Heinrich-Heine-Universität

Philosophische Fakultät

Abteilung der Computerlinguistik

Bachelor-Programm

Projektarbeit

Thema: Anwendung und Auswertung eines Coreference Resolution trainiertes Modell von Hugging-Face

verfasst von

Aldi Halili, Matr. Nr.: 3013735

Fach: Textverstehen mit Transformer Modellen SS. 2022

Betreuer: Prof. Dr. Wiebke Petersen

1 Inhaltsverzeichnis

1 Einleitung

Natural Language Processing (NLP) ist ein umfangreicher Bereich, der viele Herausforderungen mit sich bringt. Eine davon ist die Koreferenzauflösung, auch Coreference Resolution (CR) genannt. Die CR-Aufgabe der besteht darin, alle sprachlichen Ausdrücke (Mentions) in einem Text zu finden, die sich auf dieselbe Entität beziehen. Alle diese vorkommenden referierenden Ausdrücke, die sich zur Reale Entitäten beziehen, werden verknüpft. Dieses Verfahren wird Koreferenzauflösung (CR) bezeichnet. Um dieses Verfahren maschinell durchzuführen, wurden im Laufe der Jahre verschiedene Methoden und CR-Architekturen angewendet. (Kevin Clark, 2015)

Um einen umfassenden Überblick über die gebildeten Systeme der CR zu schaffen, werden in diesem Projekt alle grundlegende CR- Modelle und Architekturen bezüglich ihrer Funktionalität und Problemlösung erläutert. Des Weiteren werden die modernen Architekturen wie die Feed-Forward-Neuronale Netzwerke und End-to-End Methoden vorgestellt, die mithilfe modernerer Maschinelles-Lernen und Deep-Learning Verfahren im Rahmen der Künstlichen Intelligenz (KI) eine bessere Lösung für die CR-Aufgabe bieten.

Wichtig zu erwähnen ist es, dass die CR eine umfangreiche Aufgabe in NLP ist und in verschiedenen NLP-Anwendungen wie Maschinelle Übersetzung, Sentiment-Analysis, Dokument- Summarization und anderen eingesetzt werden kann. (Jurafsky & Martin, 2023)

Das vorliegende Projekt zielt darauf ab, ein ausgewähltes, von Hugging-Face vortrainiertes Modell namens „Neuralcoref“ [1] zu testen. Es wurden verschiedene Phänomene untersucht, um eine detaillierte Bewertung der Leistung dieses CR-Modells zu erstellen. Der Test wurde auf Google Colab [2]durchgeführt.

Des Weiteren wurde von mir ein selbstprogrammiertes Evaluations-Systems verwendet, um eine genauere Bewertung des Modells durchzuführen, wobei Precision, Recall und F-Score als Bewertungsmaßnamen angewendet werden. Die Ergebnisse werden für jedes einzelne Phänomen separat prozentuell bewertet, um die Schwachstellen oder Stärken des Models zu identifizieren.

[1] huggingface/neuralcoref: ✨ Fast Coreference Resolution in spaCy with Neural Networks (github.com)
[2] Neurocoref_Projekt_Aldi_Halili.ipynb - Colaboratory (google.com)

2 Was ist Coreference Resolution

Unter Koreferenz versteht man die Beziehengen zwischen zwei linguistischen Ausdrücken, die auf dieselbe Entität in einem Diskurs-Modell verweisen. (Karttunen, 1969)
CR hat zwei Hauptaufgaben. Als erstes müssen all diese linguistischen Ausdrücke (Mentions oder Erwähnung von Entitäten) im Text identifiziert und erkannt werden. Nach dieser Erkennung werden diese Mentions verknüpft und in einer Koreferenzkette gruppiert (geclustert). Hierbei werden alle linguistischen Eigenschaften und deren dazugehörige grammatikalische Kategorien sowie die Koreferenzbeziehungen zwischen den Mentions (wie z.B. pronominale Anaphern, bestimmte und unbestimmte Nominalphrasen usw.) berücksichtigt.

