Maria Tabernig

Evaluation of Annotation Strategies

Bachelor's Thesis

to achieve the university degree of

Bachelor of Science

Bachelor's degree programme: Computer Science

submitted to

Graz University of Technology

Supervisor

Ass.Prof. Dipl.-Ing. Dr.techn. Roman Kern

Institute for Interactive Systems and Data Science
Head: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Ainet, Jänner 2021

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Dissertation identisch.

18.01.2021

Datum

	Unterzeichner	Maria Tabernig
	Datum/Zeit-UTC	2021-01-19T10:27:53Z
Hinweis	Diese qualifizierte elektronische Signatur ist gemäß § 4 Abs. 1 Signaturgesetz einer handschriftlich erstellten Unterschrift grundsätzlich rechtlich gleichgestellt. https://sign.sproof.io	

Unterschrift

Abstract

Verschiedene Forscher- und Entwicklerteams beschäftigen sich mit der Frage, wie man Maschinen die menschliche Sprache beibringen kann. Um Maschinen sprechen zu lernen, muss man sie mit großen Mengen an Daten füttern. Diese müssen für die Computer aufbereitet und mit Metadaten erweitert werden. Das Hinzufügen solcher Anmerkungen nennt man Annotation.

Die vorliegende Bachelorarbeit gibt einen Überblick über Annotation und den Vergleich von manuellen und semi-automatischen Ansätzen. Der erste Teil der Arbeit befasst sich mit dem theoretischen Hintergrund von Annotation und verschiedener Anwendungsgebiete. Einige Begriffe wie Natural Language Processing, Human Language Technologies und Distant Supervision werden erläutert und der Zusammenhang mit Annotation erklärt. Im Hintergrund-Kapitel werden auch verschiedene Tools und Techniken zur Annotation vorgestellt und ihre Anwendungsgebiete erläutert.

Im praktischen Teil der Arbeit wird ein Webtool, mit dem man Dokumente annotieren kann, vorgestellt. Das Ziel der Evaluation ist herauszufinden, welche der zwei Methoden des Webtools besser für Annotation geeignet ist. Es werden zwei Ansätze gegenüber gestellt, nämlich eine manuelle und eine semi-automatische. Methode 2 stellt Suchbegriffe in ihrem Kontext dar. Diese Begriffe können per Mausklick zur Annotation hinzugefügt oder verworfen werden. Um die Forschungsfrage zu beantworten, annotieren sieben Personen zwei wissenschaftliche Dokumente, einmal mit der manuellen, danach mit der semi-automatischen Methode. Außerdem muss jede Person einen Fragebogen zur Selbsteinschätzung ausfüllen. Das Ergebnis der Evaluation hat gezeigt, dass die manuelle Methode mehr Zeit erfordert, dafür aber ein genaueres Ergebnis liefert. Die zweite Methode ist schneller, dafür aber ungenauer. Es war zu beobachten, dass der Großteil der Probanden sich wesentlich schlechter einschätzte als ihr tatsächliches Können in Wirklichkeit war.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

Annotation ist eine der wichtigsten Grundlagen für Natural Language Processing (siehe 2.1) [23]. Natural Language Processing beschreibt im wesentlichen die maschinelle Verarbeitung von natürlicher Sprache. Das Ziel von NLP ist die Kommunikation von Mensch und Maschine mittels natürlicher Sprache zu ermöglichen [23]. Damit der Computer überhaupt Sprache verstehen kann müssen Dokumente und andere textbasierte Daten erst aufbereitet werden und mit Metadaten versehen werden. Da manuelle Annotation, also Annotation, die Personen händisch durchführen, sehr zeitaufwendig ist, wird häufig auf semi-automatische Annotation zurückgegriffen. Bei semi-automatischer Annotation wir mithilfe von maschinellem Lernen der Text annotiert. Eine Person muss dann nur mehr den Output überprüfen und gegebenenfalls korrigieren. Es gibt zahlreiche Annotation Tools, sei es in Form von Web-Tools oder als Computer-Anwendungen. Auch semi-automatische bzw. automatische Annotation wird von manchen dieser Anwendungen unterstützt. [14].

Diese Arbeit beschäftigt sich mit Annotation mithilfe von Software und Webtools. Im ersten Teil der Arbeit werden wichtige Grundlagen und Begriffe erläutert. Neben Natural Language Processing wird auch die Annotation im Allgemeinen mit einigen Techniken und wichtigen Begriffen vorgestellt. Außerdem wird der Unterschied zwischen manueller und semi-automatischer Annotation genauer erklärt. Weiters werden Brat und BioQRator, zwei Tools zur anwendungsbasierten Annotation, vorgestellt und deren Funktion erläutert.

Der zweiten Teil der Arbeit besteht aus einer Evaluation. Bei dieser Bewertung werden mithilfe eines Web-Tools Dokumente zuerst manuell annotiert und anschließend semi-automatisch. Ziel der Evaluation ist herauszufinden, welche Methode des Tools besser für die Annotation geeignet ist. 7 Personen annotieren anhand einiger festgelegter Regeln und einem genau definierten Modell wissenschaftliche Dokumente. Untersucht wird die verwendete Zeit und das Ergebnis der Annotation.

Auch auf den Vergleich von Selbsteinschätzung und tatsächlichem Resultat wird eingegangen.

2 Theoretischer Hintergrund

In diesem Kapitel werden einige wichtige Grundlagen der Annotation und verschiedene Anwendungsgebiete vorgestellt. Der Unterschied zwischen manueller und semi-automatischer Annotation wird erklärt und es werden zwei Tools vorgestellt, die bei der Annotation verwendet werden können.

2.1 Natural Language Processing

Natural Language Processing (kurz NLP) beschäftigt sich mit Techniken zur maschinellen Verarbeitung natürlicher Sprachen. Das Ziel von NLP ist eine direkte Kommunikation von Mensch und Maschine zu ermöglichen. Mithilfe von Algorithmen und bestimmten Regeln wird versucht, natürliche Sprache computerbasiert zu bearbeiten. Natural Language Processing muss gesprochene bzw. geschriebene Sprache erkennen, analysieren und für weitere Verwendung aufbereiten. Damit NLP Sprache verstehen und verwenden kann, genügt es nicht einzelne Wörter zu erkennen und zu verstehen, es müssen ganze Textzusammenhänge und Sachverhalte erkannt werden. Eines der größten Herausforderungen bei Natural Language Processing ist die Mehrdeutigkeit und Komplexität der menschlichen Sprache. Außerdem beschränkt sich der Computer auf Algorithmen und Verfahren von maschinellem Lernen und kann nicht wie der Mensch auf Erfahrungen im Hinblick auf Sprache zurückgreifen. NLP muss den Sinn einer Zeichenkette erkennen und extrahieren. Dafür verwendet es verschieden Techniken und Methoden. Die einzelnen typischen Schritte dazu sind:

- Spracherkennung
- Segmentierung von Sprachen in einzelne Wörter
- Grundformen und grammatische Information erkennen
- Funktionen der Wörter erkennen (Verb, Adjektiv, ...)

- Bedeutung von Sätzen extrahieren
- Satzbeziehungen und Zusammenhänge extrahieren

Nicht immer kommt es bei NLP zu einem befriedigendem Ergebnis, da selbst modernste künstliche Intelligenz wegen diverser Stilmittel (zum Beispiel rhetorische Fragen, Metaphern, etc.) an ihre Grenzen stößt. Mittels Annotation kann Sprache so aufbereitet werden, dass sie für künstliche Intelligenz und maschinelles Lernen verwendet werden kann [23].

2.2 Annotation im Allgemeinen und verschiedene Techniken

Annotation ist ein wichtiger Grundstein für Natural Language Processing. Bei der Annotation werden zusätzliche Information zum Text hinzugefügt. Diese Informationen können in Form von Kommentaren, Markierungen oder Beziehungen zwischen Textteilen dargestellt werden.

Für das Internet spielt Annotation ebenso eine große Rolle. Im World Wide Web befinden sich zahlreiche Daten in Form von Texten, Videos, Bildern etc. Nutzer können mithilfe von Sprache Informationen im Internet verstehen und sie mit anderen Inhalten in Verbindung bringen. Der Computer kann zwar die Information exzellent für den Nutzer bereit stellen, die Sprache verstehen kann er aber sehr schlecht bis gar nicht. *Human Language Technologies (HlT)* [1] verwendet berechenbare Eigenschaften von sprachlichen Strukturen, um Programme zu entwerfen. Diese Programme erlauben dem Computer Sprache zu verstehen und mittels Sprache mit Nutzern zu interagieren. Da sehr viele Menschen das Internet verwenden, gibt es eine große Anzahl an sprachlichen Daten. Daher können *HLT-Probleme* als Herausforderung und Aufgabe für maschinelles Lernen gesehen werden. Es reicht allerdings nicht, den Computer mit einer großen Anzahl an Daten zu füttern und von ihm zu erwarten Sprache und Sprechen zu lernen. Die zur Verfügung gestellten Daten müssen zuerst bearbeitet werden, damit die Maschine Muster und Beziehungen erkennen kann. Das passiert mit Metadaten. Jegliches Hinzufügen von Metadaten in Form von Tags und Markierungen nennt man Annotation. Damit

[1] Human Language Technologies beschreibt ein breites Forschungsgebiet mit dem Ziel, den Austausch von Mensch und Maschine mittels natürlicher Kommunikation zu ermöglichen [4].

die Maschine effizient lernen kann, müssen diese Annotationen genau und akkurat sein. Automatische und semi-automatische Annotation gewährleistet dies eher als manuelle Ansätze [20].

