Praktikumsbericht

Machine Learning for Natural Language Processing

am Lehrstuhl für Informatik X

Konstantin Herud Thomas Schaffroth Maximilian Meißner

Abgabedatum:	16. August 2020
Betreuer:	Daniel Schlör
	Albin Zehe
	Konstantin Kobs
	Tobias Koopmann

Julius-Maximilians-Universität Würzburg
Lehrstuhl für Informatik X
Data Science

Zusammenfassung

Dieses Dokument soll Studenten an unserem Lehrstuhl bei der Erstellung ihrer Abschlussarbeit unterstützen. Wir zeigen eine beispielhafte Gliederung einer Arbeit und beschreiben die Inhalten der einzelnen Kapitel. Zusätzlich geben wir an vielen Stellen auch Hinweise zur Benutzung von LaTeX für die Erstellung der Arbeit. Im Anhang **??** geben wir ein paar Hinweise zum Ablauf der Betreuung von Abschlussarbeiten an unserem Lehrstuhl.

Zur Handhabung dieses Pakets. In diesem Paket sind Vorlagen für verschiedene Dokumenttypen enthalten, die sie als Ausgangspunkt für ihre Arbeit verwenden können. Es gibt jeweils Vorlagen für deutsche und englische Arbeiten.

- `template_thesis_de.tex`, `template_thesis_en.tex`: Vorlage für Bachelorarbeit bzw. Masterarbeit
- `template_seminar_de.tex`, `template_seminar_en.tex`: Vorlage für Seminarausarbeitungen und Praktikumsberichte

Der Quelltext zu diesem Leitfaden ist ebenfalls im Paket enthalten. Diesen können Sie als praktisches Beispiel dafür verwenden, wie diese Dokumentenklasse angewandt wird.

Inhalt der Zusammenfassung. Schreiben Sie hier eine Zusammenfassung der Arbeit, vergleichbar mit dem Abstract auf wissenschaftlichen Papers. Sie dient dem Leser dazu, einen groben Überblick über die Inhalte zu gewinnen (Problemstellung, verwendeter Lösungsansatz, ggf. experimentelle Ergebnisse, gewonnene Erkentnisse). Der Umfang soll ca. eine halbe Seite betragen. Für Seminararbeiten ist diese Zusammenfassung nicht erforderlich.

Achtung: Bei Arbeiten auf Englisch fordern die Prüfungsordnungen, dass es eine deutsche Zusammenfassung gibt. Schreiben Sie in diesem Fall eine englische *und* eine deutsche Zusammenfassung (mit dem gleichen Inhalt). Die passenden LaTeX-Befehle dafür finden Sie in den englischsprachigen Vorlagen.

WARNUNG: Die vorliegende Version des Leitfadens ist eine Vorabversion, die noch nicht vollständig ist. Sie bezieht sich größtenteils auf die Ausarbeitung von Bachelor- und Masterarbeiten; Seminararbeiten unterscheiden sich davon etwas in Aufbau und Inhalt.

Inhaltsverzeichnis

1. Methodik

Konstantin Herud

1.1. Task 1: Dataset Preparation

Da verschiedene Informationen das Herkunftsland von Autoren entweder direkt oder indirekt preisgeben, müssen diese aus der Datenbasis entfernt werden. Die Zielsetzung sieht hierbei vor Namen der Autoren, E-Mails, Institutionen und Firmen, Herkunftsländer, Förderungen und Danksagungen, Persönliche Daten sowie Referenzen zu entfernen. Referenzen sind beispielsweise indirekte Hinweise, weil Arbeitsgruppen der gleichen Herkunft und somit dem selben Fachgebiet die selbe Literatur referenzieren.
Da diese Aufgabe essentieller Bestandteil der Validität späterer Ergebnisse ist, wurde eine mehrstufige Daten-Pipeline entwickelt, um alle Informationen bestmöglich aus den Texten zu entfernen.

Textextraktion Im ersten Schritt muss der rohe Text der Dokumente eingelesen werden. Die Daten liegen sowohl in PDF- als auch in Textform vor, wobei letztere durch das Befehlszeilenprogramm *pdftotext* erstellt wurde und somit bereits Potenzial für Fehler bietet. Ursprünglich wurde deshalb ein Ansatz mit dem Werkzeug *pdfminer* verfolgt, um den Text der Dokumente selbstständig zu extrahieren und dabei direkt Strukturinformationen ausnutzen zu können (z. B. um Referenz-Blöcke zu erkennen). Bei einigen Dokumenten entstanden jedoch Fehler mit pdfminer, weshalb dieser Ansatz verworfen wurde. So werden in der Pipeline zunächst die Textdateien sowie zugehörige Metadaten eingelesen [1]. Da verschiedene spätere Schritt außerdem Informationen über die Zeichenposition des Abstract-Starts brauchen, wird dieser zudem bereits mit einem regulären Ausdruck, unterstützt durch verschiedene Fehlermaßnahmen, bestimmt. So wird beispielsweise überprüft ob der Titel des Dokuments das Wort Abstract enthält. Falls der Start nicht gefunden werden kann oder eine bestimmte Obergrenze überschreitet, wird ein mittelwert-basierter Defaultwert verwendet.

