

Praktikumsbericht

# Machine Learning for Natural Language Processing

am Lehrstuhl für Informatik X

Konstantin Herud

Thomas Schaffroth

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



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

# Zusammenfassung

Diese Arbeit untersucht die Fragestellung, ob die Muttersprache von Autoren englischsprachiger, wissenschaftlicher Texte durch linguistische Merkmale ermittelt werden kann. Die Grundlage dafür bildet ein Korpus mit über 18.000 Veröffentlichungen auf Konferenzen des künstlichen Intelligenzbereichs. Hierzu werden zunächst umfangreich direkte und indirekte Informationen, die Nationalitäten preisgeben, wie Namen, Adressen oder Referenzen, entfernt. Anschließend werden zahlreiche Klassifikationsverfahren basierend auf Stylometrie- und Deep-Learning-Methoden untersucht. Die stylometrischen Modelle erreichen über zwölf Muttersprachen hinweg einen Makro-F1-Wert von 44 %. Neben verschiedenen anderen Deep-Learning Ansätzen erreicht ein End-to-End-LSTM-Modell einen Wert von 54 %. Die Idee hinter dessen Architektur ist, globale Vorhersagen als Summe gewichteter, lokaler Entscheidungen zu treffen. Zudem ist es möglich, einzelne Textzeilen zu klassifizieren und die erlernte Gewichtung zu visualisieren, um zu überprüfen, welche Informationen zu Vorhersagen geführt haben.

# Inhaltsverzeichnis

<b>1. Einleitung</b>	<b>4</b>
<b>2. Vorausgehende Arbeit</b>	<b>5</b>
<b>3. Methodik</b>	<b>6</b>
3.1. Task 1: Dataset Preparation . . . . .	6
3.2. Task 2: Learning to Discriminate . . . . .	9
3.2.1. Stylometrische Klassifikation . . . . .	9
3.2.2. End-to-End Klassifikation . . . . .	10
<b>4. Ergebnisse</b>	<b>14</b>
<b>5. Zusammenfassung</b>	<b>16</b>
<b>Literaturverzeichnis</b>	<b>17</b>
<b>A. Wort- und Satzlängenhistogramme</b>	<b>21</b>
<b>B. Visualisierung der gelernten Subwortgewichtung</b>	<b>22</b>
<b>C. PCA-Projektion des internen Modellzustands</b>	<b>23</b>
<b>D. Klassifikation einzelner Zeilen</b>	<b>24</b>

# 1. Einleitung

## Thomas Schaffroth

Durch aufkommende Neuerungen zu Beginn der 2000er Jahre, wie der Herausforderung von *Big Data* sowie der Entfaltung des Internets und der mobilen Kommunikation, wurden seit dem Beginn der *künstlichen Intelligenz* (KI) große Fortschritte in Feldern, wie der Computer Vision, Intelligenten Agenten und Mustererkennung, erzielt. Es sind unter anderem ebendiese Weiterentwicklungen, die den Weg für moderne Herausforderungen, wie Spracherkennung, selbstfahrende Fahrzeuge und *Natural Language Processing* (NLP), geebnet haben [31]. Besonders das Themengebiet des NLP nimmt einen wichtigen Stellenwert in vielen Software-Anwendungen unseres täglichen Lebens ein. Einige der herausragendsten Beispiele sind E-Mail-Plattformen, wie Microsoft Outlook (Spam-Klassifikation, Auto-Complete etc.), sprachbasierte Assistenzsysteme, wie Apple Siri und Amazon Alexa oder Services für maschinelle Übersetzung, wie Google Translate [43].

Die zunehmende Bedeutung von NLP lässt sich neben seiner Rolle in unserem Alltag auch in der Forschung beobachten. Die ACL Anthology (AA) ist ein digitales Repository, dass zehntausende Veröffentlichungen von NLP-Papern aus der Familie der ACL- sowie anderer NLP-Konferenzen beinhaltet. Lag die Anzahl der Veröffentlichungen im Jahr 2000 noch bei 1050, wurden 2018 bereits 4173 publizierte Arbeiten verzeichnet [29].

Ziel dieser Arbeit ist die Bearbeitung der Fragestellung, ob die Erkennung der Muttersprache von Autoren in anderssprachigen, wissenschaftlichen Texten möglich ist. Dieses Problem zu lösen ist deshalb relevant, weil so beispielsweise Autoren besser identifiziert, forensische Analysen unterstützt oder pädagogische Anwendungen stärker personalisiert werden können [8, 10, 32]. Der hierfür zur Verfügung stehende Datensatz beinhaltet 18000 Dokumente verschiedener NLP- und AI-Konferenzen in PDF- und Text-Form.

Das Praktikum gliedert sich in zwei Bestandteile. Zunächst müssen die Dokumente einer grundsätzlichen Datenbereinigung unterzogen werden. Informationen, von denen direkt oder indirekt auf die Herkunft der Autoren geschlossen werden kann, müssen dabei aus den Arbeiten entfernt werden. Beispiele hierfür sind Autorennamen, E-Mail-Adressen, Ländernamen, Institutionen oder Referenzen im Text. Der Grund dafür ist, dass für den zweiten Teil der Arbeit sichergestellt werden muss, dass die Klassifikation der Dokumente aufgrund allgemeiner Texteigenschaften, wie der Syntax und dem verwendeten Vokabular, stattfindet. Informationen, die sich auf die Herkunft des Autors beziehen, würden für eine Verfälschung der Ergebnisse sorgen.

Der folgende Schritt beschäftigt sich mit der Feature-Modellierung, dem Training und der Evaluierung im Kontext verschiedener Techniken der KI und NLP auf dem genannten Korpus in Bezug auf die Klassifikationsaufgabe. Ziel soll sein, einen Klassifikator zu trainieren, der für eine möglichst große Zahl der Dokumente das Herkunftsland des jeweiligen Autors korrekt vorhersagt.

## 2. Vorausgehende Arbeit

Die Aufgabe der Muttersprachenerkennung basierend auf Texten einer zweiten Sprache ist in der Literatur als *Native Language Identification* (NLI) bekannt. Die Association for Computational Linguistics (ACL) hat dafür einen eigenen Wettbewerb zur Entwicklung des besten Systems [40, 24], welcher das Interesse an diesem Forschungsgebiet mit seiner Einführung massiv steigerte. Generell wird NLI als überwachtes Trainingsproblem behandelt. Als eine der ersten extrahieren Koppel et al. 2005 eine Vielzahl an linguistischen Merkmalen, wie Zeichen und Part-Of-Speech-N-Gramme, Funktionswörter und insbesondere Rechtschreib- und Grammatikfehler. Damit trainieren sie eine *Support-Vektor-Maschine* (SVM) und erreichen auf dem ICLE-Korpus mit fünf Muttersprachen etwa 80 % Genauigkeit [20].