2.2.1 Part of Speech (POS) Tagging Annotation

Part of speech ist die einfachste Art von Text-Annotation. Beim POS-Tagging wird jeder lexikalischen Einheit im Text ihr Part of speech in Form eines POS-Tag zugeordnet. Als Part of speech bezeichnet man die Beschreibung eines Terms mittels seiner Grammatik (zum Beispiel: Nomen, Adjektiv, Past participle, etc.) [15]. Pos-Tagging wird beispielsweise für die Suche nach grammatikalischen oder lexikalischen Mustern verwendet, ohne dafür ein konkretes Wort zu definieren (zum Beispiel: Finde alle Nomen im Plural ohne vorangehendem Artikel). Die Sammlung von POS-Tag nennt man Tagset. Die Abbildung 2.1 zeigt einen Ausschnitt aus dem *Stuttgart-Tübingen-TagSet (STTS)*. Das Stuttgart Tübigen Tagset wurde von der Universität Stuttgart und der Universität Tübigen entworfen. Durch den Entwurf eines einheitlichen Tagsets können bereits annotierte Texte ohne komplizierte Anpassung der Tags gegenseitig verwendet werden. Das *STTS* besteht aus 54 Tags und umfasst 11 Hauptwortarten [2].

Beim POS-Tagging wird jedem Wort die richtige Klasse aus dem Tagset zugeordnet (vgl. Abbildung 2.2). Schwierigkeiten beim Tagging sind vor allem *Out of Vocabulary (OOV)* Fälle: unbekannte Wörter, die nicht im Lexikon stehen und *Ambiguitäten*: ein Wort kann mehrere Tags tragen (zum Beispiel: Sucht $\longrightarrow$ NN vs. VVFIN, ab $\longrightarrow$ ADP vs. PTKVZ, etc). Um diese Probleme zu lösen, müssen der Wortkontext und die Worteigenschaften mit einbezogen werden [21]. Die Pos-Tagging Variante ist eine der meist verwendeten Annotationen. Ein Vorteil ist, dass dieser Ansatz auch von Computern sehr genau ausgeführt werden kann. Das zeigen auch das Programm TAGGIT, das mit einer Genauigkeit von 71% Wörter taggt, oder das Programm CLAWS, das eine Genauigkeit von 95% erreicht [15].

2.2.2 Eigennamenerkennung

Als Named Entity Annotation oder auch Eigennamenerkennung bezeichnet man die Klassifikationen von Eigennamen in einem Text. Vor allem durch die Förderung

der MUC-Konferenzen (1998) [16] stellt Eigennamenerkennung ein zentrales Thema in vielen Arbeiten dar. Neben dem Datensatz von MUC gibt es auch neuere Ansätze für Named Entity Annotation. Die *Computational Natural Language Learning* (kurz CoNLL) Konferenz ist eine jährlich stattfindende Konferenz, die sich mit verschiedenen Forschungsfragen von Natural Language Processing beschäftigt. Die CoNLL-2002 und CoNLL-2003 thematisiert Named Entity Recognition [3]. CoNLL-2002 behandelt Named Entity Annotation in Spanisch und Niederländisch, bei der CoNLL-2003 liegt der Fokus auf der Sprachen unabhängigen Annotation von Eigennamen [22].

Mit dem Datensatz von MUC werden folgende Kategorien für Eigennamen definiert: Personen, Unternehmen, geographische Ausdrücke, Datums- und Maßangaben. Mit diesem Datensatz konnte eine 97 prozentige Vollständigkeit und eine 95 prozentige Korrektheit bei der Erkennung von Eigennamen für die englische Sprache erzielt werden. Im Deutschen ist Named Entity Annotation etwas komplizierter und schneidet daher schlechter ab. Ein Grund für die schlechtere Durchführung im Deutschen ist, dass Eigennamen nicht durch Groß- und Kleinschreibung von Nomen unterschieden werden können. Deshalb können einige Regeln im Deutschen nicht angewendet werden. Eine dieser Regeln lautet: Ein Vorname, gefolgt von einem groß geschriebenen Wort, ist ein Personenname. In den Sätzen "schreibt Lisa Bücher" oder "repariert Karl Autos", wird Bücher und Autos irrtümlich als Eigenname erkannt. Oft wird Pos-Tagging und Named Entity Annotation kombiniert, um mehr Kontext zu den Wörtern hinzuzufügen [6].

2.3 Manuelle vs. Semi-automatische Annotation

Bei der manuellen Annotation bearbeitet eine Person einen Textkorpus mit zuvor definierten Regeln und Anleitungen oder ohne irgendwelche Richtlinien. Selbst mit sehr gut definierten Regeln beeinflusst der Annotator/die Annotatorin das Ergebnis der Annotation stark. Vor allem die Struktur und der Aufbau des Satzes ist ausschlaggebend für die daraus resultierende Annotation. Der Annotator/die Annotatorin muss auf dem Gebiet der Sprachwissenschaft und für den Syntax von Sätzen über ein gutes Vorwissen verfügen, um eine schnelle und qualitativ hochwertige Annotation durchzuführen. Aufgrund der Mehrdeutigkeit von Wörtern muss immer

auch der Kontext betrachtet werden, um richtig zu annotieren. Der persönliche Hintergrund und die Erfahrung des Annotators/der Annotatorin sind ausschlaggebend für die Geschwindigkeit und Qualität der Annotation. Aber auch die Definition der Richtlinien spielt eine Rolle für eine schnelle und zufriedenstellende Annotation. Je komplizierter der Satzbau ist, desto zeitaufwendiger gestaltet sich die Annotation. Mithilfe von gut definierten Richtlinien können für komplizierte Sätze bestimmte Regeln festgelegt und so die benötigte Zeit minimiert werden. Manuelle Annotation ist sehr zeitintensiv und vor allem für große Datensätze in der Realität schlecht praktikabel. Aus diesem Grund greifen viele auf semi-automatische Annotation zurück [26]. Dabei korrigiert im Allgemeinen eine Person einen zuvor automatisch annotierten Text. Die Grundidee von semi-automatischer Annotation ist, einen Teil des Annotation-Prozesses zu automatisieren. Ein Tool wird mittels großer Datensätze trainiert, um ein besseres Ergebnis zu erzielen. Diese Datensätze können beispielsweise manuell oder mit Distant Supervision (siehe 2.3.1) erstellt werden. Anschließend wird der Text maschinell annotiert und eine Person überprüft den automatisch generierten Output. Fehlerhafte Ergebnisse werden gegebenenfalls verworfen [10]. Mittels semi-automatischer Annotation kann man bis zu 50% der Zeit einsparen [14].

2.3.1 Distant Supervision/Weak Supervision

Damit Annotation Tools automatisch laufen, müssen sie mit sehr großen Trainingsdatensätzen trainiert werden. Ein Datensatz kann zum Beispiel mit manueller Annotation erstellt werden. Da dies sehr viel Zeit erfordert und der Mensch fehleranfällig ist, eignet sich diese Methode für die Praxis eher schlecht. Ein alternativer Ansatz dazu wäre Distant Supervision bzw. Weak Supervision. Distant/Weak Supervision ist ein effizienter Ansatz, um Trainingsdaten zu sammeln. Bei Distant/Weak Supervision wird ein Datensatz mit einer bereits bestehenden Open Source Datenbank automatisch erstellt. Ein Beispiel für eine solche Datenbank ist Freebank [17]. Mithilfe von Distant Supervision und Freebank kann man einen Trainingsdatensatz erstellen, um beispielsweise Beziehungen zwischen Eigennamen zu kennzeichnen. Abbildung 2.3 zeigt ein Beispiel für einen solchen Datensatz. Die Eigennamen Apple und Steve Jobs sind mit der Beziehung *Firmengründer* gekennzeichnet, d.h. alle Sätze, die diese beiden Wörter enthalten, werden als Trainingsdaten ausgewählt. Distant/Weak Supervision ist zwar eine wirksame Methode zur Erstellung großer Datensätze, aber sie hat auch Nachteile. Es kommt vor, dass falsche Labels gewählt

werden, da auf den Inhalt der Eigennamen nicht eingegangen wird. Nur weil zwei Eigennamen in einem Satz vorkommen, heißt es nicht zwingend, dass dieser Satz die entsprechende Beziehung der Wörter ausdrückt. Nehmen wir das vorherige Beispiel 2.3: Der erste Satz wurde richtig ausgewählt und drückt die Beziehung (*company/founders*) zwischen den beiden Eigennamen *Steve Jobs* und *Apple* aus. Der zweite Satz beinhaltet zwar auch die beiden Eigennamen, hat aber nichts mit der Beziehung *Firmengründer* zu tun. Beide Sätze werden aber als Trainingsdaten ausgewählt. Durch solche falsch gewählten Daten wird die Performance der trainierten Modelle verschlechtert [27].

2.4 Vorstellung von Annotation Tools

Mittlerweile gibt es eine große Anzahl von Annotation Tools, die sich sich in ihrer Funktion, ihrem Aussehen und auch in zahlreichen anderen Aspekten unterscheiden. Einige dieser Anwendungen sind web-basiert und somit für alle Geräte geeignet, da keine spezielle Installation nötig ist. Andere müssen lokal installiert werden und deswegen sind sie vom Betriebssystem abhängig. Neves et al. [18] haben 78 Annotation Tools ausgewählt und nach 26 verschiedenen Kriterien untersucht. Die Kriterien sind in vier Kategorien eingeteilt: *Publication criteria*, *Technical criteria*, *Data criteria*, *Functional criteria*. Nur zwei Anwendungen decken 80% der Kriterien ab, 15 weitere Tools decken 60% der Kriterien ab. Die am häufigst erfüllten Kriterien sind die Unterstützung aller Standard-Datenformate, qualitativ hochwertige Dokumentationen und die Möglichkeit Textteile zu markieren. Source Code nicht verfügbar, schlechte bzw. fehlende Dokumentation und stark wechselnde Qualität im Ergebnis waren die am öftesten fehlenden Kriterien im Vergleich zu den anderen Tools. Im folgenden Abschnitt werden zwei Anwendungen näher betrachtet.

