Part-of-Speech-Tagger

Elke Smith und Nilay Karagözoglu

Überblick

1. Definition -Was ist POS-Tagging?
2. Stuttgart-Tübingen-Tag-Sets
3. Wozu benötigt man POS-Tagging?
4. Verfahren

4.2. Tagging mit Supervised Learning

4.3. Tagging mit Unsupervised Learning

5. Vorstellung des Stanford Log-linear Part-Of-Speech-Tagger

POS- Tagger


Text wird in Sätze zerlegt
Wort wird einer Wortkategorie zugeordnet
Informationen werden gewonnen

Definition

- Part-of-Speech-Tagging ist ein maschineller Vorverarbeitungsschritt, um Informationen aus Texten im Internet herauszulesen und zu filtern.
- Texte werden analysiert und in Sätze zerlegt. Die Wörter in den Sätzen, werden den richtigen Wortkategorien zugeordnet.
- Für die Wortarten-Zuordnung (das Taggen) gibt es verschiedene Tagsets.
- Je nach Sprache kann eine Einteilung der Wörter in unterschiedlich viele Klassen erfolgen.

Stuttgart-Tübingen-Tag-Set

- Das Standard-Tagset für die deutsche Sprache ist das Stuttgart-Tübingen-Tag-Set (STTS) mit 54 POS-Tags
- Es gibt verschiedene Hauptkategorien wie z.B. Substantiv, Verb, Adverb, Präposition.
- Die Hauptkategorien werden noch weiter unterteilt z.B. in Imperativ, Infinitiv oder Partizipien.

ADJA	attributives Adjektiv	PDS	subst. Demonstrativpron.	PTKVZ	abgetrennter Verbzusatz
ADJD	adverb./ prädikat. Adj.	PDAT	attr. Demonstativpron.	PTKANT	Antwortpartikel
ADV	Adverb	PIS	subst. Indefinitpron.	PTKA	Partikel bei Adj./Adv.
APPR	Präposition	PIAT	attr. Indef.pron. o. Determ.	TRUNC	Kompositions-Erstglied
APPRART	Präp. mit Artikel	PIDAT	attr.Indef.pron.m.Determ.	VVFIN	finites Verb, voll
APPO	Postposition	PPER	irreflex. Personalpron.	VVIMP	Imperativ, voll
APZR	Zirkumposition rechts	PPOSS	subst. Possessivpron.	VVINF	Infinitiv, voll
ART	best. oder unbest. Artikel	PPOSAT	attr. Possessivpron.	VVIZO	Infinitiv mit „zu“, voll
CARD	Kardinalzahl	PRELS	subst. Relativpron.	VVPP	Partizip Perfekt, voll
FM	fremdspr. Material	PRELAT	attr. Relativpron.	VAFIN	finites Verb, aux
ITJ	Interjektion	PRF	reflex. Relativpron.	VAIMP	Imperativ, aux
KOUI	uo. Konj. mit „zu“ und Inf.	PWS	subst. Interrogativpron.	VAINF	Infinitiv, aux
KOUS	uo. Konj. mit Satz	PWAT	attr. Interrogativpron.	VAPP	Partizip Perfekt, aux
KON	nebenordn. Konjunktion	PWAV	adv. Interr.-/Rel.pron.	VMFIN	finites Verb, modal
KOKOM	Vergleichskonjunktion	PAV	Pronominaladverb	VMINF	Infinitiv, modal
NN	normales Nomen	PTKZU	„zu“ vor Infinitiv	VMPP	Partizip Perfekt, modal
NE	Eigennamen	PTKNEG	Negationspartikel	XY	Nichtwort, Sonderzeich. enth.
\$,	Komma	\$.	satzbeend. Interpunktion	\$(	sonst. Satzzeichen, satzintern

Stuttgart-Tübingen-Tag-Set

- Früher geschah die Zuordnung manuell, mittlerweile ist der Vorgang automatisiert.
- Mehrdeutigkeiten können auftreten, die von den Taggern nicht immer korrekt aufgefasst werden können.
- Einem Wort können mehrere Wortarten zugeordnet werden.
- Der syntaktische und semantische Kontext muss hinzugezogen werden.

Beispielsatz:

Satz:	Sie	haben	meinen	Freunden	geholfen
STTS-Tag:	PPER	VAFIN	PPOSAT	NN	VVPP
STTS-Tag:		VAFIN	VVFIN		
STTS-Tag:		VVINF	VVINF		

- „haben“ ist in diesem Fall ein Hilfsverb .
- „haben“ kann ein Verb im Infinitiv sein.