Spätere Arbeiten versuchen syntaktische Merkmale miteinzubeziehen. Wong und Dras nutzen 2011 Parsebäume und Parsereranking, um von syntaktischen Fehlern auf die Muttersprache zu schließen, und erreichen damit über sieben Muttersprachen 80 % Genauigkeit [46]. Tetreault et al. sowie Swanson und Charniak untersuchen 2012 jeweils Tree-Substitution-Grammars und erreichen damit auf sieben Muttersprachen ebenfalls etwa 80 % Genauigkeit [41, 38]. Jarvis et al. gewannen 2013 den ersten NLI-Workshop der ACL mit einer SVM und einer Vielzahl linguistischer Merkmale, wobei sie fast 84 % Genauigkeit erreichen [16].

Bis zum nächsten ACL-Workshop 2017 entstanden zahlreiche weitere Ansätze [24]. Dabei zeigten sich verschiedene Erkenntnisse. Zum einen setzten sich Deep-Learning Ansätze nicht durch. Stattdessen ging die Forschung weiter in Richtung extrahierter Merkmale und Klassifikation mittels Support-Vektor-Maschinen und anderen linearen Modellen. Zum anderen zeigte sich, dass Systeme mit mehreren Klassifikatoren gut funktionieren. Der beste Ansatz von Cimino und Dell’Orletta 2017 nutzt beispielsweise lineare Regression zur Klassifikation auf Satzebene und anschließend eine SVM, die mit diesen Ergebnissen eine Vorhersage auf Dokumentenebene trifft, wodurch sie eine Genauigkeit von etwa 88,2 % auf 11 Muttersprachen und 1000 Dokumenten pro Sprache erreichen [5]. Ein weiterer Ansatz von Kulmizev et al. nutzt eine lineare SVM mit Zeichen-N-Grammen und erreicht dabei etwa 87,6 % Genauigkeit [21]. Dabei weisen sie darauf hin, dass bei einer Aufgabe wie NLI Zeichen-N-Gramme besonders geeignet sind, um Muster in Rechtschreibfehlern zu erkennen, und dabei wesentlich dünner besetzt sind, als Wort-N-Gramme. Ein Deep-Learning-Ansatz mit LSTM-Netzwerken von Bjerva et al. erreicht hingegen nur 83 % Genauigkeit [2].

## 3. Methodik

### Konstantin Herud

Zunächst wird das Vorgehen bei der Datenbereinigung erläutert und anschließend die Implementierung verschiedener Verfahren zur Bestimmung der Autornationalitäten.

### 3.1. Task 1: Dataset Preparation

Da verschiedene Informationen die Nationalität 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. Jeder folgende Paragraph entspricht einer der Stufen.

**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 <sup>1</sup>. Da verschiedene spätere Schritte außerdem Informationen über die Zeichenposition des Abstract-Anfangs 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, um zu frühe Treffer des regulären Ausdrucks zu überspringen. 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

---

<sup>1</sup>[https://github.com/marekrei/ml\\_nlp\\_paper\\_data](https://github.com/marekrei/ml_nlp_paper_data)

Herausforderungen der Aufgabe. Insbesondere Namen, E-Mails, Herkunftsländer, Institutionen und persönliche Daten beziehen sich auf diesen Schritt. Initial wurde dafür die Idee verfolgt, *Named-Entity-Recognition* mittels *spaCy* [14] 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 klassifizierte wurde ein weiterer Ansatz entwickelt. So wurde ein *LSTM*-basiertes [13], neuronales Netzwerk implementiert, um jede Zeile des Dokumentkopfs in der Textdatei 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 3.3–3.6, nutzt jedoch nach  $X_{zd}$  ein zweites LSTM zur zeilenweisen Klassifikation.

**Referenzen** Referenzen unterteilen sich in zwei Typen: Verweise innerhalb des Texts und ausführlich Referenzblöcke. Erstere sind aufgrund ihrer einheitlichen Struktur leicht mittels regulären Ausdrücken zu ermitteln. Dafür wurden zahlreiche zufällige Dokumente betrachtet und drei Strukturtypen identifiziert, für die eigene reguläre Ausdrücke entwickelt wurden: Klammer-Ausdrücke, die Jahreszahlen im 20. und 21. Jahrhundert enthalten, Phrasen mit “et. al.” und zuletzt Aufzählungen (z. B. verbunden durch “und”, “,” oder “&” etc.), gefolgt von Jahreszahlen.

Referenzblöcke zu identifizieren stellt sich jedoch als zweitschwerste Aufgabe heraus. Ein initialer Ansatz sah vor, die zuvor im Text gefundenen Referenzen zusammen mit verschiedenen Schlüsselwörtern dazu zu nutzen, Zeilen-Blöcke als Referenzen zu erkennen. Da durch die Konvertierung von zweiseitigen, mehrseitigen Dokumenten in eindimensionalen Text jedoch häufig Strukturfehler auftreten, sind Teile der Referenzblöcke oftmals stark über die Dokumente verteilt, weshalb dieser initiale Ansatz zu keinen befriedigenden Ergebnissen führte. Deshalb wurde ebenfalls ein neuronales Netz, mit einer Architektur äquivalent zur vorherigen, trainiert, um zeilenweise zu annotieren, ob es sich um Referenzen handelt. Dafür wurden etwa 50 Dokumente manuell annotiert, wobei neben zufällig gewählten Werken insbesondere darauf geachtet wurde, Fehlerfälle zu verwenden. Das Netz verarbeitet den Text iterativ in Stücken von 80 Zeilen (um Padding und Grafikspeicher-Anforderungen möglichst gering zu halten, wurde kein kompletter *End-to-End* Ansatz gewählt). Um dennoch von Informationen über die Position dieser 80 Zeilen im Dokument zu profitieren, wird auf die  $d$ -dimensionalen  $(1, \dots, i, \dots, d)$ , internen

Zustände des Netzwerks ein Positionssignal  $PE$  addiert (vgl. [44], Gleichung 3.1 & 3.2).

$$PE(zeile, 2i) = \sin\left(\frac{zeile}{10000^{2i/d}}\right) \quad (3.1)$$

$$PE(zeile, 2i + 1) = \cos\left(\frac{zeile}{10000^{2i/d}}\right) \quad (3.2)$$

Das Modell erreicht auf 20% separaten, zufälligen Daten einen Makro-F1-Wert von beinahe 99,7%. Sowohl das Dokumentkopf- als auch das Referenzblocknetz sind mit ihren Gewichten aufgrund der hoch-rekurrenten Struktur unter einem Megabyte groß.