Es gibt drei Arten von Mentionserkennung: Pronomen (z.B. I, your, she, he, him, he, it. Usw.), Entitäten (z.B. Menschen, Orten wie Victoria Chen, Apple) und Nominalphrasen (z.B. a car, the red car.). Anhand eines Beispiels wird unten eine Erläuterung der CR vereinfacht, wie die linguistischen Ausdrücke analysiert werden, welche Bezeichnung die Mentions bekommen und wann sie koreferent sind und wann nicht. (Kevin Clark ,2015)

"Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks." (Jurafsky & Martin, 2023, Chapter 26, S1)

In unserem Beispiel oben sind folgende Koreferenzketten entstanden:

{"Victoria Chen", "her", "the 38-year-old", "She"}
{Megabucks Banking", "the company, Megabucks"}
{"her pay"}
{"Lotsabucks"}

Wichtige Fachbegriffe hinsichtlich der CR werden in diesem Abschnitt erläutert. Wie bereits erwähnt, sind mindestens zwei oder mehrere Mentions erforderlich, die auf dieselbe Entität (Bezugswort) referieren, um eine Koreferenz bilden zu können. In einem vorgestellten Diskurs gibt's auch *Anaphora*. Wenn eine Entität zum ersten Mal im Text erwähnt wird, wird sie als *Antezedens* bezeichnet, z.B. Victoria Chan. Nachfolgende Mentions werden als *Anapher* bezeichnet und sind in unserem obigen Beispiel die Mentions "She", "her". und die bestimmte

Nominalphrase “the 38-year-old”. Sie verweisen auf dieselbe Referentin, sind aber nicht die erste Mention dieser Referenz.

Das Gegenteil von Anaphora ist Kataphora. Dies bedeutet, dass die erste Mention im Text ein Pronomen ist und die Nominalphrase, die sich auf dieses Pronomen referiert zu nachfolgenden Mentions gehört. Z.B “because he is a good student, John has always loved learning."
Das Pronomen „he“ ist eine Kataphora und der Name „John“, der sich auf Kataphora referiert, wird als Postezedens bezeichnet.

Außerdem gibt es auch Entitäten in einem Text, die nur einmal erwähnt werden und keine anderen Referenzen haben. Diese Mentions werden als Singleton bezeichnet und in unserem obigen Beispiel ist das Wort “Lotsabucks”. Die Singeltons sind nicht koreferent. (Jurafsky & Martin, 2023)

3 Coreference Resolution Haupt- Modellen und Architekturen

Um Coreference Resolution Systeme zu entwickeln, wurden im Laufe der Jahre verschiedene Modelle und Architekturen eingesetzt, wie zum Beispiel nicht-neuronale statistische Klassifikatoren, einfache neuronale Netzwerke sowie fortgeschrittene Modelle wie LSTM, Attention und Transformers. (Jurafsky & Martin, 2023)

3.1 Rule base (pronominal anaphora resolution)

Das bekannteste Modell für die CR ist der sogenannte „Hobbs naive Algorithmus“, der im Jahre 1976 entwickelt wurde. Dieser Algorithmus basiert sich hauptsächlich auf dem syntaktischen Parse-Baum der Sätze. Dieser Algorithmus ist ein Regel-basierter Ansatz zur CR und versucht die Bedeutung vom Text zu erfassen, wobei er Pronomen und Nominalphrasen auf mögliche Antezedenzien zurückführt.
Hobbs naive Algorithmus wurde für die Pronomen Anaphora-Resolution lange Zeit verwendet, bis die entwickelte Maschinelles Lernen basierte Systeme für die CR zunutze gemacht wurden. Dieser Algorithmus war einfach zu implementieren und konnte menschliches Wissen miteinbeziehen, da er regelbasiert war. Allerdings wird er auch benachteiligt, weil er auf feste definierte Regeln angewiesen ist und somit zu Fehlern führen kann. Zudem erkennt er nicht alle Phänomene und versteht den Text auf pragmatische oder semantische Weise nicht gut genug. (Jurafsky & Martin, 2023)

3.2 Mention-Pair Modell

Diese Methode basiert auf dem Training eines binären Klassifikators, der die Wahrscheinlichkeit von zwei Mentions berechnet und darauf basierend entscheidet, ob sie koreferent sind.

