Semantische Analyse von gesprochenen und geschriebenen Dokumenten mit Methoden des tiefen Lernens

Prof. Dr.-Ing. Elmar Nöth
Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen-Nürnberg

Ein Ansatz zur frühen Erkennung von Kriminalität ist die zielgerichtete Erfassung, Verarbeitung, Fusion und Aggregation aller legal zugänglichen Daten und Informationen, die mehrheitlich als unstrukturierte, multilinguale Sprach- und Textdokumente vorliegen. Automatisierung ist zur rationellen Verarbeitung solcher Massendaten zwingend erforderlich.

Moderne Methoden des Maschinellen Lernens erlauben die automatische Analyse großer Sammlungen sprachlicher Dokumente. Während dies früher meistens wissensbasiert durchgeführt wurde, werden heutzutage fast ausschließlich tiefe neuronale Netze zur Analyse verwendet. In diesem Beitrag diskutieren wir die Vorgehensweise bei der flachen linguistischen Erschließung von sprachlichen Dokumenten mit maschinellen Lernverfahren.

Elmar Nöth

Elmar Nöth ist Professor für Angewandte Informatik an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). Er studierte in Erlangen sowie am M.I.T. und erhielt den Dipl.-Inf. und den Dr.-Ing. Abschluss an der FAU 1985 bzw. 1990. Er erhielt seine Habilitation im Jahr 2006.

Seit 1990 war er Assistenzprofessor am Lehrstuhl für Mustererkennung der FAU. Seit 2008 ist er ordentlicher Professor am selben Institut und Leiter der Sprachgruppe. Er ist Autor oder Co-Autor von mehr als 500 Artikeln.

Seine aktuellen Interessen sind Prosodie, Analyse von pathologischer Sprache, Emotionsanalyse und Analyse von Tierkommunikation.

elmar.noeth@fau.de

Wir danken der Initiative von Philip Morris International, das Projekt im folgenden Sinne mit ca. 1,5 Mio. US$ maßgeblich zu fördern: Combating illegal trade, together. A global initiative to support projects against illegal trade and related crimes.

Sprachliche Dokumente können als Audiodatei, Bilddatei (gescanntes Dokument) oder als maschinenlesbares Dokument vorliegen. Bei der Analyse dieser Dokumente sollen verschiedene Fragenstellungen untersucht werden, die das Dokument unterschiedlich tief erschließen.

In einem ersten Schritt müssen die Dokumente in eine maschinenlesbare Form umgewandelt und vorverarbeitet werden:

Audiodaten

Weltweit existieren ca. 6000 verschiedene Sprachen. Zumindest für die großen Verkehrssprachen (z.B. für Englisch, Spanisch, Portugiesisch, Französisch, und Deutsch) aber auch für Sprachen mit vergleichsweise wenigen Sprechern wie z.B. das Slowenische mit ca. 2,5 Mio. Sprechern gibt es leistungsfähige Erkenner, welche das gesprochene Signal in die Folge der am wahrscheinlichsten gesprochenen Wörter umsetzen (Speech-to-Text). Die Fehlerraten hängen sehr stark von der Qualität des aufgenommenen Signals und dem verwendeten Vokabular ab, können durchaus im Bereich von über 95% liegen. Auch für „kleine Sprachen“ werden sehr gute Ergebnisse erzielt, denn die Erkenner werden mit großen Sprachen initialisiert und können dann mit weniger Trainingsdaten auf die Sprache angepasst werden. Während bis vor kurzem typischerweise Hidden Markov Modelle verwendet wurden, kommen heutzutage meistens rekurrente Neuronale Netze (RNNs) zum Einsatz. Je nach Gesamtanalyse-Ziel kann es sein, dass nicht nur erfasst wird, was der Sprecher gesagt hat (Wortfolge), sondern auch wie (z.B. verärgert oder traurig).

Gescannte Dokumente

Mit ähnlichen Verfahren wie bei der Spracherkennung werden gescannte Dokumente in die Folge der am wahrscheinlichsten geschriebenen Wörter umgesetzt. Unter Umständen muss eine Segmentierung in Teildokumente vorgenommen werden, etwa in gedruckten Text, Abbildungen und handgeschriebene Anmerkungen.