2.4.1 Brat

Nach Neves et al. [18] erfüllte Brat 16 ihrer 26 definierten Kriterien. Viele moderne Annotation Tools weisen eine nicht zufriedenstellende Benutzeroberfläche auf. Dieses Manko kann das Endergebnis einer Annotation stark beeinträchtigen. Eine schlechte Handhabung vergrößert logischerweise den Kosten- und Zeitfaktor

von anwendungsbasierter Annotation. Brat ist ein online Tool, das neben einer guten Dokumentation auch eine größtenteils selbsterklärende Benutzeroberfläche aufweist. Somit spart dieses Tool Zeit und Geld bei der Annotation. Brat unterstützt verschiedene Sprachen sowie eine große Anzahl von Annotations-Typen, wie man in Abbildung 2.4 sehen kann. Je nach Gebrauch und Ziel kann der Nutzer unter zahlreichen Einstellungen auswählen. Brat unterstützt die einfachsten Annotation Tasks wie zum Beispiel *POS-Tagging* oder *Named Entity Annotation*. Außerdem besteht die Möglichkeit, Beziehungen zu erstellen und somit Annotationen wie *Dependenzgrammatik-Erkennung* [2] und *Verb-Frame Erkennung* [3] durchzuführen. Das Tool kann online verwendet werden und ist mit den meisten Browsern kompatibel. Außerdem besteht die Möglichkeit Brat downzuloaden und lokal zu installieren. Es hat eine sehr intuitive und benutzerfreundliche Oberfläche. Die Anwendung wurde mit Standard Web-Technologien entwickelt und vermittelt somit eine vertraute Umgebung für die meisten Nutzer. Mittels Funktionen, bekannt aus Text-Editoren und anderen Anwendungen, können Dokumentteile annotiert werden. Mit einem doppelten Mausklick auf ein Wort werden neue Labels hinzufügen oder bereits bestehende editiert. Beziehungen können mittels *drag&drop* zwischen Textteilen aufgebaut werden. Abbildung 2.5 zeigt, wie eine Beziehung in Brat erstellt wird. Brat wurde in zahlreichen Projekten als Annotation Tool verwendet. Diese Projekte beinhalteten japanische Verb-frame Annotation und Annotation von biologischen Texten [24]. Automatische Annotation-Tools werden in Brat per Mausklick integriert, womit laut Stenetorp et al. [24] die Annotations Zeit um 15% verringert werden kann.

[2] Dependenzgrammatik wurde von Lucien Tesnière gegründet und basiert auf der Abhängigkeit von einem Wort auf ein anderes. Dabei steht das Verb hierarchisch über allen anderen Wortarten. Zum Beispiel der Satz "Lea lernt."besteht laut Lucien Tesnière [8] aus dem Verb *lernt*, dem Nomen *Lea* und der Abhängigkeit von Lea und lernt [8].

[3] Verb-Framed bezeichnet die Darstellung eines Bewegungspfades in einem Verb. Im Gegensatz dazu erfolgt bei *Satellite-Framed* die Beschreibung eines Pfades außerhalb eines Verbs mit sogenannten Satelliten (beispielsweise Präpositionen oder Adverbien). Spanisch ist zum Beispiel eine Verb-Framed Sprachen. Im spanischen werden viele Bewegungsverben eingesetzt, zum Beispiel: *entrar* (sich hineinbewegen) oder *salir* (sich hinaus bewegen). Im Deutschen werden Bewegungsweisen meist außerhalb vom Verb mit Präpositionen (aus der Wohnung) oder Adverbien (heraus, vorbei) dargestellt [7].

2.4.2 BioQRator

Ein anderes, ebenfalls web-basiertes Annotation Tool ist BioQRator. Der größte Unterschied von BioQRator zu Brat und anderen Tools ist, dass es sich auf die Annotation von medizinischer Literatur beschränkt. Es wurde für die Annotation von Eigennamen und Beziehungen entwickelt und unterstützt als erstes Web-Tool das BioC Format [13] [4]. BioQRator ist einfach zu verwenden und hat eine sehr benutzerfreundliche und intuitive Oberfläche. Ähnlich wie bei Brat können per Mausklick Labels und Beziehungen hinzugefügt bzw. entfernt werden. Außerdem ist BioQRator mit den meisten modernen Browsern kompatibel, da es mit HTML5/CSS implementiert wurde. Dokumente können mit einem BioC File hinzugefügt oder in PubMed[5] gesucht und hinzugefügt werden. Für Dokumente, die mittels PubMed und BioC importiert wurden, besteht die Möglichkeit, zahlreiche vordefinierte medizinisch genormte Eigennamen zur Annotation zu verwenden. In Abbildung 2.6 ist eine Annotation mit BioQRator und den vordefinierte Eigennamen für *Protein-Protein-Interaktionen*[6] zu sehen [13].

[4] Als BioC Format bezeichnet man ein einfaches Format für den Austausch von Textdokumenten und Annotationen. BioC Files sind in XML kodiert und bestehen aus Dokumenten, gegliedert in Absätzen. Jeder Absatz und einzelne Sätze können zusätzlich Annotationen beinhalten [5].

[5] PubMed ist eine Datenbank mit über 30 Millionen Referenzen auf biomedizinischer Literatur [9]

[6] *Protein-Protein-Interaktionen (PPI)* sind für viele Prozesse in der Zelle sehr wichtig. Durch die Veränderung von Proteinen werden biologische Funktionen beeinflusst [19].

POS =	Beschreibung	Beispiele
ADJA	attributives Adjektiv	*[das] große [Haus]*
ADJD	adverbiales oder prädikatives Adjektiv	*[er fährt] schnell* *[er ist] schnell*
ADV	Adverb	*schon, bald, doch*
APPR	Präposition; Zirkumposition links	*in [der Stadt], ohne [mich]*
APPRART	Präposition mit Artikel	*im [Haus], zur [Sache]*
APPO	Postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	Zirkumposition rechts	*[von jetzt] an*
ART	bestimmter oder unbestimmter Artikel	*der, die, das,* *ein, eine*
CARD	Kardinalzahl	*zwei [Männer], [im Jahre] 1994*
FM	Fremdsprachliches Material	*[Er hat das mit "]* *A big fish [" übersetzt]*
ITJ	Interjektion	*mhm, ach, tja*
KOUI	unterordnende Konjunktion mit "zu" und Infinitiv	*um [zu leben],* *anstatt [zu fragen]*
KOUS	unterordnende Konjunktion mit Satz	*weil, daß, damit,* *wenn, ob*
KON	nebenordnende Konjunktion	*und, oder, aber*
KOKOM	Vergleichspartikel, ohne Satz	*als, wie*
NN	Appellativa	*Tisch, Herr, [das] Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativ–pronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativ–pronomen	*jener [Mensch]*
PIS	substituierendes Indefinit–pronomen	*keiner, viele, man, niemand*
PIAT	attribuierendes Indefinit–pronomen ohne Determiner	*kein [Mensch],* *irgendein [Glas]*
PIDAT	attribuierendes Indefinit–pronomen mit Determiner	*[ein] wenig [Wasser],* *[die] beiden [Brüder]*
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessiv–pronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein [Buch], deine [Mutter]*
PRELS	substituierendes Relativpronomen	*[der Hund,] der*

Abbildung 2.1: Ausschnitt aus dem Stuttgart-Tübingen-TagSet [2]

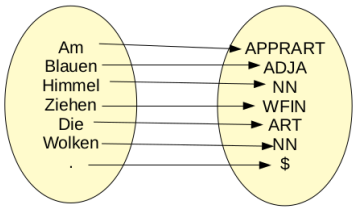

Abbildung 2.2: Beispiel zum Pos-Tagging nach STTS. Den Wörtern in der linken Ellipse werden die STTS-Tags in der rechten Ellipse zugeordnet [21].

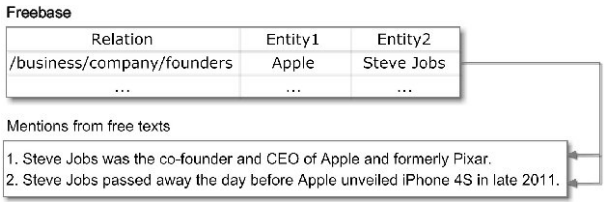

Abbildung 2.3: Datensatz generiert mittels Distant Supervision. Der obere Satz wurde richtig gekennzeichnet der untere nicht [27]

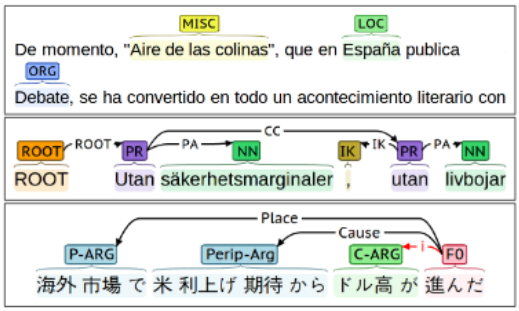

Abbildung 2.4: Ausschnitt aus Annotation mittels Brat. Ganz oben: *Eigennamenserkennung*, in der Mitte: *Dependenzgrammatik-Erkennung* und unten: *Verb-Frame Erkennung* [24]