Das Modell funktioniert in folgenden Schritten: Erstens müssen die Mentions-Paare im Satz miteinander verglichen werden. Diese Mentions-Paare bestehen aus zwei Komponenten, nämlich aus einem Antezedenz- Kandidaten und einem Anapher-Kandidaten. Das Modell bestimmt anhand der Wahrscheinlichkeitsberechnung, ob die beiden Mentions miteinander übereinstimmen und somit koreferent sind. Durch die Koreferenz-Wahrscheinlichkeiten wird für jedes mögliche Mention-Paar ein Wert zugewiesen, wobei für die koreferente Paare ein Wert grösser als 0.5 vorhergesagt wird und somit als positiv bezeichnet wird. Anderseits entsprechen die Werte, die kleiner als 0.5 sind, einer negativen Bewertung und gehören nicht zu koreferente Paaren.

Einerseits ist dieses Modell vorteilhaft, weil es eine einfache CR-Architektur hat und leicht zu implementieren lässt. Als Nachteile zählen hier zwei wichtige Probleme, die dazu führen, dass die CR nicht gut gelöst werden kann. Der Klassifikator kann nicht direkt den Antezedenz-Kandidaten zueinander vergleichen, weil er nicht trainiert worden ist, um zu entscheiden, welche von denen anhand der Wahrscheinlichkeit ein Antezedens ist. Zum anderen ignoriert das Modell das Diskursmodell vollständig, und basiert sich nur auf Mentions, sodass es nicht in der Lage ist, andere Mentions derselben Entitäten zu berücksichtigen. (Jurafsky & Martin, 2023)

3.3 Mention-Ranking Architektur

Diese Architektur ermöglicht es, die Antezedenzkandidaten direkt miteinander zu vergleichen, was dazu führt, dass jedem potenziellen Paar von Antezedenzien und Koreferenten ein Score zugewiesen wird. Am Ende dieses Verfahrens wählt das Modell das Paar mit dem besten Score aus. Es handelt sich um denselben Prozess wie bei Mention-Paar Modell, mit der Ausnahme, dass jede Anapher nur einem Antezedenten mit dem besten Score, zugeordnet wird. Um den besten Score zu finden, wird die Softmax Funktion angewendet. Die Lernwahrscheinlichkeit der Mention-Pair und Mention-Rank Architekturen wird auf gleiche Weise berechnet, wie bei anderen maschinellen Lernarchitekturen. (Jurafsky & Martin, 2023)

3.4 Einfaches Neuronales Netzwerk

Zu dieser Kategorie gehört das neuronale Mention-Paar Modell, welches ein standardmäßiges Deep Feedforward neuronales Netzwerk ist. Dieses Modell verwendet vortrainierte Word Embedding Features und kategoriale Features als Input, um Ähnlichkeiten zwischen Antezedent-Kandidaten und Mentions zu berechnen. Die erwähnten Features füttern die feedforward Layers des neuronalen Netzwerkes und geben einen endgültigen Score aus. Insgesamt gibt es drei hidden Layers und ReLU-Funktionen, die die Wahrscheinlichkeiten berechnen, ob die Mentions koreferent sind oder nicht. Clark & Manning (2016) haben ein erfolgreiches Modell namens „neural coref“ entwickelt, welches gute Ergebnisse erzielt hat. Darüber hinaus wird das Modell mit moderneren Mechanismen wie LSTM und Attention verbessert, um eine bessere Leistung zu erzielen. Mehr dazu wird in folgenden Abschnitten erklärt.

3.5 End-to-end Neuronale Koreferenz-Modell

Die fortgeschritteneren gebildete Systeme für CR sind die end-to-end Neuronale Modelle. Als Verbesserung der einfachen Neuronalen Netzwerk werden hier zusätzlich andere Methoden angewendet wie z.B. LSTM, Attention. Normalerweise sind diese Systeme end-to-end, das bedeutet, dass sie keinen separaten Mention-Erkennung Prozess durchführen, sondern stattdessen betrachtet dieses Modell jeden möglichen Span von Text als Mention. Als Eingabe füttert man den rohen Text und endet als Ausgabe mit den besten Score Paaren von Anaphern und Antezedenten. (Kenton Lee et al. 2017)

Diese Methode funktioniert in folgenden Schritten: Erstens erhält das Modell als Input die Wortsequenzen, die von Wort-Embeddings repräsentiert werden. Zunächst werden die Word-Embeddigs als Input für die Anwendung eines bidirektionalen LSTM-Modells gefüttert, das daraus eine hidden Layer für jeden möglichen Textspan erzeugt. Die erzeugten Span-Repräsentationen werden in zwei unterschiedliche Feed Forward Neuronalen Netzwerke (NN) weitergeleitet. Das erste NN ist für die Mention-Erkennung zuständig und gibt als Ausgabe eine Wahrscheinlichkeit aus, ob ein bestimmter Textspan eine Mention ist oder nicht. Das zweite NN ist für die Identifikation des Antezedenten zuständig und gibt als Ausgabe eine Wahrscheinlichkeit dafür aus, ob ein anderer Textspan als Antezedent für die aktuelle Mention sein könnte oder nicht.