Maschinenlesbare Dokumente

Auch hier kann eine Umwandlung notwendig sein. Liegt z.B. die Ausgabe einer Tageszeitung vor, so müssen die Artikel segmentiert werden. Druck-, Formatierungs- oder auch HTML-Befehle werden entfernt. Ebenso wie bei Audiodaten können gewisse Symbole als störend entfernt werden oder wichtig für die Gesamtanalyse sein, wie z.B. Emojis in Social Media Beiträgen, welche Information enthalten über die Einstellung des Autors.

Linguistische Analyse

Das Dokument liegt nun als (fehlerhafte) Wortfolge vor. Die linguistische Analyse wird mit Verfahren durchgeführt, die anhand großer Datenmengen trainiert werden. Analyseziele werden im Folgenden beschrieben.

Übersetzung des Dokumentes

Die Übersetzung ist insbesondere sehr gut, wenn eine der Sprachen Englisch ist, da ca. 60% der Texte im Internet in English geschrieben sind. Daher liegen für praktisch alle Sprachen mehr parallele Texte der Sprache und Englisch zur Verfügung als zu irgendeiner anderen Sprache. Das Übersetzungsprogramm trainiert anhand paralleler Texte, eine englische Folge von Wörtern zu erzeugen, wenn ein Dokument in der Quellsprache vorliegt. Englisch nimmt die Rolle einer „Interlingua“ ein.

Flache linguistische Interpretation oder Kategorisierung

Ziel dieses Schrittes ist es, nicht zu sagen, was gesagt/geschrieben wurde, sondern was gemeint war in dem vorliegenden Dokument. Der Detaillierungsgrad der Analyse hängt von der Anwendung und der Größe der Trainingsdaten ab. Angenommen, zwei Dokumente sollen in der Analyse in eine Kategorie fallen, unterscheiden sich aber deutlich in Länge und Vokabular. Hier helfen Verfahren wie Bert und Word2vec, die an Milliarden von Sequenzen fortlaufender englischer Wörter unabhängig von irgendwelchen Anwendungen trainiert wurden. Zunächst einmal werden Funktionswörter (der, es, eine, …) entfernt und Wörter in ihre Grundform (gab → geben) umgewandelt. Dann erstellt man eine Wortliste und nimmt diese als den festen Wortschatz für alle Dokumente an. Zusätzlich wird ein Wort „Nicht-im-Vokabular“ eingeführt. Somit kann jeder Text in einen Vektor, der so lang ist wie die Wortliste+1, abgebildet werden und bei dem der Wert jedes Vektorelements der Anzahl der Vorkommen dieses Wortes im Trainingsdokument entspricht. In einem Neuronalen Netz mit vielen Schichten wird nun dieser Vektor als Eingangsvektor und als Ausgangsvektor benutzt. In den verborgenen Schichten sind Vektoren mit deutlich weniger Elementen. Pro Schicht wird die Zahl der Vektorelemente kleiner bis zu einem Minimum (Flaschenhals, Bottleneck) und danach wieder größer. Der Wert in einem Knoten ergibt sich aus der gewichteten Summe der Knoten der vorherigen Schicht. Die Gewichte werden zufällig initialisiert und im Training so verändert, dass der resultierende Vektor in der letzten Schicht für jedes Trainingsdokument dem Ausgangsvektor (und somit dem eingegebenen Text) möglichst nahe kommt. Somit lernt das Verfahren, einen „beliebigen“ Text in einen Vektor der Größe der Bottleneck-Schicht abzubilden, ohne dass irgendeiner der Trainingstexte vorher gelabelt werden muss. Ersetzt man nun alle Schichten nach der Bottleneck-Schicht durch eine Ausgangsschicht, in der jedes Vektorelement einer Kategorie entspricht (One-Hot-Encoding) und trainiert das Netz mit Dokumenten nach, bei denen die Kategorie des Dokuments bekannt ist, so kann man mit relativ wenigen Daten einen Klassifikator trainieren, der einem beliebigen, zuvor noch nie gesehenen Text eine Kategorie zuordnet.