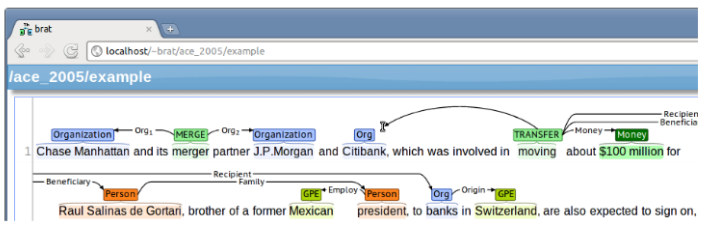

Abbildung 2.5: Erstellung von Beziehungen im Annotation-Tool Brat. Zwischen den einzelnen Eigenname (farbige Klassen) werden mittels Pfeile Beziehungen erstellt [24]

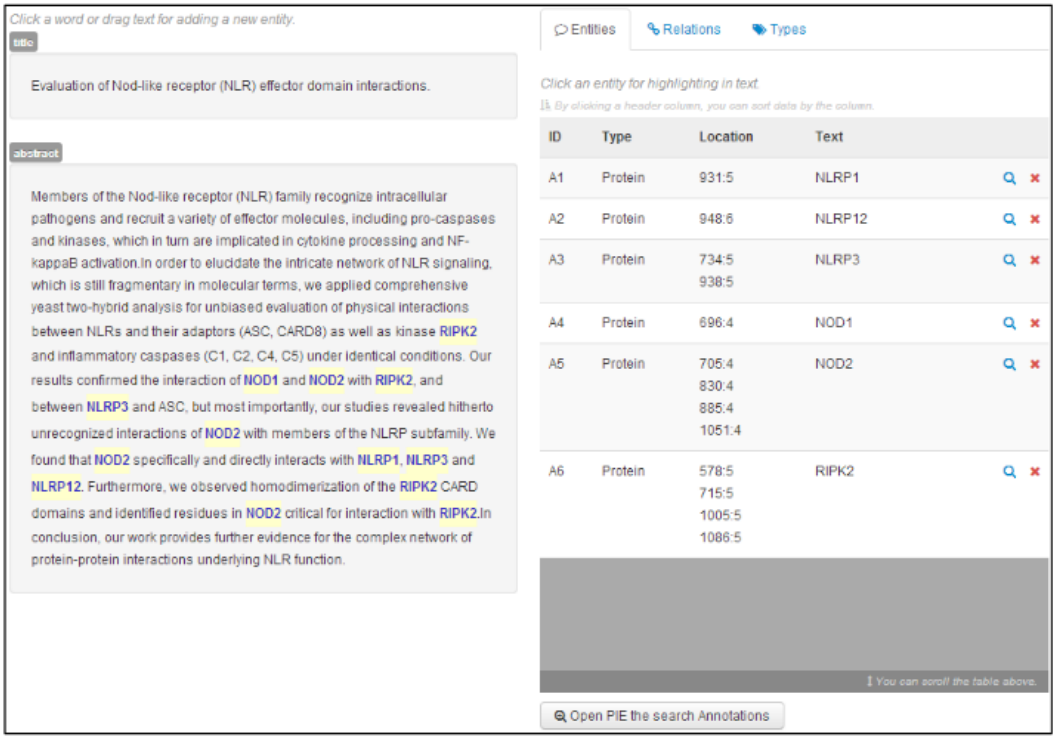

Abbildung 2.6: Annotation von Eigennamen mit dem Annotation-Tool BioQRator das speziell für medizinische Literatur entwickelt wurde. Rechts werden die vordefinierten *Protein-Protein-Interaktionen* Klassen dargestellt [13]

3 Evaluation

Ziel der Evaluation ist herauszufinden welche Methode des Webtools *CodeAnnotator* besser zur Annotation geeignet ist. Untersucht wird die verwendete Zeit und das Ergebnis im Allgemeinen. Außerdem wird die Selbsteinschätzung und die tatsächliche Genauigkeit verglichen.

3.1 Aufbau des Webtools

Der *CodeAnnotator* wurde unter der Leitung von Herrn Ass.Prof.Dipl.-Ing.Dr.techn. Roman Kern am Know-Center der TU Graz entwickelt. Man hat die Möglichkeit sich zu registrieren und eine beliebige Anzahl von Datensätzen zu erstellen. Es gibt verschiedene Möglichkeiten Datensätze zu bilden. Datensätze können mit wissenschaftlichen Dokumenten, .txt Dateien oder der eigenen Mendeley Bibliothek erstellt werden [1]. Jedem hinzugefügten Datensatz kann man ein oder mehrere Modelle zuweisen. Diese Modelle definieren die spätere Annotation und die darin verwendeten Entities. Das Webtool bietet zwei verschieden Methoden für die Annotation an. Methode 1: manuelle Annotation mit einem eingebetteten Brat Editor, Methode 2: Annotation mittels *Keyword in Context* (kurz KWIC) [2]. Bei der manuellen Annotation werden Wörter oder Wortgruppen markiert und die entsprechenden Klassen in einem Pop Up auswählt. Die Annotation mit KWIC funktioniert über ein Suchfeld. Das gesuchte Wort wird in einer Liste mit seinem Kontext dargestellt und die gewünschte Annotation-Klasse wird ausgewählt. Per Mausklick können einzelne Einträge zur Annotation hinzugefügt oder verworfen werden. Abbildung

[1] Mendeley ist ein online Literaturverwaltungssystem für Studierende und Forschende. Die Spezialisierung liegt dabei auf die Verwaltung von pdf-Dokumenten. Man kann Fachliteratur organisieren, austauschen, zitieren, kommentieren und Textteile farbig markieren [25].

[2] Darstellung eines ausgewählten Wortes als Liste in seinen Umgebungswörtern (Kontext) [1].

3.1 zeigt ein Beispiel für die KWIC Annotation mit *Christine* als Suchbegriff und der Klasse *given-name* [12].

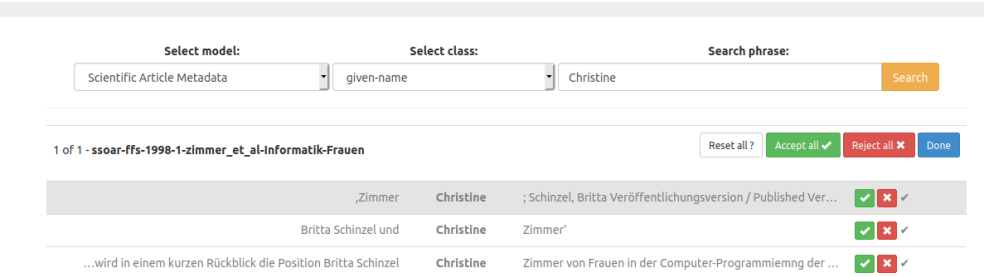

Abbildung 3.1: Beispiel für KWIC Annotation mit CodeAnnotator. Darstellung vom Suchbegriff *Christine* in seinem Kontext [12].

3.2 Vorbereitung der Evaluation und Ablauf

An der Evaluation nahmen 7 Testpersonen teil. Die Personen kommen aus dem näheren Umfeld der Autorin. Der Fokus lag auf einer möglichst gleichen Erfahrung mit Annotation der Testpersonen um das Ergebnis der Evaluation nicht mit dem Vorwissen Einzelner zu beeinflussen. Für jeden Probanden wurde ein eigener Datensatz mit wissenschaftlichen Dokumenten angelegt. Annotiert wurden zwei verschiedene Dokumente aus dem Themengebiet der Informatik: **Dokument 1: *Informatik-Frauen*** **[28]**, **Dokument 2: *Informatik - Kompetenzentwicklung bei Kindern*** **[11]** Als Annotation-Modell wurde das vorinstallierte Set für wissenschaftliche Dokumente gewählt. Das Modell besteht aus den 30 nachfolgend angeführten Klassen:

- title
- journal
- subtitle
- academic-title
- given-name
- middle-name
- surname
- index
- affiliation
- email
- doi
- ref-authorGivenName
- ref-authorSurname
- ref-authorOther
- ref-editor
- ref-title
- ref-date
- ref-publisher
- ref-issueTitle
- ref-bookTitle
- ref-pages
- ref-location
- ref-conference
- ref-source
- ref-volume
- ref-edition
- ref-issue
- ref-url
- ref-note
- ref-other

Dokument 1 enthält 122 Wörter die annotiert werden sollen in Dokument 2, 215 Wörter. Tabelle 3.3 zeigt die Verteilung der einzelner Klassen. Außerdem kann man in Abbildung 3.2 bis 3.5 eine Visualisierung des Ergebnis sehen. Die Reihenfolge der Dokumente und Annotation-Methoden wurde zufällig gewählt (siehe 3.1). Das Webtool unterstützt zwei verschiedene Verfahrensweisen: **Methode 1: manuelle Methode mithilfe von Brat**, **Methode 2: Annotation mittels KWIC**. Jeder Proband/jede Probandin hatte eine unterschiedliche Erfahrung mit der Bedienung von Computern bzw. Webtools und der Annotation von wissenschaftlichen Dokumenten und Texten im Allgemeinen. Bei Methode 1 hatten die Testpersonen keinen Einblick auf das pdf-Dokument, bei der zweite Methode durften sie das pdf-Dokument verwenden. Nach der Annotation wurde jeder aufgefordert, einen Fragebogen auszufüllen. Das Template des Fragebogens befindet sich im Appendix 5.1.