- „meinen“ ist in diesem Fall ein attribuierendes Possessivpronomen.
- „meinen“ kann ein Verb in der dritten Person Plural sein.
- „meinen“ kann ein Verb im Infinitiv sein.

- Fremdwörter, Umgangssprache oder seltene Satzkonstruktionen können das Taggen erschweren.

Wozu benötigt man POS-Tagger?

- Erleichterung der Mensch-Maschine-Kommunikation
- POS-Tagging ist ein Teil von Natural Language Processing
- Analysieren von Wörtern kann helfen, Informationen aus Texten herauszulesen
- Suchmaschinen und automatischen Übersetzungssystemen die Arbeit erleichtern
- POS-Tagging ist grundlegend für maschinelles Lernen von Sprache und die automatisierte Sprach- und Informationsverarbeitung

Quellenangabe

- https://books.google.de/books?id=73WDoBCvQBkC&pg=PA39&lpg=PA39&dq=regelbasiert+p&hl=de#v=onepage&q&f=false
- https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/germantagsets/#id-cfcbf0a7-0
- https://de.wikipedia.org/wiki/Part-of-speech-Tagging
- https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/STTS_Tagset_Tiger
- Carstensen, Kai-Uwe - 3., überarb. und erw. Aufl., 2010 „ Computerlinguistik und Sprachtechnologie : eine Einführung“

Part-of-speech-Tagging

Nilay Karagözoglu, Elke Smith

15.12.2020

INHALTE

Verfahren

VERFAHREN

- Manuelles Taggen (früher)

  . . .

- Automatisiertes Taggen
  - *Supervised Learning*
    (überwachtes maschinelles Lernen)
  - *Unsupervised Learning*
    (unüberwachtes maschinelles Lernen)
  - *Weak Supervision*
    (u.a. mittels Verwendung von *Parallel Text* oder *Tag Dictionaries*)

VERFAHREN

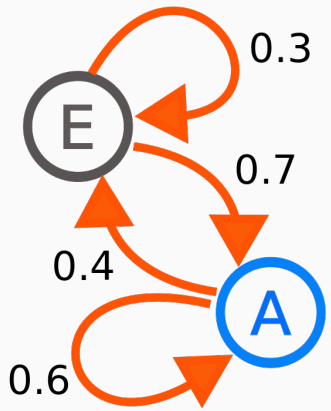
– SUPERVISED LEARNING

- Alle Tags sind in einem zuvor erstellten Tagset definiert
- Das Lernen der Tag-Wahrscheinlichkeiten erfolgt stets ausgehend von einem Textkorpus
- Zum Einsatz kommen u.a. Verfahren wie Entscheidungsbäume und *Hidden Markov Models*

VERFAHREN

– HIDDEN MARKOV MODELS

- Eine Markowkette ist ein stochastisches Modell
- Modellierung eines Systems als Markowkette (*Markov Chain*)



img source: `https://en.wikipedia.org/wiki/Markov_chain`

VERFAHREN

– HIDDEN MARKOV MODELS

$$p(x, y) = \prod_{i=1}^{length(x)} p_t(y_i|y_{i-1}, y_{y-2}) p_o(x_i|y_i) \quad (1)$$

$p(x, y)$ Wahrscheinlichkeit eines Satzes x und einer bestimmten *Hidden State Sequence* y