Anschließend erhält man die Mentions- und Antezedens-Scores, die sich aus beiden Netzwerken ergeben und zusammen summiert werden, um einen Koreferenzscore zu berechnen. Um die endgültige Koreferenzketten zu erhalten, wird für die Inputsequenzen berechneten Koreferenzscores eine Softmax-Funktion angewendet. (Jurafsky & Martin, 2023)

4 Neoralcoref Modell von Hugging-Face

Das von mir für das Projekt ausgewählte vortrainierte Modell heißt Neuralcoref. Dieses neuronale CR-System [3]wurde für Chatbots von offiziellem Entwickler von Hugging-Face [4]als Open-Source-Software entwickelt und ist in spaCy NLP-Pipeline integriert.
Frühere traditionelle Algorithmen für die CR-Aufgabe verwendeten ursprünglich handgefertigte linguistische Features, die sehr umfangreich waren. Die Magie moderner NLP-Techniken besteht darin, handgeschriebene Features automatisch generieren zu können. Das Neuralcoref-Modell basiert sich auf der Anwendung von Wortvektoren und neuronalen Netzwerken.

Für dieses Model wurden Word Embeddings angewendet und es wurde auf dem größten annotierten Korpus der Coreference Resolution (OntoNotes 5.0-Datensatz[5]), trainiert. Dieser Datensatz wurde grundsätzlich aus Nachrichten und Webartikeln erstellt, die normalerweise eine formellere Sprache als Chatbot-Benutzer enthalten. Das bedeutet, dass dieses Koreferenzsystem einerseits für einige Mentions-Paare, die durch eine formale Sprache repräsentiert werden, gut funktioniert. Anderseits kann dieses System für die Mentions-Paare, die weniger formal sind, möglicherweise fehlschlagen.

Word Embeddings beinhalten Informationen über Geschlecht der Substantive, aber allein diese Informationen reichen normalerweise nicht aus, um eine CR durchzuführen. Deshalb werden zusätzliche Informationen wie z.B. der Mentions-Kontext und externes Allgemeinwissen hinzugefügt. (die Länge der Mentions, Ort der Mentions, Informationen über den Sprecher usw.)

Anschließend werden diese Repräsentationen in zwei neuronale Netzwerke weitergeleitet. Das erste neuronales Netz gibt ein Score für alle Mentions-Paare, sowie eine mögliche Antezedenten

[3] State-of-the-art neural coreference resolution for chatbots | by Thomas Wolf | HuggingFace | Medium
[4] Hugging Face – The AI community building the future.
[5] OntoNotes Release 5.0 - Linguistic Data Consortium (upenn.edu)

aus, während das zweite NN für Mention ohne Antezedenten ein Score ausgibt. Am Ende wird der höchste Score verwendet, um zu bestimmen, ob eine Mention eine Antezedenten hat und welche davon ist.

Dieses Modell wurde auf eine nicht probabilistische „slack-rescaled-max-margin" Ziel trainiert. Das bedeutet, dass das System ein Score für jedes Mentions-Paar und ein Score für jede einzelne Mention berechnet. Diese Scores sind keine Wahrscheinlichkeiten, sondern nur Punktzahlen in einer beliebigen Einheit (je höher, desto besser).

