Toolbeitrag: Stanford Named Entity Recognizer

Mareike Schumacher [1]

1. Universität Regensburg

Thema:	Named Entity Recognition	DOI:	10.48694/fortext.3767
Jahrgang:	1	Ausgabe:	9
Erscheinungsdatum:	30-10-2024	Erstveröffentlichung:	2018-09-20 auf fortext.net
Lizenz:	CC BY SA		open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

Abb. 1: Der Workflow des Stanford-NER: Textdatei im TXT-Format und sprachspezifische Classifier im ZIP-Format über die grafische Nutzeroberfläche laden, dann die NER durchführen und die Ergebnisse direkt im Tool anschauen oder als TXT-Datei mit HTML-Tags herunterladen.

- **Systemanforderungen:** Läuft auf Windows und Mac, benötigt aktuelle Java-Version
- **Stand der Entwicklung:** Seit der Erstveröffentlichung 2006 laufend aktualisiert und für weitere Sprachen angepasst
- **Herausgeber:** Stanford Natural Languages Processing Group
- **Lizenz:** Open Source Tool, das kostenfrei genutzt werden kann
- **Weblink:** https://nlp.stanford.edu/software/CRF-NER.html#About
- **Im- und Export:** Import einzelner Texte als TXT-Datei (vgl. Reintext-Version), Export als TXT-Datei mit HTML-Tags
- **Sprachen:** Deutsch, Englisch, Spanisch, Chinesisch, Italienisch, Ungarisch

1. Für welche Fragestellungen kann Stanford-NER eingesetzt werden?

Mit Stanford-NER können vor allem Fragen nach quantitativen Aspekten von Figurennamen, Orten und Organisationen bearbeitet werden (Schumacher 2024). Dazu gehören Fragen wie: Wie viele Figuren werden in einem Text benannt? Welche Figuren werden am häufigsten erwähnt? Wie ist die Verteilung von Ortsnennungen im Text? Welche Orte werden erwähnt? In welchem Kontext werden Organisationen genannt?

2. Welche Funktionalitäten bietet Stanford-NER und wie zuverlässig ist das Tool?

Funktion: Eigennamenerkennung (vgl. Named Entities) in Texten zahlreicher Sprachen.

Zuverlässigkeit: Die höchste Zuverlässigkeit wird in Sachtexten erreicht. Hier liegt die Erkennungsquote für deutschsprachige Texte bei rund 70% F-Score (mehr zur Methode der Named Entity Recognition und ihren Qualitätskriterien finden Sie bei Schumacher (2024)). Auf der Stanford-NER-Homepage wird darauf hingewiesen, dass das deutsche Modell von 2018 erheblich besser ist, es werden aber keine genauen Zahlen genannt. Damit erreicht der Stanford-NER eine vergleichsweise hohe Zuverlässigkeit. Bei der Anwendung auf literarische Texte wird eine weit geringere Zuverlässigkeit erreicht. Diese kann allerdings durch die Anpassbarkeit (vgl. Machine Learning) des Tools erhöht werden.

3. Ist Stanford-NER für DH-Einsteiger*innen geeignet?

Checkliste	✓ / teilweise / –
Methodische Nähe zur traditionellen Literaturwissenschaft	✓
Grafische Benutzeroberfläche	✓
Intuitive Bedienbarkeit	✓
Leichter Einstieg	✓
Handbuch vorhanden	–
Handbuch aktuell	–
Tutorials vorhanden	teilweise
Erklärung von Fachbegriffen	–
Gibt es eine gute Nutzerbetreuung?	teilweise

Das Stanford-NER-Tool ist in seiner Grundfunktionalität sehr einsteigerfreundlich. Bisher wurde das Tool hauptsächlich in der Computerlinguistik eingesetzt. Um das Tool für die Literaturwissenschaft anzupassen, sind einige technische Grundkenntnisse vonnöten. Die Nutzerbetreuung findet hauptsächlich in der recht aktiven NLP-Community statt und kann je nach Frage in Schnelligkeit und Qualität der Antwort variieren.

4. Wie etabliert ist Stanford-NER in den (Literatur-)Wissenschaften?

Stanford-NER ist ein sehr gängiges computerlinguistisches Tool, das Gegenstand in zahlreichen Publikationen ist. In der Literaturwissenschaft ist es noch nicht etabliert, da eine Domäneadaption hier gerade erst beginnt.

5. Unterstützt Stanford-NER kollaboratives Arbeiten?

Stanford-NER ist ein Java-basiertes Desktop-Tool, das ohne weitere Installation offline über den eigenen PC ausgeführt wird. Kollaboratives Arbeiten wird dadurch nicht unterstützt.

6. Sind meine Daten beim Stanford-NER sicher?

Ja. Da es sich um ein desktopbasiertes Tool handelt, ist keine Anmeldung und/oder Angabe personenbezogener Daten für die Nutzung notwendig. Texte werden auf dem eigenen Rechner analysiert und Ergebnisse werden lokal gespeichert. Durch die Nutzung des Stanford-NER ergeben sich also keine datenschutz- oder urheberrechtlich bedenklichen Situationen.

Externe und weiterführende Links

- Stanford Named Entity Recognizer (NER): https://web.archive.org/save/https://nlp.stanford.edu/software/CRF-NER.html#About (Letzter Zugriff: 10.10.2024)

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

CSV CSV ist die englische Abkürzung für *Comma Separated Values*. Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv`, sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.

Domäneadaption Domäneadaption beschreibt die Anpassung einer in einem Fachgebiet entwickelten digitalen Methode an ein anderes Fachgebiet.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z. B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

NLP *Natural Language Processing* (NLP), maschinelle Sprachverarbeitung zu Deutsch, ist ein Teilgebiet der Linguistik, der Informatik und der künstlichen Intelligenz, welches sich damit beschäftigt, wie Computer so programmiert werden, dass sie große Mengen an natürlichsprachlichen Daten verarbeiten und analysieren können.

POS PoS steht für *Part of Speech*, oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Reintext-Version Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und CSV.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

Bibliographie

Schumacher, Mareike. 2024. Methodenbeitrag: Named Entity Recognition (NER). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 9. Named Entity Recognition (30. Oktober). doi: 10.48694/fortext.3765, https://fortext.net/routinen/methoden/named-entity-recognition-ner.