**Gutachter** Generell wurden wenige Gutachter innerhalb der Dokumente festgestellt. Diese geringe Menge folgt jedoch einem festen Schema, welches mittels eines regulären Ausdrucks ermittelt wird.

**Danksagungen** Danksagungen sind meist wenige Zeilen in geschlossenen Blöcken und somit leichter zu ermitteln als Referenzblöcke. Vorerst stand im Raum, ebenfalls ein neuronales Netz einzusetzen, allerdings genügen reguläre Ausdrücke. So wird nach dem letzten Vorkommen der gängigen Überschrift “Acknowledgment” gesucht (wobei unterschiedliche Schreibweisen zu beachten sind, i. e. Acknowledge?ments?). Das Ende der Danksagung wird durch eine Liste mit Schlüsselwörtern, kombiniert mit den zuvor gefunden Referenzen sowie einem regulären Ausdruck für Referenzen (da Danksagungen diesen häufig vorausgehen), ermittelt. Kann dennoch kein Ende gefunden werden, wird der Text bis zum Ende des Paragraphen gewählt.

**E-Mails** Der Großteil aller E-Mails wurde in diesem Schritt bereits durch die Dokumentkopfstufe entfernt (insbesondere schwer identifizierbare Fälle, deren gezielt irreführende Schreibweise z. B. Spam verhindern soll). Der Rest aller E-Mails wird in diesem Schritt mit zwei weiteren regulären Ausdrücken entfernt.

**Herkunftsländer** Länder und Nationalitäten werden mit dem Python-Modul *geotext*<sup>2</sup> ermittelt. Das Modul arbeitet mit regulären Ausdrücken und der geographischen Datenbank *GeoNames*, die jegliche Länder und über elf Millionen Ortsnamen enthält.

**Fußnoten** Fußnoten sind unscharf definiert und aufgrund der fehlenden visuellen Struktur sehr schwer im rohen Text zu identifizieren. Um diese Informationen dennoch zu entfernen, wird der Text zunächst mittels spaCy in Sätze unterteilt. Anschließend werden alle Sätze mit einer Liste von Schlüsselwörtern, wie “sponsored” oder “internship”, sowie bereits ermittelten Autor- und Organisationsnamen untersucht. Wird ein Schlüsselwort gefunden, wird der entsprechende Satz entfernt.

---

<sup>2</sup><https://github.com/elyase/geotext>



**Sprache** Zuletzt wird nicht-englische Sprache entfernt. Diese Aufgabe stellt ebenfalls eine Herausforderung dar, da viele Rohtext-Fragmente aufgrund kryptischer Formeln oder anderen Artefakten von gängigen Sprachidentifikationssystemen fälschlicherweise als nicht-englisch erkannt werden. Dennoch wird die spaCy-Erweiterung *CLD* eingesetzt, welche die *Compact Language Detection 2* Bibliothek<sup>3</sup> anbindet. Diese ermöglicht die Klassifikation von rohem Text in über 80 verschiedene Sprachen mittels Zeichen-N-Grammen und einem Naiven-Bayes-Modell. Dieser Ansatz führt zu einer akzeptablen Sensitivität, allerdings zu einer schlechten Genauigkeit. Um letztere zu verbessern, wird lediglich eine geringe Teilmenge relevanter Sprachen der 80+ möglichen berücksichtigt.

Die einzelnen Stufen der Pipeline lassen sich nicht nur jeweils über verschiedene Prozesse verteilen und vervielfältigen, die Pipeline lässt sich auch als Ganzes parallelisieren. Zusätzlich ermöglicht die Anwendung Hardware-Beschleunigung für die neuronalen Netzwerke. Das Wissen über zu entfernende Informationen wird akkumuliert und letztlich in eine CSV-Datei geschrieben, außerdem wird der inkrementell bereinigte sowie der ursprüngliche Text zwischen Stufen weitergegeben und abschließend in eine Textdatei geschrieben.

## 3.2. Task 2: Learning to Discriminate

Um die bereinigten Dokumente hinsichtlich der Muttersprache ihrer Autoren zu klassifizieren, wurden zwei Ansätze verfolgt: Zum einen Klassifikation mittels verschiedenen *stylometrischen* Merkmalen, zum anderen ein *End-to-End-Deep-Learning*-Modell.

### 3.2.1. Stylometrische Klassifikation

Stylometrie befasst sich mit der Untersuchung von Sprachstil in der Regel zur Bestimmung des Autors [35]. Diesem Abschnitt liegt die Annahme zugrunde, dass sich die selben Mittel auch auf die Bestimmung des Herkunftslands übertragen lassen. Statistische und rechnerische Zuweisung der Urheberschaft haben generell eine lange Geschichte, die bis ins 19. Jahrhundert reicht [37]. Dementsprechend existieren viele Methoden, die von einfachen Ideen, wie Mendenhalls Kurven zu Wortlängen [27], über Kilgariffs  $\chi^2$ -Tests zu Vokabularähnlichkeit [17] bis zu fortgeschritteneren Ideen, wie Burrows Delta Statistik [9], reicht. In dieser Arbeit wurden verschiedene linguistische Merkmale extrahiert.

- **Wortschatz:** Für verschiedene Maße zur lexikalischen Diversität einzelner Dokumente wurde das Python-Modul *lexicalrichness* eingesetzt<sup>4</sup>. Insgesamt wurden die Maße Type-Token-Ration (TTR) [39], Guirauds Index [11], Corrected-TTR [4], Herdans Maß [12], Somers Maß [34], Dugasts Index [7], Maas Index [23], Mean-Segmental-TTR [42], Moving-Average-TTR [6], Measure of Textual Lexical Diversity [26] und die Hypergeometrische-Verteilungs-Diversität [25] ermittelt.