5 Testen des neuralcoref Modells und Koreferenz-Phänomenen:

Für dieses Modell wurden verschiedene Phänomene getestet, um zu sehen, wie diese Phänomene vom Neuralcoref erkannt und aufgelöst werden. Die Hauptphänomene, die getestet wurden, waren Pronomen, bestimmte und unbestimmte Nominalphrase, sowie Named Entities. Diese Phänomene gehören zu referierenden Ausdrücken und sie werden als Anaphora, Kataphora, koreferente Nominalphrasen, gebundene Antezedenten und gebundene Variablen bezeichnet. Darüber hinaus gibt es auch nicht-referierende linguistische Ausdrücke, die von allen CR-Systemen erkannt werden müssen, um die CR-Aufgabe auflösen zu können. Diese sogenannten nicht-referierende Ausdrücke sollen vom CR-System nicht als Mention erkannt werden. Dazu gehören Appositionen, prädikative NPs, Generics, Expletives wie Clefts, Extraposition und Pleonastik.
Für das Testen und die Evaluation wurden die Textbeispiele aus Wikipedia [6]und dem 26. Chapter der Coreference Resolution [7]von Jurafsky & Martin Buch aus Stanford Universität verwendet. Damit die gewünschte CR-Phänomen mit dem Text angepasst werden könnten, wurden die Sätze von mir teilweise bearbeitet und geändert.

Um die Leistungsfähigkeit des Modells bewerten zu können, wurde die Evaluation anhand der metrischen Bewertung der Precision und Recall angewendet. Precision[8] in CR gibt an, wie viele der vom Modell als Koreferentpaare tatsächlich korrekt sind. Eine hohe Precision bedeutet, dass das System nur wenige falsche Koreferenzpaare gefunden hat und somit genauer arbeitet. Der Recall bezieht sich darauf, wie viele der vom Golden Standard Koreferenzpaare vom System

[6] Coreference - Wikipedia
[7] 26.pdf (stanford.edu)
[8] Precision and recall - Wikipedia

gefunden wurden. Ein höherer Recall-Wert bedeutet, dass das System in der Lage ist, mehr relevante Koreferenzpaare zu erkennen. Der F-Score [9]ist das harmonische Mittel zwischen Precision und Recall und wird berechnet, um eine Aussage darüber zu treffen, wie gut das System sowohl präzise als auch umfassend arbeitet. Ein höherer F-Score bedeutet, dass das System sowohl präzise als auch umfassend arbeitet und somit besser in der Coreference Resolution-Aufgabe ist. Die Werte des F-Scores variieren von 0 bis 1, wobei 1 die beste Leistung ist und 0 auf eine nicht korrekte Vorhersage hinweist. (Goutte & Gaussier, 2005)

5.1 Anaphora

Anaphorische Pronomen gehören zu den Hauptphänomenen für die CR und sie referieren auf Nominalphrasen oder Entitäten. Eine wesentliche Voraussetzung für ein gut funktionierendes CR-System besteht darin, die korrekten Antezedenten für jedes anaphorische Pronomen im Text zu identifizieren. (Jurafsky & Martin, 2023)

Um dieses Phänomen zu evaluieren, wurden zehn anaphorische Sätze für das Neuralcoref-Modell getestet. Für die getesteten Sätze scheint das Modell die Pronominalresolution mit einem F-Score von 0.85 einer Precision von 0.91 und einem Recall von 0.80 gut zu lösen. Es ist zu beachten, dass dieser Score textabhängig ist, was bedeutet, dass das Ergebnis sich verbessert oder verschlechtert werden kann, wenn das Modell mit anderen Daten getestet wird.

5.2 Kataphora

Die Kataphorische Resolution könnte durch das Neuralcoref-Modell anhand der getesteten Satzbeispiele nur teilweise gelöst werden. Das Modell funktioniert in der Regel gut, wenn die Beziehungen zwischen den beiden linguistischen Ausdrücken eindeutig sind, allerdings hängt die Identifizierung stark vom Kontext ab. Beim Testen von 10 Sätzen wurden nur 5 richtig erkannt und es wurde F-Score von 0.60 erreicht, wobei die Precision 0.90 und der Recall 0.45 betrugen.

5.3 Gebundene Variablen (bound anaphors)

Die gebundenen Variablen sind linguistische Ausdrücke, die durch Quantoren gebunden sind. In der Realität sind die gebundenen Variable nicht koreferent, sie müssen aber trotzdem bei CR

[9] F-score - Wikipedia

mit Antezedenzien verknüpft werden, um die Bedeutung im Text zu verstehen. (Jurafsky & Martin, 2023)

Um die Fähigkeit des CR-Systems zur Erkennung von gebundenen Variablen zu evaluieren, wurden 10 Sätze getestet. In 9 Fällen wurden die gebundenen Variablen korrekt erkannt, was auf eine gute Leistung des Systems hinweist. Das Ergebnis zeigte einen F-Score von 0,84, eine Precision von 0,89 und einen Recall von 0,80.