Eine wesentliche Einschränkung der Evaluation war eine eher geringe Anzahl an Teilnehmern. 7 Personen spiegeln ein Ergebnis recht gut wieder, aber je höher die Anzahl der Probanden, desto aussagekräftiger ist ein Ergebnis. Außerdem hatten die Personen keine bzw. wenig Erfahrung mit Annotation, deshalb ist das Ergebnis der Annotation nicht nur auf die angewandten Methoden zurückzuführen, sondern auch stark beeinflusst von der Erfahrung der Testpersonen.

	erster Durchgang	zweiter Durchgang
Testperson 1	Methode 2 Dokument 1	Methode 1 Dokument 2
Testperson 2	Methode 1 Dokument 1	Methode 2 Dokument 1
Testperson 3	Methode 1 Dokument 1	Methode 2 Dokument 1
Testperson 4	Methode 1 Dokument 2	Methode 2 Dokument 2
Testperson 5	Methode 2 Dokument 1	Methode 1 Dokument 2
Testperson 6	Methode 2 Dokument 1	Methode 1 Dokument 2
Testperson 7	Methode 1 Dokument 2	Methode 2 Dokument 2

Tabelle 3.1: Reihenfolge der Annotation Durchläufe bei der Evaluation

3.3 Ergebnisse und Diskussion

Im Allgemeinen lieferten alle Probanden bei der Annotation ein ansprechendes Ergebnis. Alter, Ausbildung und unterschiedliche Erfahrungen mit der Verwendung von Computern beeinflussten das Ergebnis nicht nennenswert. Nur eine Person hatte schon Erfahrungen mit Annotation. Dieser Probant annotierte bei beiden Methoden mehr als 80% der Wörter richtig.

Die durchschnittliche Zeit für die manuelle Methode beträgt mit Brat 1:20:51, für Methode 2 benötigte man im Durchschnitt 49 Minuten und 11 Sekunden (eine Visualisierung der einzelnen Zeiten inklusive Durchschnitt siehe 3.6 und 3.7). Bei Dokument 1 wurde mit der Methode 1 170 Wörter von 215 richtig annotiert. 140 Wörter von 215 wurden mit der zweiten Methode richtig annotiert. Bei Dokument 2 wurde mit der manuellen Methode 91 Wörter von 122 richtig annotiert, mit der KWIC Methode 49 Wörter von 122. 21 Wörter wurden im Durchschnitt bei der Methode 1 falsch annotiert und 18 bei der Methode 2. Hier kann man gut erkennen, dass die Methode mit Brat zwar etwas längere Zeit in Anspruch nimmt (31:40

Minuten mehr), dafür aber genauer ist als die KWIC Methode. Es gibt keinen signifikanten Unterschied, ob ein Dokument zuerst mithilfe von Brat oder der KWIC annotiert wurde. Vor allem die Formatierung der Dokumente führte bei der manuellen Annotation zu einigen Schwierigkeiten. Überschriften wurden häufig nicht als solche erkannt oder Textteile, die nicht zur Klasse *title* gehören als solche markiert. Zu beobachten war auch, dass aufgrund der Formatierung Überschriften häufig als *subtitle* markiert wurden. Klassen, die sehr eindeutig sind, wie zum Beispiel *given-name*, *surname*, *authorGivenName*, *authorSurname* oder *email* wurden in den meisten Fälle richtig erkannt. Schwierigkeiten machten Klassen wie zum Beispiel *index* (Methode 1: durchschnittlich 18 von 27 richtig, Methode 2: durchschnittlich 0,3 richtig).Eine genaue Auflistung der Anzahl von richtig annotierten Wörtern einzelner Klassen siehe 3.3 und 3.1.

Nach der Annotation musste jede Testperson einen Fragebogen ausfüllen. Teil dieses Fragebogen war auch eine Skala zur Selbsteinschätzung der einzelnen Methoden von 1 - 6 (1= sehr gut, 2 = gut, 3 = eher gut, 4 = eher schlecht, 5 = schlecht, 6 = sehr schlecht).Die Selbsteinschätzung passt bei der manuellen Annotation in den meisten Fällen nicht zum tatsächlichen Ergebnis. Hier bewerteten sich die meisten Probanden zwischen *sehr schlecht* und *eher schlecht*, nur zwei Personen stuften sich mit *eher gut*. Das tatsächliche Ergebnis der Annotation war aber wesentlich besser. Für die Methode 2 war die Selbsteinschätzung besser das tatsächliche Ergebnis, war aber im Vergleich zur manuellen Annotation nicht so treffend. Fünf der Personen bewerteten sich mit *eher gut* oder *gut*, die anderen *eher schlecht* oder *schlecht*. Die schlechte Selbsteinschätzung im Kontrast zum guten Ergebnis lässt sich wahrscheinlich mit der Unerfahrenheit der Probanden im Bezug auf Annotation begründen. Der Großteil der Personen hatte nämlich zuvor noch nie annotiert und daher auch keine Einschätzung, ob etwas richtig oder falsch sein könnte. Außerdem erhält man kein direktes Feedback von der Webseite. Vier der Personen würden das Web Tool nicht privat nutzen oder weiterempfehlen, zwei Probanden gaben als Grund *Kein Bedarf* an, die restlichen zwei meinten, dass das Webtool zu kompliziert sei und dass zu viele Klicks benötigt würden. Zwei der Testpersonen führten zusätzlich an, dass ihnen das Markieren der Wörter/Wortgruppen bei der Methode mit Brat schwer gefallen sei. Außerdem gab eine Person an, dass Wörter mit deutschen Umlauten bei der KWIC Methode nicht gefunden werden konnten. Sechs der Getesteten würden das nächste Mal die Methode 2 verwenden und nur eine Person würde diese wegen ihrer Meinung nach *unnötigem Mehraufwand* nicht wieder gebrauchen. Interessant ist, dass diese Person um mehr als 20% besser war

als alle anderen. Zusammenfassend kann man feststellen, dass die Anwendung der ersten Methode etwas länger dauert, dafür das Ergebnis aber besser ausfällt als bei der zweiten Methode, die dafür um einiges weniger Zeit in Anspruch nimmt. Die KWIC Methode war bei den Testpersonen beliebter als das Annotieren mit Brat. Das kann man auf die um einiges unkompliziertere Bedienung zurückführen. Methode 1 war weiters deshalb nicht so beliebt, weil das Markieren etwas schwierig war und die Formatierung oft für Verwirrung sorgte. Vor allem für kurze Dokumente und unerfahrene Annotatoren/Annotatorinnen ist die Methode 2 besser geeignet als die erste. Wenn man ein sehr exaktes Ergebnis will und mehr Zeit zur Verfügung hat, sollte man eher die Methode 1 wählen. Vor allem für die Klassen *given-name* oder *surname* eignet sich die Methode 2 aber besser, weil auf diese Weise alle Vor- und Nachnamen schnell annotiert werden können. Für biografische Texte ist daher die Methode mittels KWIC sicher passender.

Tabelle 3.2: Annotationsergebnis Methode 1

	Referenz Dokument 2	Referenz Dokument 1	Testperson 1 (Dokument 2)	Testperson 4 (Dokument 2)	Testperson 5 (Dokument 2)	Testperson 6 (Dokument 2)	Testperson 7 (Dokument 2)	Testperson 2 (Dokument 1)	Testperson 3 (Dokument 1)
title	11	6	11	9	2	11	1	6	6
journal	0	0	0	0	0	0	0	0	0
subtitle	0	0	0	0	0	0	0	0	0
academic-title	0	2	0	0	0	0	0	2	0
given-name	5	46	5	5	5	5	5	46	46
middle-name	0	6	0	0	0	0	0	6	3
surname	7	61	6	7	7	6	6	61	60
index	27	0	27	22	27	27	27	0	0
affiliation	0	0	0	0	0	0	0	0	0
email	2	0	0	2	2	2	0	0	0
doi	1	0	1	1	1	1	1	0	0
ref-authorGivenName	4	10	4	3	4	4	4	9	8
ref-authorSurname	4	17	4	4	4	4	4	17	9
ref-authorOther	0	2	0	0	0	0	0	0	0
ref-editor	0	0	0	0	0	0	0	0	0
ref-title	11	12	8	9	10	3	5	9	7
ref-date	13	15	13	12	12	10	12	15	4
ref-publisher	0	1	0	0	0	0	0	0	0
ref-issueTitle	0	0	0	0	0	0	0	0	0
ref-bookTitle	0	0	0	0	0	0	0	0	0
ref-pages	8	7	7	8	7	6	8	8	1
ref-location	2	4	1	0	0	0	0	4	4
ref-conference	0	0	0	0	0	0	0	0	0
ref-source	18	15	17	10	8	13	3	0	0
ref-volume	7	4	1	1	0	0	0	1	0
ref-edition	0	0	0	0	0	0	0	0	0
ref-issue	0	0	0	0	0	0	0	0	0
ref-url	2	7	0	0	0	0	0	7	1
ref-note	0	0	0	0	0	0	0	0	0
ref-other	0	0	0	0	0	0	0	0	0
richtig annotiert in Prozent			86%	77%	73%	75%	62%	89%	70%
Selbsteinschätzung			schlecht	Sehr schlecht	Eher schlecht	Eher gut	Eher schlecht	schlecht	Eher gut
Gesamt	122	215	105	93	89	92	76	191	149

Tabelle 3.3: Anzahl der richtig annotierten Wörter jeder Testperson inkl. Prozentsatz und Selbsteinschätzung bei Methode 1