---

<sup>3</sup><https://github.com/CLD2Owners/cld2>

<sup>4</sup><https://github.com/LSYS/lexicalrichness>

- **Wort- und Satzlängen:** Für Wort- und Satzlängen werden Histogramme der Dokumente gebildet, die mittels der Anzahl an Wörtern beziehungsweise Sätzen insgesamt normalisiert wird. Wortlängen werden in Zeichen gemessen und auf 50 beschränkt, Satzlängen mittels ihrer Anzahl an Wörtern bis 100 (Anhang A zeigt solche Histogramme für alle Dokumente kombiniert).
- **Interpunktion:** Häufigkeiten von genutzten Interpunktionszeichen (jegliche in Python durch *string.punctuation* definierte Zeichen), normalisiert über die gesamte Zeichenanzahl im Dokument.
- **Funktionswörter:** Häufigkeiten von insbesondere Stopwörtern und weiteren Kandidaten (insgesamt 277 Funktionswörter <sup>5</sup>).
- **N-Gramme:** Den wichtigsten Teil bilden die N-Gramm-Histogramme. Dafür werden jeweils die 200 häufigsten 2- und 3-Gramme über die Lemmata der Wörter, über OntoNotes5-POS-Tags und über universelle POS-Tags mittels spaCy ermittelt (später in dieser Reihenfolge mit *Lemmata*, *Tag* und *POS* bezeichnet) [45].

Für jedes Dokument werden alle Merkmale extrahiert, über den gesamten Korpus standardisiert und anschließend mit einem einfachen neuronalen Netzwerk klassifiziert. Das Netzwerk besteht aus einer beliebigen Anzahl Blöcken, die jeweils aus einer verdeckten Schicht, Batch-Normalisierung [15], SELU-Aktivierung [19] und anschließend Dropout [36] bestehen. Aufgrund der einfachen und schnell zu trainierenden Architektur werden die Hyperparameter (wie Anzahl Blöcke, Anzahl Neuronen, Dropout etc.) mit einer umfangreichen Zufallssuche ermittelt. Neben einem Netzwerk mit der Kombination aller Merkmale existieren separate Netzwerke für alle Merkmalstypen alleinstehend, mit der Intention diese vortrainierten Subnetzwerke später mit dem *End-To-End*-Modell zu kombinieren.

### 3.2.2. End-to-End Klassifikation

Als *End-to-End* werden Modelle bezeichnet, die keine vorherigen oder manuellen Schritte, wie Merkmalsextraktion, benötigen, sondern ausschließlich anhand der rohen, bereinigten Texte klassifizieren. Dafür ist zunächst eine Methode zur Textrepräsentation notwendig. Gängige frühe Verfahren zur Repräsentation ganzer Wörter sind latente *Word2Vec*- oder *GloVe*-Vektoren [28, 30]. Diese sind jedoch aus zahlreichen Gründen problematisch, insbesondere werden die Wörter dabei in der Regel auf Lemmata, Wortstamm oder Ähnliches normalisiert, um die Anzahl benötigter Vektoren zu reduzieren. Zudem existieren für falsche Schreibweisen, seltene oder zusammengesetzte Wörter und andere Eigenheiten keine individuellen Vektoren. All diese Informationen spielen aber potentiell eine wichtige Rolle zur Klassifizierung. *FastText* löst diese Probleme teilweise, indem Wörter als Summen von Zeichen-n-Gramm-Vektoren dargestellt werden und somit auch unbekannte Sequenzen repräsentiert werden können [3]. Der gegenwärtige

---

<sup>5</sup><https://semanticssimilarity.wordpress.com/function-word-lists>

Standard und auch in dieser Arbeit gewählte Ansatz ist jedoch eine *Byte Pair Encoding*-Subwortdarstellung [33]. Hierbei werden ausgehend von einer Repräsentation auf Zeichenebene inkrementell die häufigsten Symbolfolgen zu einzelnen Subwörtern kombiniert, wodurch letztlich häufige Wörter als eigene Vokabeln existieren, unbekannte Wörter jedoch auch durch Subwortsequenzen ausgedrückt werden können. Dies ermöglicht trotz geringer Vokabulargröße Texte kompakt zu repräsentieren. Aufgrund der eigenen wissenschaftlichen Semantik der Konferenzen und der spezifischen Syntax der Textdateien mit ihren Strukturfehlern und Artefakten, wird keine vortrainiertes Embedding verwendet, sondern während dem Trainingsprozess von Grund auf erlernt. Zudem wurde probiert, das Embedding eigenständig vorzutrainieren, indem das spätere, anschließend weiterverwendete LSTM (Gleichung 3.3 & 3.4) mittels *Masked Language Modelling* trainiert wird. Die Idee dabei ist, einzelne Wörter zu maskieren, um anschließend mit dem unmaskierten Kontext vorherzusagen, welche Wörter an den jeweiligen Stellen standen, wodurch das Modell Syntax und teilweise Semantik erlernt und die Embeddings entsprechend anpasst. Dabei zeigte sich zwar keine Verbesserung der späteren Ergebnisse, allerdings eine erhebliche Beschleunigung des Trainingsfortschritts, was eine effiziente Hyperparametersuche erlaubt.

Um die zentrale, für diese Arbeit entwickelte End-to-End-Netzarchitektur zu beschreiben werden zunächst einige Indizes definiert.

- $z$ : Zeilen in der Textdatei des Dokuments (dynamisch)
- $w$ : (Sub-)Worte pro Zeile (dynamisch, Padding kürzerer Zeilen)
- $d$ : Dimension des Modells (150)
- $e$ : Embedding-Dimension (100)
- $c$ : Anzahl Klassen (12)

Die Netzarchitektur lässt sich dann durch Gleichung 3.3 – 3.9 beschreiben (die Batchdimension wird hierbei ausgelassen).