5.4 Gebundene Antezedenten

Die gebundene Antezedenten (Split Antecedents) sind zwei oder mehrere NPs, wobei auf die nur ein Pronomen verweist. Das Neuralcoref-Model erkennt diese Phänomene beim Testen unserer Beispielsätze sehr gut. Es wurden fünf Sätze getestet, die alle als coreferent markiert wurden und zufriedenstellende Ergebnisse lieferten. Es gibt aber jedoch einige Koreferentpaare, die von Koreferenzkette weggelassen wurden, wodurch insgesamt Grund der F-Score bei 0.89, der Precision bei 1.00 und der Recall bei 0.80 liegen.

5.5 Koreferente Nominalphrasen

Es scheint für das neuralcoref-Modell eine Herausforderung zu sein, koreferente Nominalphrasen zu erkennen. Das Modell erkennt Sätze mit bestimmten NPs, die "the" enthalten, nicht als koreferent. Demonstrative Nominalphrasen in unserem Textbeispiel erkennt das System teilweise gut, während andere NPs falsch annotiert werden können. Ein Beispiel für einen falsch annotierten Satz falsch annotierter Satz ist folgender: Input: "Some of our colleagues are going to be supportive. These kinds of people will earn our gratitude. Die Ausgabe lautet "Output: [our: [our, our]]. Wenn man das obige Textbeispiel betrachtet, erkennt man, dass die Annotation falsch gelabelt wurde. Die zweite Nominalphrase "These kinds of people" bezieht sich auf die erste Nominalphrase "Some of our colleagues".

Beim Testen erkennt das Modell von fünf Sätzen nur zwei richtig und ergibt einen Score von 0,44, einen Recall von 0,40 und eine Precision von 0,50, was keine zufriedenstellenden Ergebnisse liefert.

5.6 Testen einiger nicht referierenden Ausdrücken

Es gibt auch einige Nicht referenzierende Ausdrücke wie Nominale Phrasen oder andere Nomen, die diese Aufgabe erschweren. Diese Ausdrücke sollten nicht als Mention annotiert werden und werden nicht als koreferent bezeichnet. Dazu gehören Appositionen, prädikative NPs, Generics und Expletives, die ich für Neuralcoref-Modell getestet habe.

5.6.1 Appositionen

Appositionelle NPs sind keine referierenden Ausdrücke, sondern sie sind Nomen oder Nomens-Gruppe, die ein Bezugswort näher beschreiben. Die Appositionen stehen direkt hinter dem Bezugswort. (Jurafsky & Martin, 2023)

Alle fünft oben getesteten Sätze werden vom System erkannt, dass es sich nicht um eine Korefenz handelt, sondern um eine appositionelle NP. Dabei ist in unserem ersten Satzbeispiel zu betrachten, dass das Bezugswort John und die appositionelle NP miteinander verknüpft werden und es keine verweisenden Ausdrücke gibt, die zu einer Entität beziehen, sondern dass „a student“ zu „my youngest brother John“ referiert.

5.6.2 Prädikative Nominalphrasen.

Die prädikative NPs beschrieben die Eigenschaften des Hauptsubstantivs, in unserem Beispiel: „In United is a unit of UAL“. Die Nominalphase „a unit of ual“ beschreibt die Eigenschaft von United und referiert nicht auf eine Entität im Satz. In unseren Textbeispiele erkennt das System alle vier getestete prädikative NPs als richtig. Das System gibt jedoch ein „False“ aus, weil diese Sätze nicht als Mention markiert werden sollten. (Jurafsky & Martin, 2023)

5.6.3 Generische Nominalphrasen

Generics sind andere Ausdrücken, die nicht explizit auf eine bestimmte Entität referiert, sondern sie referieren den realen Welten im Allgemeinen. Das System erkennt drei von vier Sätzen in unseren Textbeispielen als richtig, wobei es für die letzten drei Sätze ein "False" und den ersten Satz ein "True" ausgibt.

Beispiel-Satz: "I love mangos. They are very tasty."