Stimmungsanalyse

Ziel dieser Analyse ist es, die Einstellung des Autors bzw. des Dokumentes festzulegen. Die Vorgehensweise ist so wie bei der linguistischen Analyse, nur dass die Knoten der Ausgangsschicht nun den verschiedenen Stimmungsklassen entsprechen. Die Netze können mit derselben Parametereinstellung bis zum Bottleneck initialisiert werden. Dann werden alle folgenden Schichten durch eine Ausgangsschicht ersetzt, welche statt der linguistischen Kategorien die Stimmungskategorien kodiert. Bei gesprochenen Dokumenten kann dies mit einem Emotionserkenner aufgrund akustischer Information verknüpft werden.

Beispielszenario

Im Folgenden soll ein kurzes Beispiel vorgestellt werden, welches den Einsatz aller Aspekte der inhaltlichen Erschließung, die oben vorgestellt wurden, beinhaltet. Ein Konzern möchte nach einer wichtigen Unternehmensentscheidung (z.B. Ankündigung eines größeren EU-weiten Stellenabbaus) die Stimmung in der Bevölkerung und in politischen und

sozialen Gruppen innerhalb der EU zu dieser Entscheidung einschätzen. Hierzu werden verschiedene Kanäle analysiert:

- Zeitungen und Magazine
- Fernsehnachrichten
- Fernseh-Talkrunden
- Internet-Foren
- Social Media.

Die Kanäle werden aufbereitet (Spracherkennung und Textaufbereitung) und ins Englische übersetzt. Danach werden die Dokumente als relevant/nicht relevant kategorisiert. Diese Kategorisierung kann verschiedene Dimensionen enthalten, wie Art des Kanals, Art des Dokuments (Kommentar vs. Berichterstattung), Einstellung zu der Entscheidung des Konzerns (z.B. positiv bis negativ auf einer 5-Punkte-Skala), Grundeinstellung des Verfassers (eine positive Berichterstattung von einer gewerkschaftlich orientierten Quelle ist anders zu behandeln als von einer Arbeitgeberquelle). Aus all diesen einzelnen Beurteilungen wird dann ein Gesamtbild erstellt: Wie stark war die Reaktion in den verschiedenen Bevölkerungsgruppen? Wie schnell war das Abklingen des Interesses an dem Thema? Ist eine Kampagne zu erwarten, die dem Konzern schaden kann? Das Ergebnis dieser Analyse kann dem Konzern als Entscheidungshilfe bei der weiteren Konzernpolitik dienen.

Fazit

Die vorgestellten Verfahren sind auf einer gewissen Abstraktionsebene nicht neu. Was neu ist, ist die Verfügbarkeit der Softwarepakete in einem vortrainierten Zustand. Jeder Entwickler kann Module aus dem Internet herunterladen, die (meistens für englischsprachige) große Korpora trainiert wurden. Durch die Übersetzungsverfahren können eigene, szenarioabhängige und annotierte Korpora benutzt werden, um diese Module auf die eigenen Bedürfnisse sprachenunabhängig anzupassen. Dies geschieht durch Nachtrainieren der initialisierten Module, welches mit wesentlich weniger Daten und Zeitaufwand durchgeführt werden kann, als wenn man bei null anfangen müsste.

Der Zugang zu frei zugänglichen sprachlichen Quellen, die Möglichkeit, ein vortrainiertes State-of-the-Art-System auf die eigenen Bedürfnisse anzupassen, der preiswerte Zugang zu Rechen- und Speicherleistung, all dies macht die automatische Erschließung von Dokumenten seit wenigen Jahren viel einfacher möglich. Die Interpretation eines Dokumentes ist weiterhin sehr flach. Selbst bei einem perfekt übersetzten Dokument weiß man noch nicht, was gemeint ist in dem Text. Selbst bei einer guten Kategorisierung in Bezug auf Inhalt und Stimmung kann man Detailfragen nicht beantworten. Aber aus dem Zusammenbringen von sehr vielen flach analysierten Daten und ihren Metadaten (wie, wer, was, wann) lassen sich Tendenzen erkennen, die notwendige Entscheidungshilfen sein können. Durch die reduzierte Handarbeit müssen die Verfahren nicht aufwändig auf veränderte Situationen angepasst werden, so dass der finanzielle Aufwand im Rahmen bleibt.