Dokumentkopf Als Dokumentkopf definieren wir jeglichen Text (in Rohform), der vor dem Abstract erscheint. Die hohe Dichte unterschiedlicher zu entfernender Informationen, gepaart mit Artefakten und Strukturfehlern, machen diese Stufe zu einer der größten Herausforderungen der Aufgabe. So kommen beziehen sich vor allem Namen, E-Mails, Herkunftsländer, Institutionen und persönliche Daten auf diesen Schritt. Initial wurde

[1] `https://github.com/marekrei/ml_nlp_paper_data`

dafür die Idee verfolgt, *Named-Entity-Recognition* mittels *spaCy* [**?**] einzusetzen, um all diese Informationen (außer E-Mails, welche leicht durch reguläre Ausdrücke zu erkennen sind) zu entfernen. SpaCy bietet dafür verschiedene vortrainierte Modelle, welche für diese Arbeit mit im Text durch reguläre Ausdrücke erkannten Informationen nachtrainiert wurden. Da dieser Ansatz Entitäten nicht nur zu unverlässlich erkannte (insbesondere keine Phrasen, die über mehrere Tokens reichen), sondern auch zu schlecht klassifzierte wurde ein weiterer Ansatz entwickelt. So wurde ein LSTM-basiertes [**?**], neuronales Netzwerk implementiert, um jede Zeile (im Textdokument, also keine ganzen Sätze) des Vor-Abstract-Texts in die Klassen Autor, E-Mail, Private Informationen, Institution oder Anderes zu unterteilen. Länder werden in einem späteren Schritt entfernt. Zum Training wurden 250 Dokumente zeilenweise mit einer Kommandozeilenanwendung annotiert. Dabei wurden zufällig Dokumente aus dem Korpus gezogen, allerdings wurde sichergestellt, dass jede Konferenz vorkommt. Zudem sind verschiedene Fehlerfälle enthalten, bei denen der Start des Abstracts nicht korrekt erkannt wurde und somit späterer Text enthalten ist. Das Modell erreicht auf weiteren 50 Dokumenten einen Makro-F1-Wert von etwa 96%. Die Architektur verfolgt einen ähnlichen Ansatz wie Gleichung 1.2–1.5, nutzt jedoch nach X_{zd} ein zweites LSTM zur zeilenweisen Klassifikation.

Referenzen

Gutachter

Danksagungen

E-Mails

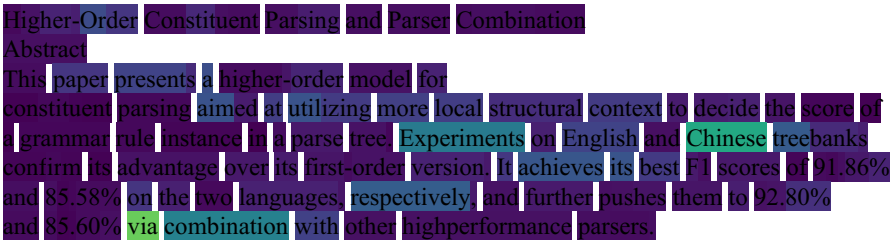
Geolokationen

Fußnoten

Sprache

1.2. Task 2: Learning to Discriminate

- z: Zeilen im Dokument
- t: Tokens pro Zeile (Padding kürzerer Zeilen)
- d: Dimension des Modells (z. B. 150)
- e: Embedding-Dimension
- c: Anzahl Klassen



(a) Bereinigter Text

(b) Originaltext

Abb. 1.1.: Visualisierung der gelernten Gewichtung einzelner Subtokens zur Klassifikation des Herkunftslandes mit bereinigtem und ursprünglichem Text.

$$X_{zte} = \text{Embedding}(X_{zt}) \tag{1.1}$$

$$X_{ztd} = \text{LSTM}_t(X_{zte}) \tag{1.2}$$

$$A_{zt} = \text{softmax}_t\left(\sum_d X_{ztd} W_d^1\right) \tag{1.3}$$

$$X_{zd} = \sum_t X_{ztd} A_{zt} \tag{1.4}$$

$$A_z = \text{softmax}_z\left(\sum_d X_{zd} W_d^2\right) \tag{1.5}$$

$$X_d = \sum_z X_{zd} A_z \tag{1.6}$$

$$Y_c = Y_d W_{dc}^3 \tag{1.7}$$

$$W_{L,C_i} = \left(\frac{|C_{\max}|}{|C_i|}\right)^{\alpha}, \quad \alpha = 0.6 \tag{1.8}$$

		Voraussage										
		AU	CA	CN	FR	DE	IN	IL	JP	SG	CH	UK	USA
Wahrheit	AU	6		8		1						7	6
	CA	1	8	4					2		1	7	32
	CN	1	1	188		1			2	1		1	17
	FR			1	36	3			1			2	4
	DE		2	2	2	47			1			12	11
	IN			1			6					2	11
	IL		1		1			14	2				10
	JP			2	1				45		1	1	6
	SG		1	6						9		2	8
	CH			3		7					4	4	8
	UK	3	5	1	6	1	2		1		2	66	18
	USA	3	13	38	17	6	1	2	8	2	3	32	746

Tab. 1.1.: Konfusionsmatrix der Klassen Australien (AU), Canada (CA), China (CN), Frankreich(FR), Deutschland (DE), Indien (IN), Israel (IL), Japan (JP), Singapur (SG), Schweitz (CH), Großbritannien (UK), Amerika (USA). Leere Zellen implizieren Nullen.

A. Formelles/LaTeX

Erklärung

Hiermit versichere ich die vorliegende Abschlussarbeit selbstständig verfasst zu haben, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben, und die Arbeit bisher oder gleichzeitig keiner anderen Prüfungsbehörde unter Erlangung eines akademischen Grades vorgelegt zu haben.

Würzburg, den 3. August 2020

. . . . . .
Konstantin Herud Thomas Schaffroth Maximilian Meißner