Outout: [mangos: [mangos, They]]

Beim obigen Satz macht das System eine falsche Zuordnung, wobei "They" auf "mangos" mit einem hohen Score referiert. Das Wort "They" bezieht sich jedoch nicht explizit auf bestimmte Mangos, sondern auf die Kategorie von Mangos im Allgemeinen. (Jurafsky & Martin, 2023)

5.6.4 Expletives

Expletives gehören zu den nicht-anaphorischen pronominalen Referenzen. Eine Herausforderung für die moderne CR-Systeme ist die Identifikation und Eliminierung dieser Mentions, die keine Referenz tragen, bzw. nicht auf einen Antezedent verweisen. Zu diesen Referenzen gehören Clefts, Expletive oder Pleonastic "it". (Jurafsky & Martin, 2023)

5.6.4.1 Clefts

Clefts Sätze werden als komplexer Sätze betrachtet, wobei eine Bedeutung hat, die mit einem einfachen Satz ausdrücken lässt. Im Englischen ist oft der Fall, dass das Pronomen "it", der redundant und keine eigene Bedeutung trägt. Für dieses Phänomen wurden vier Sätze getestet und alle richtig mit einem "False" erkannt.

5.6.4.2 Extraposition

Extrapositionen gehören auch zu komplexen Phänomenen. Sie werden verwendet, um die Struktur des Satzes zu verändern, wobei sie ein Element, das üblicherweise am Anfang des Satzes stehen würde, am Ende des Satzes verschoben wird. Die Extrapositionen müssen für die CR-Aufgabe auch nicht annotiert werden, sondern mit "False" erkannt werden. Normalerweise fällt CR-Systeme im Allgemeinen diese Phänomene schwer zu identifizieren, da dieses Wort entfernt vom Bezugswort im Satz sein kann. Das Neuralcoref-Modell scheint mit solchen Phänomenen keine Probleme zu haben, da es alle vier getesteten Sätze richtig mit einem "False" erkannt hat.

5.6.4.3 Pleonastic

Dieses Pronomen, genannt „pleonastic“ trägt auch keine Referenz und werden wie andere nicht-referierende Ausdrücke für die CR-Systeme nicht annotiert. Es wurden auf dieses Phänomen vier Sätze getestet, die vom Modell richtig erkannt und mit "False" klassifiziert wurden.

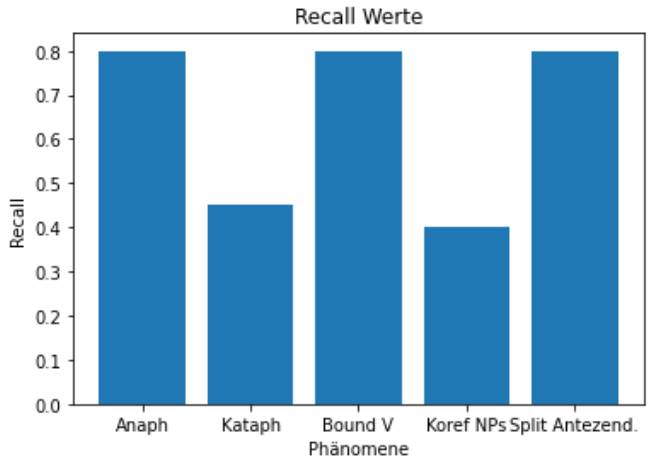
6 Ergebnisse

Das Ergebnis des durchgeführten Projekts lässt sich anhand der Darstellungen Grafik 1, Grafik 2 und Grafik 3 erläutern. Das Neuralcoref-Modell liefert für die getesteten Phänomene zufriedenstellende bis hervorragende Ergebnisse. Gebundene Antezedenten haben einen F-Score von 89.0 erreicht, was das beste Ergebnis im Vergleich zu anderen Phänomenen ist. Anaphora und gebundene Variablen haben auch gute Ergebnisse mit einen F-Score über 0.84 erzielt. Siehe Grafik 3.
Kataphora Phänomen wurde mit einem leicht erhöhten durchschnittlichen Wert, insgesamt mit einem F-Score von 0.60 bewertet, wobei der Recall mit einem nicht zufriedenstellenden Score von 0.45 benachteiligte. Im Gegensatz dazu wurde der Precision-Score für Kataphora mit einem Wert von 0.90 bewertet. Siehe Grafik 1 und 2.
Die koreferenten Nominalphrasen wurden mit einem Score von 0,44 am niedrigsten bewertet, was bedeutet, dass das Modell Schwierigkeiten bei diesem Phänomen hat. Siehe Grafik 3.