Tabelle 3.4: Annotationsergebnis Methode 2

	Referenz Dokument 2	Referenz Dokument 1	Testperson 1 (Dokument 1)	Testperson 2 (Dokument 1)	Testperson 3 (Dokument 1)	Testperson 4 (Dokument 2)	Testperson 5 (Dokument 1)	Testperson 6 (Dokument 1)	Testperson 7 (Dokument 2)
title	11	6	4	2	6	4	4	4	4
journal	0	0	0	0	0	0	0	0	0
subtitle	0	0	0	0	0	0	0	0	0
academic-title	0	2	0	0	0	0	0	0	0
given-name	5	46	44	40	40	5	19	40	5
middle-name	0	6	5	3	0	0	0	2	0
surname	7	61	49	70	60	4	16	40	4
index	27	0	0	0	0	1	0	0	1
affiliation	0	0	0	0	0	0	0	0	0
email	2	0	0	0	0	0	0	0	0
doi	1	0	0	0	0	0	0	0	0
ref-authorGivenName	4	10	10	6	6	0	10	9	4
ref-authorSurname	4	17	17	0	7	2	17	9	4
ref-authorOther	0	2	0	0	0	0	0	0	0
ref-editor	0	0	0	0	0	0	0	0	0
ref-title	11	12	12	0	5	10	8	4	8
ref-date	13	15	15	15	1	0	15	8	0
ref-publisher	0	1	1	1	0	0	1	1	0
ref-issueTitle	0	0	0	0	0	0	0	0	0
ref-bookTitle	0	0	0	0	0	0	0	0	0
ref-pages	8	7	7	5	7	4	5	3	5
ref-location	2	4	2	0	0	2	0	2	1
ref-conference	0	0	0	0	0	0	0	0	0
ref-source	18	15	0	0	0	15	2	3	10
ref-volume	7	4	4	0	2	0	0	4	0
ref-edition	0	0	0	0	0	0	0	0	0
ref-issue	0	0	0	0	0	0	0	0	0
ref-url	2	7	7	7	0	2	6	6	2
ref-note	0	0	0	0	0	0	0	0	0
ref-other	0	0	0	0	0	0	0	0	0
richtig annotiert in Prozent			87%	69%	62%	42%	48%	63%	39%
Selbsteinschätzung			schlecht	Eher schlecht	Eher gut	Eher gut	Eher gut	Eher gut	gut
Gesamt	122	215	177	149	134	49	103	135	48

Tabelle 3.5: Anzahl der richtig annotierten Wörter jeder Testperson inkl. Prozentsatz und Selbsteinschätzung bei Methode 2

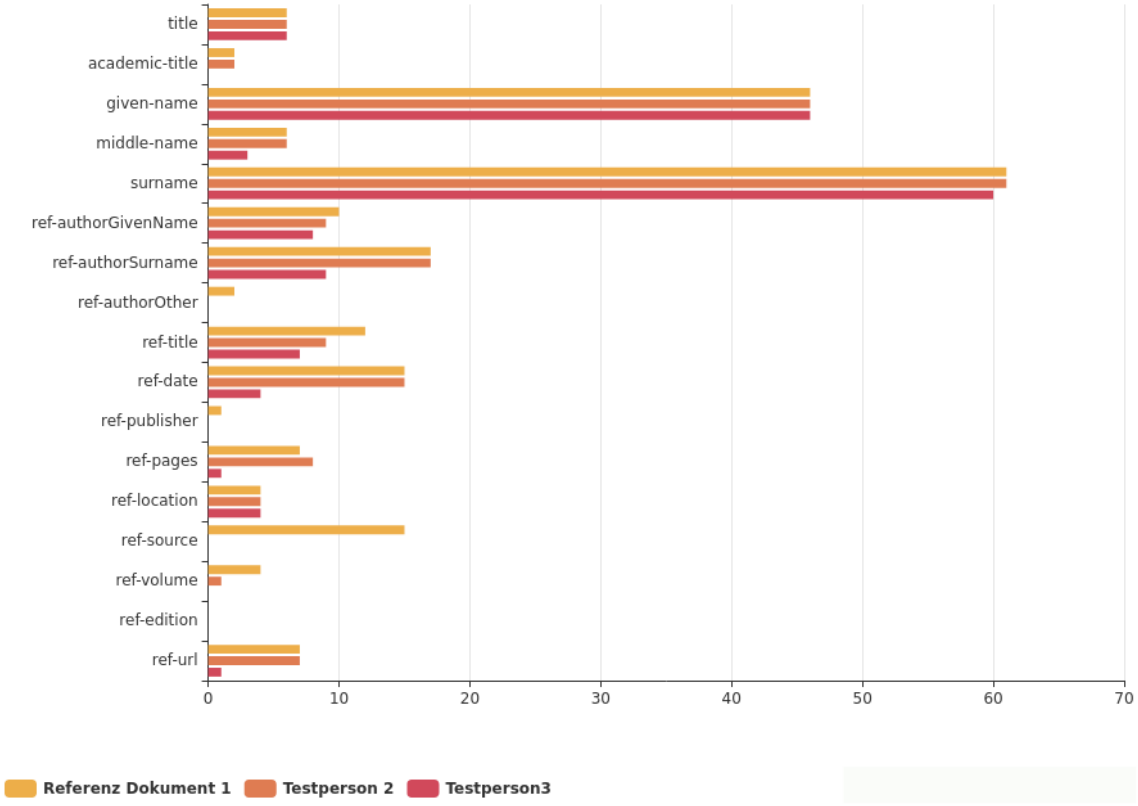

Abbildung 3.2: Visuelle Darstellung der Anzahl von richtig annotierter Wörter der Einzelnen Klassen bei Methode 1 Dokument 1. Die einzelnen Testpersonen und die Referenz-Annotation sind farblich dargestellt. Die x-Achse enthält die Anzahl der richtig annotierten Wörter, auf der y-Achse werden die einzelnen Klassen dargestellt.

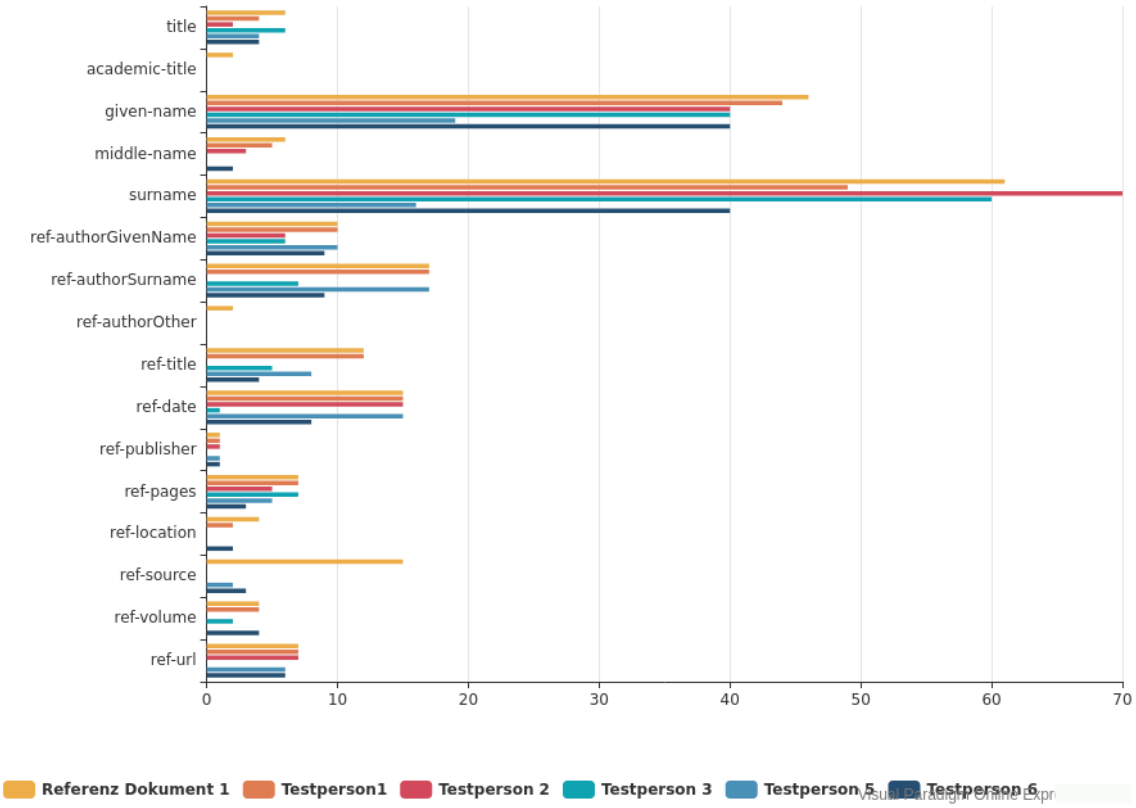

Abbildung 3.3: Visuelle Darstellung der Anzahl von richtig annotierter Wörter der Einzelnen Klassen bei Methode 2 Dokument 1. Die einzelnen Testpersonen und die Referenz-Annotation sind farblich dargestellt. Die x-Achse enthält die Anzahl der richtig annotierten Wörter, auf der y-Achse werden die einzelnen Klassen dargestellt.

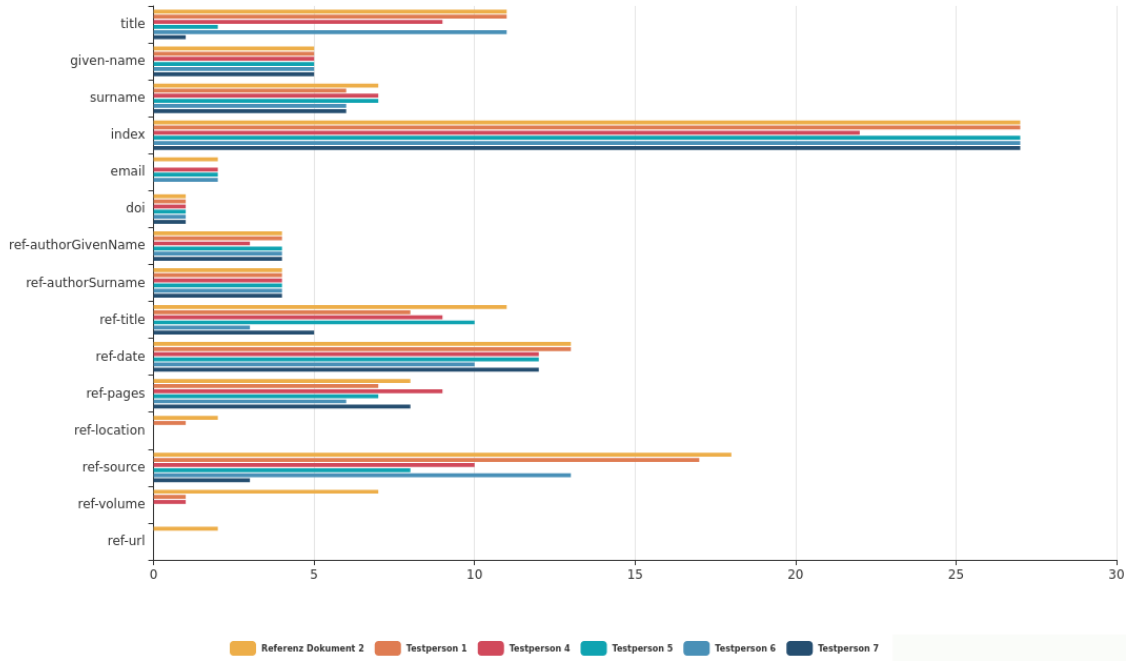

Abbildung 3.4: Visuelle Darstellung der Anzahl von richtig annotierter Wörter der Einzelnen Klassen bei Methode 1 Dokument 1. Die einzelnen Testpersonen und die Referenz-Annotation sind farblich dargestellt. Die x-Achse enthält die Anzahl der richtig annotierten Wörter, auf der y-Achse werden die einzelnen Klassen dargestellt.