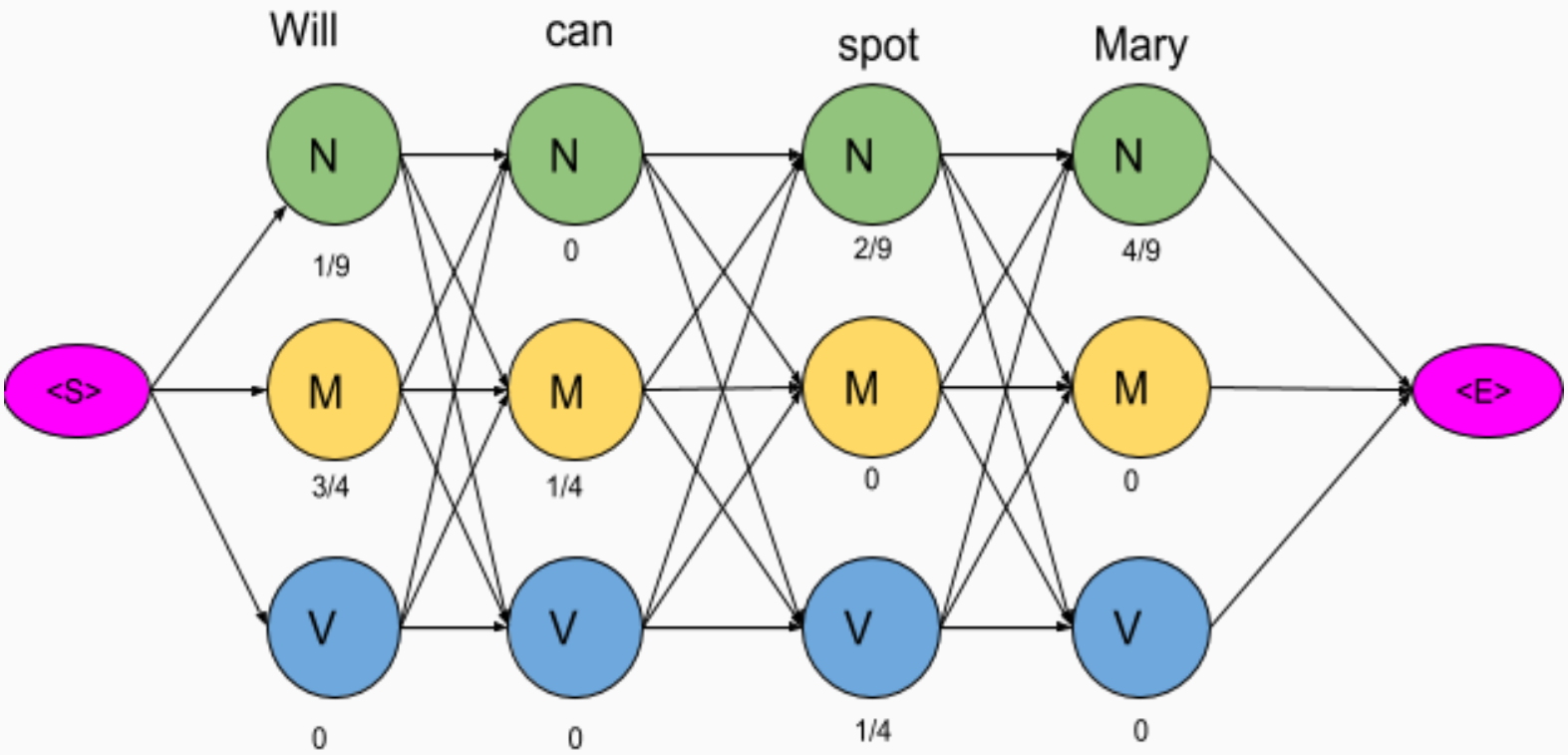
$p_o(x_i|y_i)$ Wahrscheinlichkeit, Wort x_i in Zustand y_i zu beobachten (Emissionswahrscheinlichkeit)

$p_t(y_i|y_{i-1}, y_{i-2})$ Wahrscheinlichkeit, in Zustand y_i zu sein, gegeben den beiden vorherigen Zuständen y_{i-1}, y_{i-2} (Übergangswahrscheinlichkeit, SHMM)

Li, Graça, Taskar (2012). Wiki-ly Supervised Part-of-Speech Tagging.