$$X_{zwe} = \text{Embedding}(X_{zw}) \quad (3.3)$$

$$X_{zwd} = \text{LSTM}_w(X_{zwe}) \quad (3.4)$$

$$A_{zw} = \text{softmax}_w \left( \sum_d X_{zwd} W_d^1 \right) \quad (3.5)$$

$$X_{zd} = \sum_w X_{zwd} A_{zw} \quad (3.6)$$

$$A_z = \text{softmax}_z \left( \sum_d X_{zd} W_d^2 \right) \quad (3.7)$$

$$X_d = \sum_z X_{zd} A_z \quad (3.8)$$

$$Y_c = X_d W_{dc}^3 \quad (3.9)$$

So läuft zunächst eine LSTM-Schicht parallel und individuell über die Embedding-Vektoren jeder Zeile im Dokument. Für diesen Schritt wird die Zeilen- als Batchaxe

behandelt (d. h. beide Dimensionen werden zusammengeführt), wodurch alle Zeilen in einem einzigen Vorwärtsschritt des Netzwerks verarbeitet werden können, was die Architektur extrem schnell macht. Das Netzwerk lernt anschließend eine Gewichtung  $W_d^1$  (in Form eines einzelnen Neurons oder Subnetzwerks mit einem Ausgabeneuron, daher eindimensional), wie wichtig jeder verdeckte Zustand der LSTM-Schicht  $X_{zwd}$  zu jedem Subwort ist. Dabei werden die mit der *Softmax*-Funktion normalisierten Anteile für jedes Subwort  $A_{zw}$  innerhalb einer Zeile bestimmt, um damit die LSTM-Zustände zu gewichten und anschließend zu einem Vektor pro Zeile  $X_{zd}$  aufzuaddieren. Das Netzwerk lernt danach eine zweite, äquivalente Gewichtung  $W_d^2$ , um damit alle Zeilen gemäß ihrer Relevanz zu gewichten und zu einem finalen Vektor  $X_d$ , der das gesamte Dokument repräsentiert (siehe Anhang C), zu summieren. Eine letzte lineare Schicht ergibt dann die Klassen-Logits  $Y_c$ . Eine zweite LSTM-Schicht zwischen Gleichung 3.6 und 3.7, um potentiell Abhängigkeiten zwischen Zeilen besser zu erfassen, führte empirisch zu keiner Verbesserung. Dokumente satz- statt zeilenweise zu verarbeiten ergibt vergleichbare Ergebnisse. Durch zufällige Suche wurde  $e = 100$  und  $d = 150$  ermittelt, das Vokabular besteht aus 12000 Subwörtern.

Um extremes Padding zu vermeiden, wird mit einer Batch-Größe von eins trainiert (andernfalls müssten alle Daten im Batch auf die Anzahl des Dokuments mit den meisten Zeilen angepasst werden). Die Subwortaxe wird auf die längste Zeile im Dokument, höchstens jedoch auf 50 Subwörter mittels Padding angepasst (was 99,9% aller Zeilen einschließt). Um sowohl Speicher- als auch Zeitkomplexität zu reduzieren, wird *Automatic Mixed Precision*<sup>6</sup>, also auf Float16 statt Float32 basierende Parameter, eingesetzt. Zudem wurde mit Normalisierung und Dropout nach Gleichung 3.6 und 3.8 experimentiert, jedoch ohne resultierende Verbesserungen.

Da sich die Häufigkeiten der Dokumente jeder eindeutigen Nationalität extrem unterscheiden (Tabelle 3.1), wurden verschiedene Maßnahmen für das Training untersucht. Sowohl Auslassen von Datensätzen größerer Klassen (*Downsampling*) als auch Vervielfachen von Dokumenten seltener Länder (*Upsampling*) zeigten keinen Erfolg. Erstes verschlechterte die Ergebnisse maßgeblich, letzteres führte hauptsächlich zu einer längeren Trainingsdauer. Stattdessen wurde die Loss-Funktion (Kreuzentropie) gewichtet (Gleichung 3.10). Der Wert von  $\alpha$  wurde dabei durch Ausprobieren aller Werte von 0.0–1.0 in Zehntelschritten ermittelt.

Land	USA	CN	UK	DE	JP	CA	FR	IL	AU	CH	SG	IN
Häufigkeit	8699	2119	1042	770	557	542	470	278	277	256	254	200

**Tab. 3.1.:** Häufigkeiten der eindeutigen Herkunftsländer im Korpus

fachen von Dokumenten seltener Länder (*Upsampling*) zeigten keinen Erfolg. Erstes verschlechterte die Ergebnisse maßgeblich, letzteres führte hauptsächlich zu einer längeren Trainingsdauer. Stattdessen wurde die Loss-Funktion (Kreuzentropie) gewichtet (Gleichung 3.10). Der Wert von  $\alpha$  wurde dabei durch Ausprobieren aller Werte von 0.0–1.0 in Zehntelschritten ermittelt.

$$W_{C_i}^L = \left( \frac{|C_{\max}|}{|C_i|} \right)^\alpha, \quad \alpha = 0.6 \quad (3.10)$$

<sup>6</sup><https://github.com/NVIDIA/apex>

Außerdem wurden verschiedene andere Architekturen ausprobiert. Generell sind Transformer der vorherrschende Architekturstandard. Da deren Komplexität jedoch quadratisch mit der Anzahl an Eingabewörtern skaliert und die Parameteranzahl von vortrainierten Modellen meist in den Gigabyte-Bereich reicht, eignen sich entsprechende Modelle nur bedingt. Zudem ist der Domänentransfer problematisch, da sich die bereinigten Textdateien mit oftmals fehlenden (oder schlichtweg nicht vorhandenen, z.B. Titel, Überschriften etc.) Satzgrenzen und zahlreichen Artefakten stark von den strukturierten Daten, auf denen die Transformer vortrainiert wurden, unterscheiden.

- **CNN:** Hierbei wurden Gleichung 3.4–3.6 durch ein Text-CNN ersetzt [18]. Dabei wurden Faltungen mit Breite zwei bis sechs und jeweils 20 Filtern verwendet. Anstatt den Text eindimensional in das CNN zu geben, wurde die zweidimensionale Form  $z \times w$  beibehalten, was die Zeitkomplexität dank Parallelisierung von  $O(zw)$  maßgeblich auf  $O(w)$  reduziert. Hierbei werden N-Gramme über Zeilen hinweg nicht beachtet, weshalb statt Zeilen  $z$  ganze Sätze  $s$  verwendet wurden, welche vorab mittels spaCy ermittelt wurden.
- **RoBERTa:** Dieses 125 Millionen Parameter starke, vortrainierte Sprachmodell basiert auf der Transformer-Architektur und kann Sequenzlängen bis 512 verarbeiten [22]. Der Text wird deshalb in entsprechende Stücke aufgeteilt, welche separat mit dem Modell verarbeitet werden. Anschließend läuft eine LSTM-Schicht über die gepoolte Ausgabe dieser Stücke, um das Dokument letztlich zu klassifizieren.
- **Longformer:** Dieses ebenfalls vortrainierte Modell verbessert die Transformer-Architektur, sodass Speicher- und Zeitkomplexität linear mit der Sequenzlänge skalieren [1]. Dadurch wird es möglich, gleichzeitig 4096 Subwörter zu verarbeiten, was durchschnittlich 53% einzelner Dokumente abdeckt. So werden die ersten 4096 der Subtokens mit dem Modell und einer einfachen linearen Schicht klassifiziert.