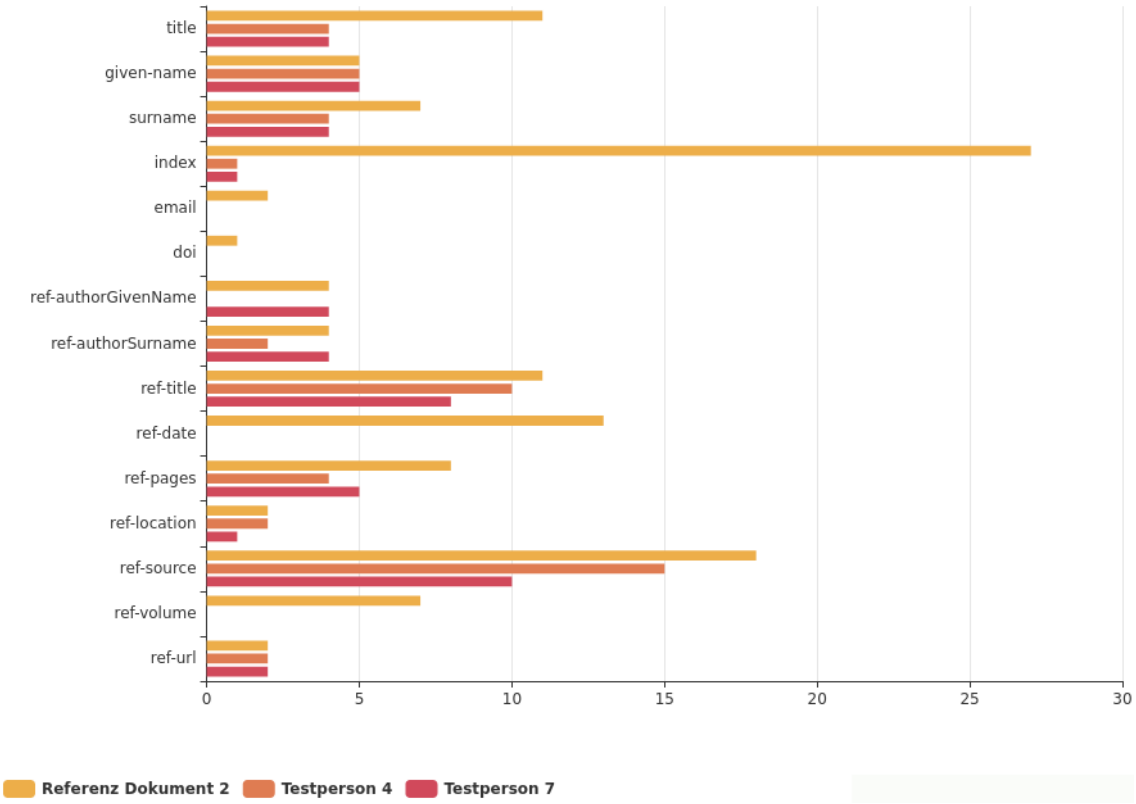

Abbildung 3.5: Visuelle Darstellung der Anzahl von richtig annotierter Wörter der Einzelnen Klassen bei Methode 2 Dokument 1. Die einzelnen Testpersonen und die Referenz-Annotation sind farblich dargestellt. Die x-Achse enthält die Anzahl der richtig annotierten Wörter, auf der y-Achse werden die einzelnen Klassen dargestellt.

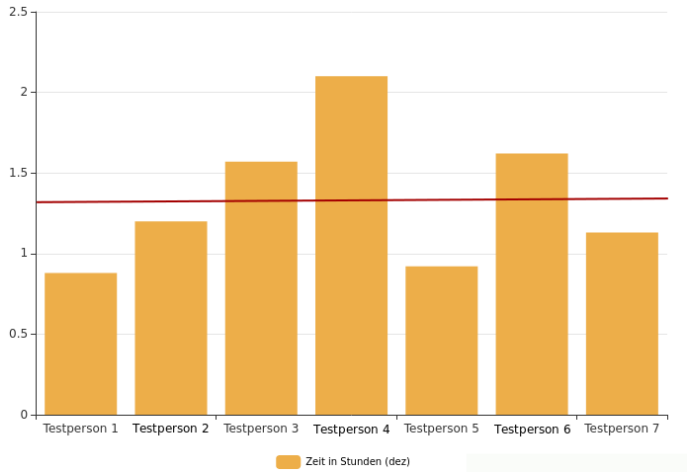

Abbildung 3.6: Visuelle Darstellung der benötigten Zeit der einzelnen Testpersonen bei Methode 1. Die x-Achse stellt die einzelnen Personen dar auf der y-Achse wird die benötigte Zeit in Stunden (dezimal) dargestellt. Die rote Linie visualisiert die Durchschnittliche Zeit von 1,35 Stunden.

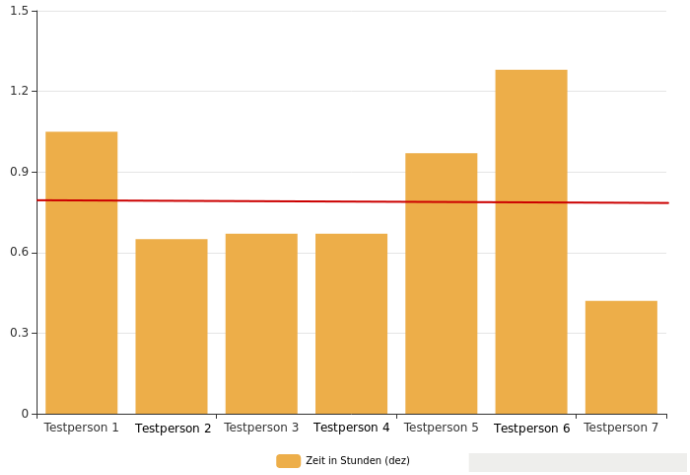

Abbildung 3.7: Visuelle Darstellung der benötigten Zeit der einzelnen Testpersonen bei Methode 2. Die x-Achse stellt die einzelnen Personen dar auf der y-Achse wird die benötigte Zeit in Stunden (dezimal) dargestellt. Die rote Linie visualisiert die Durchschnittliche Zeit von 0,82 Stunden.

3 Evaluation

4 Fazit

Annotation ist ein wichtiger Schritt um Sprache für *Natural Language Processing* (NLP) aufzubereiten. *NLP* beschäftigt sich mit der Verarbeitung natürlicher Sprache und will die direkte Kommunikation von Mensch und Maschine ermöglichen [23]. Auch für die Verarbeitung von Daten im Internet spielt die Annotation eine wichtige Rolle. Die sogenannte *Human Language Technologie* (HLT) entwickelt Programme, die in der Lage sind, dem Computer die Fähigkeit Sprechen und Sprache erkennen zu erlernen. Auch hier müssen den Daten Metadaten in Form von Annotationen hinzugefügt werden. Es gibt verschiedene Ansätze, um Textteile und Dokumente zu annotieren. Zwei sehr wichtige Ansätze sind die automatischen und semi-automatischen Methoden, die eine qualitativ hochwertige Annotation von großen Datensätzen gewährleisten [20]. Zwei sehr bekannte und häufig verwendete Annotation-Techniken sind *Part of Speech (POS) Tagging* und *Named Entity* Annotation. Beim *POS Tagging* wird jeder Einheit im Text seine grammatische Klasse zugeordnet (Nomen, Adjektiv, etc.) [15]. Bei *Named Entity* Annotation werden Eigennamen vorher definierten Klassen zugeordnet. Zum einen gibt es die Möglichkeit Textteile bzw. Dokumente manuell zu annotieren, d.h. eine Person hat den Text vor sich und annotiert diesen Anhand von Regeln und Modellen. Das Ergebnis der manuellen Annotation wird stark von der Erfahrung und Ausbildung des Annotators/der Annotatorin beeinflusst. Ein gutes Vorwissen in der Sprachwissenschaft und im Satzbau sind Voraussetzungen für einen qualitativ hochwertige Annotation. Manuelle Annotation ist sehr zeitaufwendig und vor allem für große Datensätze sollte man deshalb auf semi-automatische Annotation zurückgreifen [26]. Die Grundidee von semi-automatischer Annotation ist, dass ein Teil der Annotation automatisiert wird. Ein Annotation-Tool wird mit großen Datensätzen trainiert [10]. Diese Datensätze können beispielsweise mit Distant Supervision erstellt werden. Distant Supervision erstellt automatisch einen Datensatz mit einer Open Source Datenbank [17]. Nach dem Training annotiert das Tool dann die tatsächlichen Daten, eine Person muss nur mehr die automatisch erstellten Anmerkungen überprüfen und gegebenenfalls verwerfen [10]. Seit einiger Zeit gibt es eine

große Anzahl von Tools, die für die manuelle beziehungsweise semi-automatische Annotation hilfreich sind. Ein sehr bekanntes Tool ist Brat. Dieses Webtool ist stark konfigurierbar und unterstützt viele Sprachen. Deshalb kann es in vielen verschiedenen Fällen angewandt werden [24]. BioQRator ist ein eigens für medizinische Literatur konzipiertes Tool. Es unterstützt die Annotation von Eigenname und besitzt zahlreiche andere Features, die das Annotierten von medizinischen Texten vereinfacht [13].

