sogar unumgänglich ist. Dies liegt zum Einen daran, dass die Laufzeitkomplexität von Tagging-Algorithmen niedriger als diejenige von syntaktischen Parsern ist. Zum Anderen sind die von POS-Tagger verwendeten Sprachmodelle wesentlich „billiger“ in der Erstellung als die von Parsern benötigten Grammatiken: Durch Anwendung von Techniken des maschinellen Lernens sind moderne Tagger in der Lage, ihre Sprachmodelle automatisch aus Trainingsdaten zu akquirieren, während Grammatiken mit erheblichem manuellen Aufwand und Investition von linguistischem Wissen erstellt werden müssen. Letzteres Argument ist vor allem für die Betrachtung von „exotischen“ Sprachen von Bedeutung, für die unter Umständen noch keine maschinenlesbaren Grammatiken existieren.

POS-Tagging gilt aufgrund seiner breiten Anwendbarkeit als Standardproblem in der Sprachverarbeitung. Die Disam-