## 4. Ergebnisse

Die Aufgabe der Datenbereinigung wurde von vier Personen mit jeweils 25 Dokumenten ausgewertet, wobei die entwickelte Pipeline einen Jaccard-Index von etwa 0,847 zu den manuellen Annotationen erreicht.

Tabelle 4.1 zeigt die Ergebnisse der Klassifikation. Aufgrund des Klassenungleichgewichts wurde der Makro-F1 Wert als Metrik verwendet. Hierbei existieren zwei Baselines *Häufigste Klasse* und *Zufällig*, wobei für alle Dokumente das häufigste beziehungsweise ein zufälliges Herkunftsland bestimmt wird.

Modell	AU	CA	CN	FR	DE	IN	IL	JP	SG	CH	UK	USA	Makro
Häufigste Klasse												72,0	6,0
Zufällig	3,1	6,6	9,1	4,8	5,3	1,8	2,9	5,6	2,3	3,2	7,2	13,9	5,5
Interpunktion			10,4									72,0	8,0
Wortschatz								3,5				71,9	8,2
Wort-&Satzlängen			34,3		2,5							71,7	9,5
Tag	6,2		70,2	45,5	37,6	16,0		58,5			17,4	77,3	30,7
POS	6,7		73,7	52,2	35,5	18,2	35,9	53,0			20,3	78,1	31,4
Funktionswörter			68,2	23,7	14,0	<b>46,2</b>	43,8	59,6	6,5		4,9	75,8	34,1
Lemmata		3,4	74,2	56,1	47,7	27,6	22,7	60,2	17,6		17,5	79,5	36,9
Auswahl			75,5	<b>68,1</b>	54,0	33,3	49,0	70,3	13,8	6,9	20,4	80,5	40,9
Kombination	10,8	3,5	74,4	64,2	55,6	32,3	44,4	68,8	18,6	7,4	29,6	79,4	44,2
Longformer												72,0	6,0
RoBERTa												72,0	6,0
CNN		6,1	80,0	49,5	40,7	16,2	15,8	57,1			48,6	83,2	33,1
LSTM\Gewichtung	25,0	12,3	78,7	65,9	52,9	12,9	41,7	68,4	19,5	10,0	54,3	83,5	43,8
LSTM	<b>28,6</b>	<b>18,6</b>	<b>80,7</b>	65,5	<b>65,7</b>	41,4	<b>63,6</b>	<b>76,3</b>	<b>47,4</b>	<b>22,2</b>	<b>54,8</b>	<b>85,4</b>	<b>54,2</b>

**Tab. 4.1.:** Auswertung aller Modelle mit F1-Werten jeder Klasse sowie zuletzt Makro-F1 (in Prozent). Zuerst zwei Baselines, dann alle stylometrischen Komponenten und zuletzt Deep-Learning Modelle. *Auswahl* ist ein Netzwerk aus *Tag*, *POS*, *Funktionswörter* sowie *Lemmata*. *Kombination* verbindet alle stylometrischen Komponenten. *LSTM* wurde sowohl mit Gewichtung als auch ohne durch reines Summieren getestet.

Tabelle 4.2 zeigt die Konfusionsmatrix der LSTM-Architektur. Viele Fehler sind hierbei nachvollziehbar, beispielsweise die Klassifizierung von Dokumenten der Schweiz als französisch oder deutsch sowie Dokumente aus Singapur als chinesisch.

	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	32	746

**Tab. 4.2.:** Konfusionsmatrix des LSTM-Modells über die Klassen Australien (AU), Canada (CA), China (CN), Frankreich (FR), Deutschland (DE), Indien (IN), Israel (IL), Japan (JP), Singapur (SG), Schweiz (CH), Großbritannien (UK), Amerika (USA).

Die entwickelte Architektur besticht nicht nur durch ihre Rekurrenz sowie Effizienz, sie ermöglicht auch ein gewisses Maß an Einblick in das Modell. Wie Anhang B zeigt, kann beispielsweise die gelernte, min-max-normalisierte Gewichtung jedes Subworts  $A'_{zw} = A_{zw}A_z$  visualisiert werden (hierbei ist zu beachten, dass die Gewichtung auf rekurrenten Zuständen beruht und nicht auf Subwörtern selbst). Wirkliche Erkenntnisse ergeben sich dadurch nicht, allerdings lässt sich sicherstellen, dass das Modell keinen falschen Hinweisen, wie übersehenen, zu entfernenden Informationen, folgt. Lernt das Modell mit den unzensierten Texten, zeigt sich, dass hauptsächlich E-Mail-Domänen berücksichtigt werden, welche vermutlich selten in anderen Kontexten vorkommen und das Herkunftsland somit am zuverlässigsten preisgeben.

Zudem ermöglicht die Architektur einen weiteren Weg Einblick in das Modell zu gewinnen. Der finale Dokumentvektor  $X_d$  ist lediglich eine gewichtete Summe der Zeilenvektoren  $X_{zd}$ . Deshalb können aus allen Dokumentvektoren  $X_d$  geometrische Schwerpunkte  $Z_i$  jeder Klasse  $i$  berechnet werden (nachvollziehbar in Anhang C), um anschließend für jede Zeile  $j$  mit der Kosinus-Ähnlichkeit (Gleichung 4.1) zu bestimmen, welche Nationalität am wahrscheinlichsten ist.

$$\text{sim}(X_{zd,j}, Z_i) = \frac{X_{zd,j} \cdot Z_i}{\|X_{zd,j}\|_2 \|Z_i\|_2} \quad (4.1)$$

Anhang D zeigt dies beispielhaft.