VERFAHREN

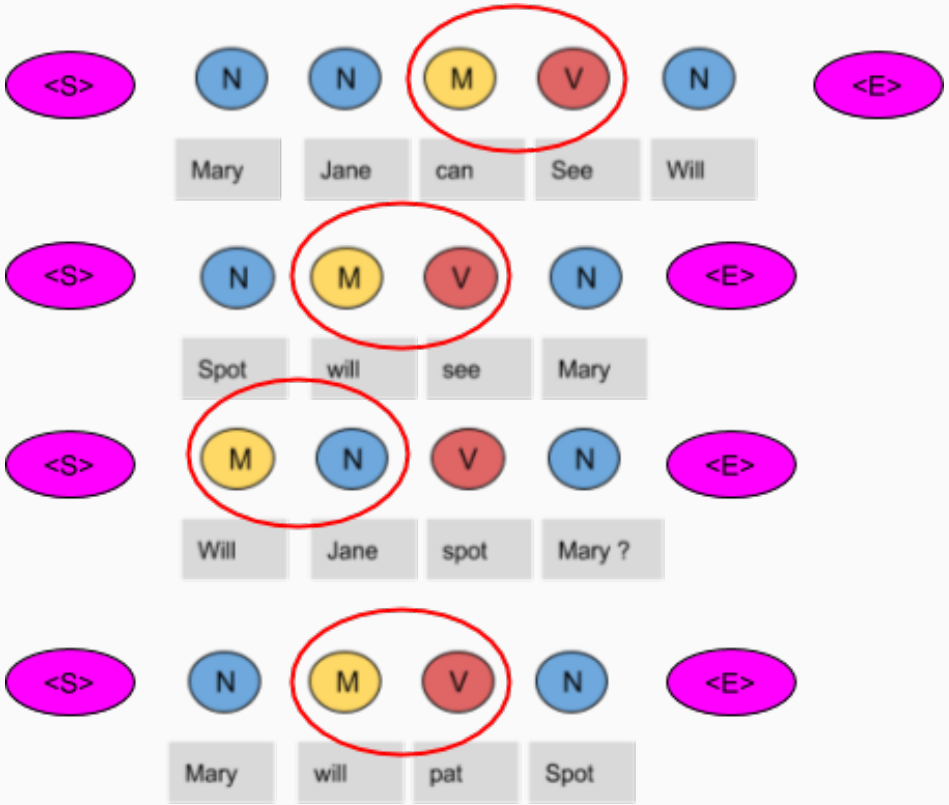
– HIDDEN MARKOV MODELS



img source: https://www.mygreatlearning.com/blog/pos-tagging/

VERFAHREN

– HIDDEN MARKOV MODELS



img source: https://www.mygreatlearning.com/blog/pos-tagging/

VERFAHREN

– UNSUPERVISED LEARNING

- Das Tagset wird nicht zuvor definiert, für das Training wird ein ungetaggter Textkorpus verwendet
- Mittels stochastischer Verfahren werden Muster, und daraus Kategorien, erschlossen
  - Artikel weisen z.B. andere statistische Eigenschaften auf als Verben, und Artikel treten in ähnlichen Kontexten auf
  - Durch viele Iterationen werden Klassen von ähnlichen Wortarten erschlossen

GEGENÜBERSTELLUNG

– SUPERVISED VS. UNSUPERVISED LEARNING

- Für manche Sprachen liegen keine (oder nicht ausreichend) gelabelten Korpora vor
  - \+ Unsupervised Learning
- Annotation ist teuer und aufwendig
  - \+ Unsupervised Learning
- Unsupervised Tagging liefert meist eine geringere Accuracy, Performanz oft nicht ausreichend für praktische Zwecke, Accuracy von Tagging mittels Supervised Learning i.d. Regel sehr hoch (Shen et al. 2007; Toutanova et al., 2003)
  - \+ Supervised Learning

Beispiel

STANFORD LOG-LINEAR PART-OF-SPEECH TAGGER

https://nlp.stanford.edu/software/tagger.shtml

The Stanford Natural Language Processing Group people publications research blog software teaching join local

Software > Stanford Log-linear Part-Of-Speech Tagger

About | Questions | Mailing lists | Download | Extensions | Release history | FAQ

About

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. This software is a Java implementation of the log-linear part-of-speech taggers described in these papers (if citing just one paper, cite the 2003 one):

> Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.

> Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

The tagger was originally written by Kristina Toutanova. Since that time, Dan Klein, Christopher Manning, William Morgan, Anna Rafferty, Michel Galley, and John Bauer have improved its speed, performance, usability, and support for other languages.

The system requires Java 8+ to be installed. Depending on whether you're running 32 or 64 bit Java and the complexity of the tagger model, you'll need somewhere between 60 and 200 MB of memory to run a trained tagger (i.e., you may need to give Java an option like `java -mx200m`). Plenty of memory is needed to train a tagger. It again depends on the complexity of the model but at least 1GB is usually needed, often more.

Current downloads contain three trained tagger models for English, two each for Chinese and Arabic, and one each for French, German, and Spanish. The tagger can be retrained on any language, given POS-annotated training text for the language.

STANFORD LOG-LINEAR PART-OF-SPEECH TAGGER

- In Java implementierter POS-Tagger
- Verfügbar für verschiedene Sprachen (u.a. Englisch, Deutsch, Spanisch)
- Package enthält GUI-Demo, Command-Line-Interface und Java API
- Der Tagger kann neu trainiert werden
- Der englische Tagger nutzt das Penn Treebank Tag Set `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.htmll`

Quellenangabe und weiterführende Literatur

Li, S., Graça, J. V., & Taskar, B. (2012, July). Wiki-ly supervised part-of-speech tagging. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 1389-1398). Association for Computational Linguistics.

https://en.wikipedia.org/wiki/Markov_chain

https://www.mygreatlearning.com/blog/pos-tagging/

https://nlp.stanford.edu/software/tagger.shtml

www.cs.columbia.edu/~mcollins/fall2014-loglineartaggers.pdf