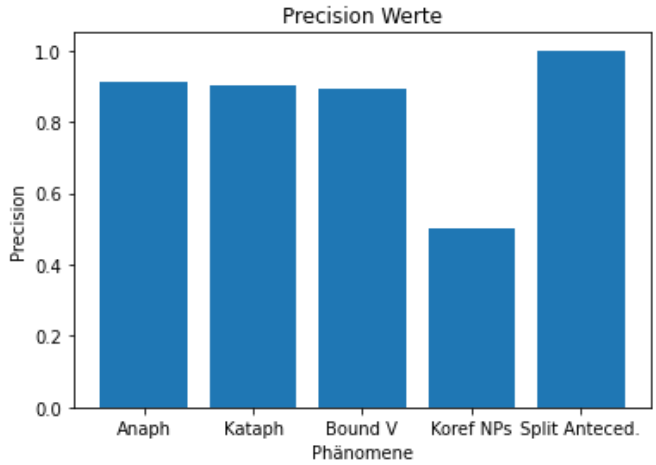
Für die folgenden Phänomene wurden hervorragende Ergebnisse erzielt, wobei das System fast alle Probleme erkannt hat. Nachfolgend findet man eine Zusammenfassung der Ergebnisse dieser Phänomene:

Das Neuralcoref-Modell hat für die Expletive-Phänomene alle 12 getesteten Sätze als korrekt erkannt und mit "False" klassifiziert, da diese Sätze, wie bereits erwähnt, keine Referenzen enthalten. Auch die Phänomene wie Appositionen und prädikative Nominalphrasen wurden vom System korrekt erkannt. Bei den Generics wurden drei von vier Sätzen richtig erkannt. Zusammenfassend kann man sagen, dass das Neuralcoref-Modell für unsere Textdaten im Allgemeinen zufriedenstellende bis sehr gute Ergebnisse liefert.

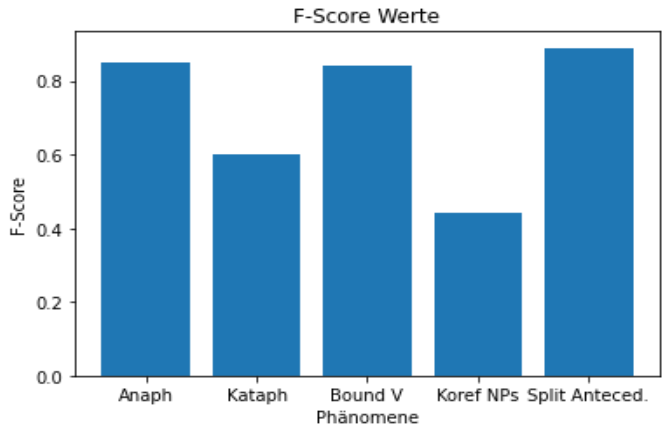
7 Anhänge



Grafik 1 Recall-Score



Grafik 2 Precision-Score



Grafik 3 F-Score

8 Literaturverzeichnis

Daniel Jurafsky & James H. Martin (2023)[10]
Speech and Language Processing. Chapter 26, Auflage 3, Stanford University

Kevin Clark (2015)
Neural Coreference Resolution, Stanford University

Hobbs, Jerry R., 1978
Resolving Pronoun References, Lingua, Vol. 44, pp. 311-338.

Kevin Clark & Christopher D. Manning (2016)
Improving Coreference Resolution by Learning Entity-Level Distributed, Stanford University

Kenton Lee et al. (2017)
End-to-end Neural Coreference Resolution, g, Univ. of Washington, n, Seattle, WA

Lauri Karttunen. (1969)
Discourse Referents. In International Conference on Computational Linguistics COLING 1969: Preprint No. 69: Collection of Abstracts of Papers, Sånga Säby, Sweden.

Claire Cardie and Kiri Wagstaff. (1999)
Noun Phrase Coreference as Clustering. In 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Goutte & Gaussier (2005)
A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation Xerox Research Centre Europe, Meylan, France

Chris Manning (WS: 22/23)[11]
Natural Language Processing course with Deep Learning, Stanford University

Chris Manning (WS: 22/23)
Lecture 17. Coreference Resolution topic [12]Stanford University

[10] Speech and Language Processing (stanford.edu)
[11] Stanford CS 224N | Natural Language Processing with Deep Learning
[12] cs224n-2023-lecture17-coref.pdf (stanford.edu)