## 5. Zusammenfassung

### **Konstantin Herud und Thomas Schaffroth**

Sprache ist vielfältig. So haben unterschiedliche Länder verschiedene Eigenheiten, die sich beim Schreiben und Sprechen einer Fremdsprache in zu wörtlichen Übersetzungen, spezifischen Fehlern oder ungewöhnlichen Sprachkonstrukten äußern. Das führt zu der Frage, ob sich diese Eigenschaften nutzen lassen, um die Muttersprache eines Autors zu identifizieren.

Dafür wurden in dieser Arbeit über 18.000 englischsprachige, wissenschaftliche Dokumente – also mit grundsätzlich hoher fremdsprachlicher Kompetenz – im Themengebiet der künstlichen Intelligenz untersucht. Um sicherzustellen, dass keine nicht-linguistischen Informationen die Herkunft der Autoren direkt oder indirekt preisgeben, werden die Dokumenttexte einer aufwändigen Bereinigung unterzogen. So werden in einer mehrstufigen Pipeline die Namen der Autoren, E-Mails, Institutionen und Firmen, Herkunftsländer, Förderungen und Danksagungen, persönliche Daten und Referenzen entfernt. Die Pipeline erreicht dabei vor allem mit neuronalen Netzen und regulären Ausdrücken einen Jaccard-Index von etwa 85 % zu vier menschlichen Prüfern.

Anschließend wurden verschiedene Klassifikationsverfahren vorgestellt. Zunächst wurden verschiedene stylometrische Methoden untersucht. So haben insbesondere Histogramme von N-Grammen über verschiedene POS-Tags, Wortlemmata und Funktionswörtern Erfolg mit einem Makro-F1 Wert von jeweils 30–37 % (was einer Genauigkeit von etwa 94 % entspricht) gezeigt. Verschiedene Maße zur lexikalischen Diversität, Wort- und Satzlängen sowie Interpunktion waren mit einem Makro-F1-Wert von unter 10 % weniger erfolgreich. Alle stylometrischen Merkmale kombiniert erreichten über 44 %.

Anschließend wurden verschiedene Deep-Learning-Ansätze untersucht. CNN- und Transformer-Architekturen konnten sich dabei nicht durchsetzen, allerdings wurde eine ergebnisreiche LSTM-Architektur mit einem Makro-F1-Wert von 54 % entwickelt. Die Idee dieser Architektur ist globale Vorhersagen als Summe gewichteter, lokaler Entscheidungen auf Zeilenebene zu treffen. Zum einen können diese globalen Entscheidungen nicht nur auf lokale Ebene zurückgeführt werden, wodurch sich einzelne Textzeilen klassifizieren lassen, zum anderen kann auch die erlernte Gewichtung visualisiert werden, um zu überprüfen, welche Informationen zu Klassifikationsergebnissen führen.

Die Frage, ob sich die Muttersprache von Autoren basierend auf Texten einer zweiten Sprache bestimmen lässt, konnte so weitestgehend beantwortet werden.

Zukünftig könnte der Fokus stärker auf vortrainierte Transformer-Architekturen gesetzt werden, welche oftmals auf mehr als einer Sprache trainiert werden und somit nicht nur englische Dokumente klassifizieren könnten, sondern dank ihres Parameterreichtums und des rapiden Forschungsfortschritts auf einer Vielzahl anderssprachiger Texte noch wesentlich bessere Ergebnisse erzielen könnten.



# Literaturverzeichnis

- [1] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer, 2020.
- [2] J. Bjerva, G. Grigonyte, R. Östling, and B. Plank. Neural networks and spelling features for native language identification. pages 235–239, 01 2017.
- [3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information, 2016.
- [4] J. B. Carroll. *Language And Thought*. Prentice-Hall, 1964.
- [5] A. Cimino and F. Dell’Orletta. Stacked sentence-document classifier approach for improving native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 430–437, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.
- [6] M. A. Covington and J. D. McFall. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100, 2010.
- [7] D. Dugast. Sur quoi se fonde la notion d’étendue théoretique du vocabulaire? In *Le français moderne*, volume 46, pages 25–32, 1978.
- [8] D. Estival, T. Gaustad, S. Pham, W. Radford, and B. Hutchinson. Author profiling for english emails. 12 2007.
- [9] B. F. ’delta’: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17:267–287, 09 2002.
- [10] J. Gibbons. *Forensic Linguistics: An Introduction to Language in the Justice System*. 01 2003.
- [11] P. Guiraud. *Les caractères statistiques du vocabulaire: essai de méthodologie*. Presses universitaires de France, 1954.
- [12] G. Herdan. *Quantitative Linguistics*. Butterworths, 1964.
- [13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.
- [14] M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

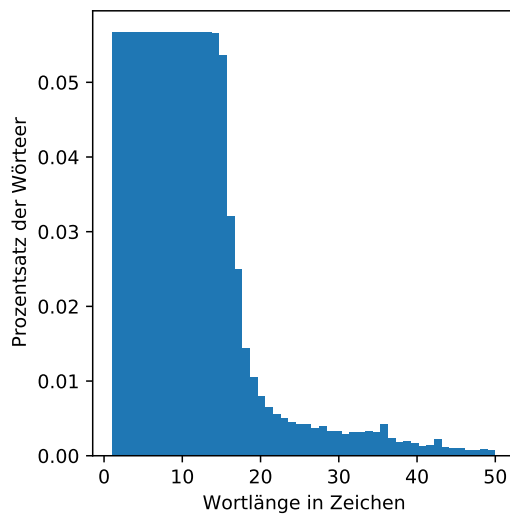
- [15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.
- [16] S. Jarvis, Y. Bestgen, and S. Pepper. Maximizing classification accuracy in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118, Atlanta, Georgia, June 2013. Association for Computational Linguistics.
- [17] A. Kilgarrieff. Comparing corpora. *International Journal of Corpus Linguistics*, 6, 11 2001.
- [18] Y. Kim. Convolutional neural networks for sentence classification, 2014.
- [19] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks, 2017.
- [20] M. Koppel, J. Schler, and K. Zigdon. Determining an author’s native language by mining a text for errors. pages 624–628, 01 2005.
- [21] A. Kulmizev, B. Blankers, J. Bjerva, M. Nissim, G. van Noord, B. Plank, and M. Wieling. The power of character n-grams in native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 382–389, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.
- [22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
- [23] H. D. Maas. Zusammenhang zwischen wortschatzumfang und länge eines textes. In *Zeitschrift für Literaturwissenschaft und Linguistik*, volume 8, pages 73–79, 1972.
- [24] S. Malmasi, K. Evanini, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, and Y. Qian. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.
- [25] P. M. McCarthy and S. Jarvis. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488, 2007.
- [26] P. M. McCarthy and S. Jarvis. Mtd, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392, May 2010.
- [27] T. C. Mendenhall. The characteristic curves of composition. *Science*, 9(214):237–249, 1887.

