INSTITUT FÜR INFORMATIK
Datenbanken und Informationssysteme

Universitätsstr. 1 D–40225 Düsseldorf

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Webtool zur Betrachtung und Annotation von Satzalignments aus Wikipedia in vereinfachtem Englisch

Alex Galkin

Bachelorarbeit

Beginn der Arbeit:	19. Mai 2019
Abgabe der Arbeit:	19. August 2019
Gutachter:	Prof. Dr. Stefan Conrad
	Prof. Dr. Stefan Conrad

Erklärung

Hiermit versichere ich, dass ich diese Bachelorarbeit selbstständig verfasst habe. Ich habe dazu keine anderen als die angegebenen Quellen und Hilfsmittel verwendet.

Düsseldorf, den 19. August 2019

Alex Galkin

Zusammenfassung

Hier kommt eine ca. einseitige Zusammenfassung der Arbeit rein.

Inhaltsverzeichnis

1 Einleitung **1**

2 Methodik **1**

3 Vorverarbeitung vom Text **1**

3.1 Lowercasing und Entfernen von Interpunktion 1

3.2 Entfernen von Stoppwörtern 1

3.3 Lemmatisierung 2

4 De Bello Hispaniensi **3**

5 Weiteres Kapitel **5**

5.1 Unterkapitel 5

5.2 Unterkapitel 5

Abbildungsverzeichnis **6**

Tabellenverzeichnis **6**

1 Einleitung

lorem

2 Methodik

lorem

3 Vorverarbeitung vom Text

Als Ausgangslage liegen uns der Wikipedia Artikel in English und der Wikipedia Artikel in Simple English vor. Damit die Algorithmen präzisere Ergebnisse erzielen können müssen die beiden Texte vorbereitet und verarbeitet werden. Zuerst werden die Texte mit Hilfe von Spacy in einzelne Sätze zerlegt und danach in zwei Listen gespeichert. Da der Benutzer nur einen Satz aus dem English Artikel aussucht, wird um Rechenkraft zu sparen, nicht der gesamte Artikel weiterverarbeitet, sondern nur der vom Benutzer ausgewählte Satz. Der Simple English Artikel wird Satz für Satz analysiert werden und muss somit in seiner gesammten Länge vorverarbeitet werden. Die einzelnen Sätze werden in Wörter zerlegt und ebenfalls in einer Liste gespeichert. Somit haben wir den vom User ausgewählen Satz, aufgeteilt in einzelne Wörter in einer Liste und eine zweite Liste welche Listen von den einzelnen Sätzen welche einzelne Wörter enthält bei dem Simple English Artikel.

Ausgewählter Satz in Liste = ['Dies', 'ist', 'der', 'ausgewählte', 'Satz', '.']
Simple English Artikel in Liste = [['Dies', 'ist', 'ein', 'Wikipedia', 'Artikel','.'], ['Er', 'enthält', 'mehrere', 'Sätze', '.']]

Nun können wir mit der Vorverarbeitung von den Texten beginnen. Alle folgenden Prozesse wurden entweder mit der Python Standard Library oder mit spaCy durchgeführt.

3.1 Lowercasing und Entfernen von Interpunktion

Damit wir gleiche Wörter erkennen können ist Lowercasing der erste Schritt um eine Uniformität für den Vergleich von Strings gewährleisten zu können. Es dient vor allem dazu damit ein Wort am Anfang von einem Satz und das gleiche Wort in der Mitte eines Satzes auch als gleich erkannt werden. In dem Programm werden somit bei der weiteren Verarbeitung alle Wörter nur im Lowercase betrachtet. Desweiteren müssen alle Satzzeichen entfernt werden, da diese nicht von Hilfe für die Algorithmen sind

3.2 Entfernen von Stoppwörtern

Stoppwörter sind Wörter welche keine Relevanz für den Inhalt und Kontext des Satzes aufweisen. Überwiegend werden Synsemantika entfernt, dies sind im Englischen Wörter wie ”the”, ”is”, ”at” etc. Dies sind Wörter die nur eine grammatische Funktion im

Text haben. Somit erhalten wir eine Liste an Wörtern die alle eine für den jeweiligen Satz inhaltliche Bedeutung aufweisen. Dies erleichtert uns die Suche nach passenden Alignments da nicht mehr so viel ”Rauschen” in unseren Daten vorhanden ist.

3.3 Lemmatisierung

Um die Uniformität weiter zu verbessern wenden wir die sogenannte Lemmatisierung an. Ein Lemma ist die Grundform von einem Wort. Ein Beispiel für die Lemmatisierung sieht wie folgt aus: ”am”, ”are”, ”is” werden zu ”be”, ”helping”, ”helps”, ”helped” wird zu ”help”. Das Wort wird auf die Form zurückgeführt die man auch in einem Wörterbuch finden würde. Auch dieser Prozess hilft uns das Rauschen einzudämmen und Alignments zu finden welche sonst vielleicht nicht gefunden worden wären.

Con97 hat ein Buch geschrieben. Es gibt auch andere Arbeiten (**PeHe97**) die referenziert sind. In Abbildung 1 ist ein Sachverhalt dargestellt.

1 Autor: **Con97** (**Con97**)
2 Autoren: **IWNLP** (**IWNLP**)
3 Autoren: **liebeck-esau-conrad:2016:ArgMining2016** (**liebeck-esau-conrad:2016:ArgMining2016**)

Online resource: **ILSVRC2016**

quotes:
„quote“.