Der praktische Teil dieser Arbeit beschäftigt sich mit dem Tool CodeAnnotator und der Frage, welche der zwei Methoden, die das Tool unterstützen, besser für die Annotation geeignet ist. Das Webtool wurde unter der Leitung von Herrn Ass.Prof.Dipl.-Ing.Dr.techn. Roman Kern am Know-Center der TU Graz entwickelt.Mit diesem Tool kann man zum einen Dokumente manuell mit einem eingebetteten Brat Editor annotieren oder mit Keyword in Context (kurz KWIC) suchen. Für die Evaluation wurden 7 Testpersonen ausgewählt, die in einer zufälligen Reihenfolge jeweils ein wissenschaftliches Dokument mit der manuellen Methode (Methode 1) und der KWIC Methode (Methode 2) annotierten. Das Ergebnis war in beiden Fällen durchaus zufriedenstellend. Für die manuelle Methode benötigte man im Durchschnitt 1:20:51 Stunden, für Methode 2 durchschnittlich 49 Minuten und 11 Sekunden. Mit der Methode 1 wurden bei Dokument 1 im Durchschnitt 170 von 215 Wörtern richtig annotiert. Im Dokument 2 wurden mit Methode 1 durchschnittlich 91 von 122 Wörtern richtig annotiert, im Dokument 1 waren mit Methode 2 140 von 215 Wörtern korrekt, bei Dokument 2 waren es 49 von 122. Das Ergebnis zeigt, dass die erste Methode zwar zeitaufwendiger ist, dafür aber das Resultat der Annotation besser ausfällt. Die KWIC Methode ist um einiges schneller, dafür aber nicht so genau. Interessant zu beobachten war, dass sich die meisten Teilnehmer schlechter einschätzten als ihre Ergebnisse tatsächlich waren. Der Großteil der Probanden würden ein nächstes Mal die zweite Methode eher wählen als die erste, weil sie schneller ausführbar und ihre Handhabung einfacher ist. Grundsätzlich ist die zweite Methode für Annotationen geeigneter, wenn wenig Zeit zur Verfügung steht und wenn nur ein einigermaßen gutes Ergebnis geliefert werden muss. Auch für kurze Dokumente und Personen mit wenig Vorwissen ist diese Methode eher geeignet. Das Annotieren mit Brat ist für Annotation zu bevorzugen, die mehr Zeit in Anspruch nehmen können, dafür aber ein gutes Ergebnis liefern müssen.

4.1 Ausblick

Die Evaluation gibt uns Hinweise darauf, wie Personen mit wenig Erfahrung mit dem CodeAnnotator annotieren. In Zukunft wäre es natürlich von Vorteil, eine Evaluation mit einer größeren Anzahl an Probanden durchzuführen. Außerdem wäre interessant, mehrere erfahrene Annotatoren/Annotatorinnen einzubinden, da geringes Vorwissen das tatsächliche Ergebnis der einzelnen Methoden beeinflusst bzw. auch beeinträchtigt. Auch die Anzahl, Länge und Art der Dokumente könnten geändert werden, um den Einfluss von Textlänge und Inhalt besser zu erfassen. Ein nächster Schritt wäre die Annotation mit anderen Modellen bzw. Annotation-Strategien durchzuführen um herauszufinden, für welche Dokumente bzw. Annotationstechniken der CodeAnnotator am besten geeignet ist. Bei dieser Evaluation wurde nämlich nur die Annotation mit wissenschaftlichen Dokumenten untersucht.

5 Anhang

Feedback Fragebogen zum CodeAnnotator

Würden Sie das Tool auch privat nutzen oder weiterempfehlen?
☐ ja ☐ nein
Falls nicht, warum nicht?
Haben Sie schon jemals einen Text annotiert?
☐ ja ☐ nein
Wie empfanden Sie das annotieren im Allgemeinen? (1= sehr einfach, 6=überhaupt nicht einfach)
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6
Wie oft haben sie mit wissenschaftlichen Dokumenten und Referenzangaben zu tun? (1= selten, 6=sehr häufig)
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6
Wenn Sie sich selbst einschätzen, wie gelang Ihnen das manuelle Annotieren (1. Methode)? (1= sehr gut, 2 = gut, 3 = eher gut, 4 = eher schlecht, 5 = schlecht, 6=sehr schlecht)
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6
Wenn Sie sich selbst einschätzen, wie gelang Ihnen das Annotieren mit KWIC (2. Methode)? (1= sehr gut, 2 = gut, 3 = eher gut, 4 = eher schlecht, 5 = schlecht, 6=sehr schlecht)
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6
Würden Sie ein nächstes Mal die KWIC Methode statt der manuellen Annotation verwenden?
☐ ja ☐ nein
Falls nicht, warum nicht
Haben Sie noch weitere Anmerkungen, inwiefern der CodeAnnotator verbessert werden kann?

Abbildung 5.1: Fragebogen zur Annotation

Literaturverzeichnis

[1] fortext literatur digital erforschen - glossar, 2016.

[2] S. Anna, T. Simone, and S. Christine. Guidelines für das tagging deutscher textcorpora mit stts (kleines und großes tagset, 1999.

[3] S. Buchholz and E. Marsi. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164, 2006.

[4] R. Cole, I. Chief, J. Mariani, H. Uszkoreit, A. Zaenen, V. Zue, G. Varile, and A. Zampolli. Survey of the state of the art in human language technology. *Language*, 76, 01 1997.

[5] D. C. Comeau, R. Islamaj Doğan, P. Ciccarese, K. B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, F. Rinaldi, M. Torii, et al. Bioc: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013, 2013.

[6] J. Didakowski, A. Geyken, and T. Hanneforth. Eigennamenerkennung zwischen morphologischer analyse und part-of-speech tagging: Ein automatentheoriebasierter ansatz. *Zeitschrift Fur Sprachwissenschaft - Z SPRACHWISS*, 26, 01 2007.

[7] S. Eisenbeiss, T. F. Jaeger, F. Lüpke, I. Ibarretxe-Antuñano, A. Kopecka, I. Tragel, J. Bohnemeyer, T. Nikitina, G. Montero-Melis, B. Narasimhan, and S. Kita. Satellite- vs. verb-framing underpredicts nonverbal motion categorization: Insights from a large language sample and simulations. *Cognitive Semantics*, 3:36–61, 02 2017.

[8] U. Engel. Heinz j. weber: Dependenzgrammatik : ein arbeitsbuch, tübingen, narr, 1992, 152 s., 1993.

[9] N. C. for Biotechnology Information. Pubmed.

[10] K. Ganchev, F. Pereira, M. Mandel, S. Carroll, and P. White. Semi-automated named entity annotation. In *Proceedings of the Linguistic Annotation Workshop*, pages 53–56, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[11] L. Humbert, A. Best, P. Micheuz, and L. Hellmig. Informatik–kompetenzentwicklung bei kindern. *Informatik Spektrum*, pages 1–9, 2020.

[12] i. p. d. w. t. C. p. Know-Center GmbH. Code annotator tool, 2012-2014.

[13] D. Kwon, S. Kim, S.-Y. Shin, and W. J. Wilbur. Bioqrator: a web-based interactive biomedical literature curating system. In *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, volume 1, pages 241–246, 2013.

[14] M. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. 1993.

[15] T. McEnery and A. Wilson. Research issues in applied linguistics, 1997.

[16] A. Mikheev, C. Grover, and M. Moens. Description of the LTG system used for MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*, 1998.

[17] S. Natarajan, A. Soni, A. Wazalwar, D. Viswanathan, and K. Kersting. Solving Large Scale Learning Tasks. Challenges and Algorithms, Essays Dedicated to Katharina Morik on the Occasion of Her 60th Birthday. pages 331–345, 2016.

[18] M. Neves and J. Ševa. An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, 12 2019. bbz130.

[19] M. Pelay-Gimeno, A. Glas, O. Koch, and T. N. Grossmann. Structure-based design of inhibitors of protein–protein interactions: Mimicking peptide binding epitopes. *Angewandte Chemie International Edition*, 54(31):8896–8927, 2015.

[20] J. Pustejovsky and A. Stubbs. Natural language annotation for machine learning. 2012.

[21] U. Reichel. Sprachsynthese: Part-of-speech-tagging, 2016.

[22] E. F. Sang and F. De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

[23] D.-I. F. L. Stefan and L. Nico. Was ist natural language processing?, 2016.

[24] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, Apr. 2012. Association for Computational Linguistics.

[25] M. und Informationszentrum der ZHdK. Literaturverwaltung mit mendeley, 2013.

[26] N. Xue, F.-D. Chiou, and M. Palmer. Building a large-scale annotated Chinese corpus. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[27] D. Zeng, K. Liu, Y. Chen, and J. Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762, 2015.

[28] C. Zimmer and B. Schinzel. Informatik-frauen. *Freiburger FrauenStudien*, (1):223–238, 1998.