- [28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.
- [29] S. M. Mohammad. The state of nlp literature: A diachronic analysis of the acl anthology, 2019.
- [30] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [31] J. A. Perez, F. Deligianni, D. Ravi, and G.-Z. Yang. Artificial intelligence and robotics, 2018.
- [32] A. Rozovskaya and D. Roth. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 924–933, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
- [33] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units, 2015.
- [34] H. H. Somers. Statistical methods in literary analysis. In *The computer and literary style*,, pages 128–140, Kent State University, 1966.
- [35] B. Spillner. *Linguistik und Literaturwissenschaft: Stilforschung, Rhetorik, Textlinguistik*. Reihe Kohlhammer. W. Kohlhammer, 1974.
- [36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.
- [37] E. Stamatatos. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556, Mar. 2009.
- [38] B. Swanson and E. Charniak. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 193–197, Jeju Island, Korea, July 2012. Association for Computational Linguistics.
- [39] M. C. Templin. *Certain Language Skills in Children: Their Development and Interrelationships*, volume 26. University of Minnesota Press, ned - new edition edition, 1957.
- [40] J. Tetreault, D. Blanchard, and A. Cahill. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative*

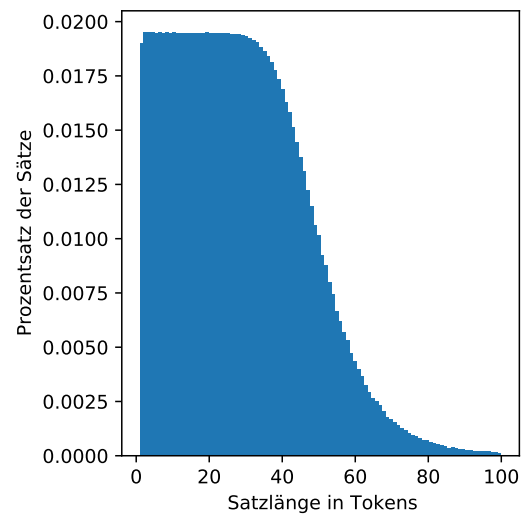
*Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

- [41] J. Tetreault, D. Blanchard, A. Cahill, and M. Chodorow. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of COLING 2012*, pages 2585–2602, Mumbai, India, Dec. 2012. The COLING 2012 Organizing Committee.
- [42] J. Torruella and R. Capsada. Lexical statistics and tipological structures: A measure of lexical richness. *Procedia - Social and Behavioral Sciences*, 95:447 – 454, 2013. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).
- [43] S. Vajjala, B. Majumder, A. Gupta, and H. Surana. *Practical Natural Language Processing*. O’Reilly Media Inc., 2020.
- [44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.
- [45] R. Weischedel et al. Ontonotes release 5.0. Technical report, LDC2013T19 Web Download. Philadelphia: Linguistic Data Consortium,, 2013.
- [46] J. Wong and M. Dras. Exploiting parse structures for native language identification. pages 1600–1610, 01 2011.

## A. Wort- und Satzlängenhistogramme



(a) Wortlängen



(b) Satzlängen

**Abb. A.1.:** Histogramme über alle Dokumente hinsichtlich durchschnittlicher Zeichenlänge von Wörtern und Wortlänge von Sätzen. Artefakte und fehlerhafte Strukturinformationen verzerren die Ergebnisse.

## B. Visualisierung der gelernten Subwortgewichtung

Higher-Order Constituent Parsing and Parser Combination  
Abstract  
This paper presents a higher-order model for constituent parsing aimed at utilizing more local structural context to decide the score of a grammar rule instance in a parse tree. Experiments on English and Chinese treebanks confirm its advantage over its first-order version. It achieves its best F1 scores of 91.86% and 85.58% on the two languages, respectively, and further pushes them to 92.80% and 85.60% via combination with other highperformance parsers.

(a) Bereinigter Text

Higher-Order Constituent Parsing and Parser Combination  
Xiao Chen and Chunyu Kit  
Department of Chinese, Translation and Linguistics  
City University of Hong Kong  
Tat Chee Avenue, Kowloon, Hong Kong SAR, China  
{cxiao2,ctckit}@cityu.edu.hk  
Abstract  
This paper presents a higher-order model for constituent parsing aimed at utilizing more local structural context to decide the score of a grammar rule instance in a parse tree. Experiments on English and Chinese treebanks confirm its advantage over its first-order version. It achieves its best F1 scores of 91.86% and 85.58% on the two languages, respectively, and further pushes them to 92.80% and 85.60% via combination with other highperformance parsers.

(b) Originaltext

**Abb. B.1.:** Visualisierung der gelernten Gewichtung des verdeckten LSTM-Zustands zu jedem Subwort zur Klassifikation des Herkunftslandes mit einem Ausschnitt des bereinigten und ursprünglichen Text.

## C. PCA-Projektion des internen Modellzustands



**Abb. C.1.:** Dreidimensionale *Principal Component Analysis*-Projektion des internen Modellzustands  $X_d$  (siehe Gleichung 3.8) aller Dokumente der Klassen USA (blau), China (rot) und UK (rosa). Zur besseren Übersicht auf drei Klassen beschränkt. Erstellt mittels <https://projector.tensorflow.org/>.

## D. Klassifikation einzelner Zeilen



**Abb. D.1.:** Dreidimensionale *Principal Component Analysis*-Projektion des internen Modellzustands jeder Zeile  $X_{zd}$  eines zufälligen Dokuments (siehe Gleichung 3.6). Über die Kosinus-Ähnlichkeit zu Schwerpunkten der Dokumentvektoren  $X_d$  kann jeder Zeile eines der Länder USA (blau), China (rot) oder UK (rosa) zugeordnet werden.

### Beispiele für Zeilen mit hoher Ähnlichkeit

- USA:

There is a large body of literature in the social sciences that studies power as a social construct (e.g., [1]).

- China:

level, when the KL divergence is adopted, our VGrow naturally gives birth to CFG-GAN while the adoption of JS

- UK:

these models still lags behind that of the aforementioned pre-trained models.