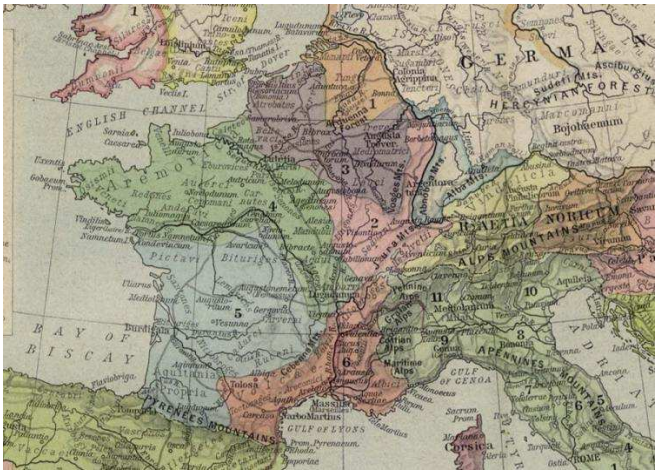

Abbildung 1: Gallien zur Zeit Caesars

Jahr	Erster Consul	Zweiter Consul
1	C. Caesar	L. Aemilius Paullus
2	P. Vinicius	P. Alfenus Varus
3	L. Aelius Lamia	M. Servilius
4	Sex. Aelius Catus	C. Sentius Saturninus
5	L. Valerius Messalla	Cn. Cornelius Cinna
suff.	C. Vibius Postumus	C. Ateius Capito
6	M. Aemilius Lepidus	L. Arruntius

Tabelle 1: Römische Konsulen

4 De Bello Hispaniensi

Pharnace superato, Africa recepta, qui ex his proeliis cum adulescente Cn. Pompeio profugissent, cum . . . et ulterioris Hispaniae potitus esset, dum Caesar muneribus dandis in Italia detinetur, . . . quo facilius praesidia contra compararet, Pompeius in fidem uniuscuiusque civitatis confugere coepit. Ita partim precibus partim vi bene magna comparata manu provinciam vastare. Quibus in rebus non nullae civitates sua sponte auxilia mittebant, item non nullae portas contra cludebant. Ex quibus si qua oppida vi ceperat, cum aliquis ex ea civitate optime de Cn. Pompeio meritus civis esset, propter pecuniae magnitudinem alia qua ei inferebatur causa, ut eo de medio sublato ex eius pecunia latronum largitio fieret. Ita pacis commoda hoste +hortato+ maiores augebantur copiae. +Hoc crebris nuntiis in Italiam missis civitates contrariae Pompeio+ auxilia sibi depostulabant.

C. Caesar dictator tertio, designatus dictator quarto multis +iterante diebus coniectis+ cum celeri festinatione ad bellum conficiendum in Hispaniam cum venisset, legatique Cordubenses, qui a Cn. Pompelo discessissent, Caesari obviam venissent, a quibus nuntiabatur nocturno tempore oppidum Cordubam capi posse, quod nec opinantibus adversariis eius provinciae potitus esset, simulque quod tabellariis, qui a Cn. Pompeio dispositi omnibus locis essent, qui certiorem Cn. Pompeium de Caesaris adventu facerent . . . multa praeterea veri similia proponebant. Quibus rebus adductus quos legatos ante exercitui praefecerat Q. Pedium et Q. Fabium Maximum de suo adventu facit certiores, utque sibi equitatus qui ex provincia fuisset praesidio esset. Ad quos celerius quam ipsi opinati sunt appropinquavit neque, ut ipse voluit, equitatum sibi praesidio habuit.

Erat idem temporis Sex. Pompeius frater qui cum praesidio Cordubam tenebat, quod eius provinciae caput esse existimabatur; ipse autem Cn. Pompeius adulescens Uliam oppidum oppugnabat et fere iam aliquot mensibus ibi detinebatur. Quo ex oppido cognito Caesaris adventu legati clam praesidia Cn. Pompei Caesarem cum adissent, petere coeperunt uti sibi primo quoque tempore subsidium mitteret. Caesar - eam civitatem omni tempore optime de populo Romano meritam esse - celeriter sex cohortis secunda vigilia iubet proficisci, pari equites numero. Quibus praefecit hominem eius provinciae notum et non parum scientem, L. Vibiurn Paciaecum. Qui cum ad Cn. P praesidia venisset, incidit idem temporis ut tempestate adversa vehementique vento adflictaretur; aditusque vis tempestatis ita obscurabat ut vix proximum agnoscere possent. Cuius incommodum summam utilitatem ipsis praebebat. Ita cum ad eum locum venerunt, iubet binos equites

conscendere, et recta per adversariorum praesidia ad oppidum contendunt. Mediisque eorum praesidiis cum essent, cum quaereretur qui essent unus ex nostris respondit, ut sileat verbum facere: nam id temporis conari ad murum accedere, ut oppidum capiant; et partim tempestate impediti vigiles non poterant diligentiam praestare, partim illo responso deterrebantur. Cum ad portam appropinquassent, signo dato ab oppidanis sunt reccepti, et pedites dispositi partim ibi remanserunt, equites clamore facto eruptionem in adversariorum castra fecerunt.

5 Weiteres Kapitel

5.1 Unterkapitel

5.2 Unterkapitel

Abbildungsverzeichnis

1 Gallien zur Zeit Caesars 2

Tabellenverzeichnis

1 Römische Konsulen 3