Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

Seminararbeit

Neueste Trends in Big Data Analytics
Natural Language Processing

Guangyu Ge

4gge@informatik.uni-hamburg.de
Studiengang Wirtschafsinformatik
Matr.-Nr. 6708726
Fachsemester 9

Betreuer: Tobias Finn
Abgabe: 30.03.2018

Inhaltsverzeichnis

1 Einleitung

Mit der kontinuierlichen Entwicklung von Informationstechnologie und Cloud-Computing, Internet der Dinge, Social Networking tauchen Neue Technologien und Dienste weiterhin auf und werden häufig genutzt. Der Datenumfang wächst sehr schnell und Datentypen nehmen zu. Die Zeitalter von Big Data ist gekommen. Bei Regierungsentscheidungen, bei der Geschäftsplanung und bei der Wissensentdeckung spielen Big Data eine wichtige Rolle. Im Laufe der Zeit ist Big Data zu einer wichtigen strategischen Ressource geworden, die von der Regierung, der Industrie und der Wissenschaft breite Beachtung gefunden hat. Deshalb liegt die Entwicklung der Informationstechnologie nicht nur in der Datenerfassung, -speicherung und -übertragung, sondern auch mehr auf die Verarbeitung, Analyse und Verwendung von Daten. Aufgrund großer Datenmengen bei Big Data wie Variabilität, dem stetigen und massiven Zuwachs an neuen Daten, potenziell schlechter Datenqualität und ihrer Komplexität verwenden Sie effektive künstliche Intelligenz-Technologie, um abstrakte Informationen aus Big Data zu erhalten und in nützliches Wissen zu verwandeln. Es ist eines der Kernprobleme, mit denen die Analyse großer Datenmengen konfrontiert ist. Big Data Analytics sind in vielen Teilgebieten geteilt, z.B.: Maschinelles lernen, Deep Learning, Data Engineering, Speicher- und Compute, Analyseverfahren und Algorithmen, Visualisierung. In dieser Seminararbeit wird der Bereich über Natural Language Processing bearbeitet.

1.1 Motivation

Im Alltag sind wir ständig mit oben genannten Phänomen konfrontiert. Das beginnt schon mit den automatischen Ergänzungsvorschlägen bei der Eingabe von Suchbegriffen in den gängigen Suchmaschinen. Dabei hat sich mir die Frage gestellt, wie ein Programm mich so gut „verstehen“ und mich „kennen“ kann. Ein Programm ist schließlich kein Mensch, sondern nur von Menschen gemacht. Dass Programme „erlernen“ wie ich denke und sogar prognostizieren, was mich in Zukunft bewegen wird, ist schon sehr erstaunlich. Das Natural Language Processing ist auch sehr weit verbreitet und erleichtert meinen Alltag sehr. Oftmals benutze ich Siri oder Cortana, um die Suche nach einer Sache oder einem Begriff zu erleichtern. Bisher hatte ich es nur genutzt, aber nicht über die dahinter verborgenen Programme, Technik oder Prozesse nachgedacht. Ein weiteres beeindruckendes Beispiel war für mich die Tatsache, dass die Produktbeschreibungen bei Amazon auch nicht mehr von Menschen formuliert, sondern automatisch programmiert sind – und das in einer sehr ansprechenden Qualität. Mit anderen Worten stellte dieses

Thema für mich eine ideale Gelegenheit dar, tiefer in die Materie einzusteigen und somit meinen Horizont und meine Wahrnehmung diesbezüglich zu erweitern.

1.2 Aufbau der Arbeit

Das Ziel dieser Arbeit ist, ein tieferes Verständnis für das Natural Language Processing als Bestandteil der Big Data Analysis/Künstliche Intelligenz zu gewinnen. Als grundlegende Quellen werden die Punkte aus der vorangegangenen Präsentation extrahiert und um Erkenntnisse aus der Fachliteratur ergänzt. Aus den gewonnen theoretischen Erkenntnissen soll abgeleitet werden, wie Künstliche Intelligenz für die verbale Interaktion zwischen Mensch und Maschine genutzt werden kann und welche Rolle das NLP dabei spielt.
In Kapitel zwei wird zunächst der Begriff des Natural Language Processing näher erklärt. Auch die geschichtliche Entstehung und Entwicklung dieses Phänomens wird in diesem Kapitel näher beleuchtet. Kapitel drei setzt sich inhaltlich mit den Vorteilen als auch Grenzen des NLP anhand konkreter Beispiele auseinander. In Kapitel vier werden die dem NLP zugrundeliegenden Programme und Prozesse detaillierter beschrieben. Abschließend werden alle gewonnenen Erkenntnisse in Kapitel 5 in einem Fazit zusammengefasst und auch ein Ausblick auf die zukünftige Entwicklung des Natural Language Processing gegeben.

2 Begriffliche Abgrenzung

2.1 Definiton

Natural Language Processing (oder auch: Natural Language Programming, kurz: NLP dt. maschinelle Verarbeitung natürlicher Sprache) bezieht sich auf die Fähigkeit der Maschine, menschliche Schreib- und Sprechmuster zu verstehen und zu erklären. Das Ziel von NLP ist es, dass Computer/Maschinen so intelligent sind wie Menschen, um Sprache zu verstehen. Ziel ist es, die Lücke zwischen menschlicher Kommunikation (natürliche Sprache) und Computerverständnis (Maschinensprache) zu schließen. Die Verarbeitung natürlicher Sprache ist ein sehr weit gefasster Begriff. Die große Mehrheit der natürlichsprachlichen Technologien kann als Verarbeitung natürlicher Sprache klassifiziert werden, einschließlich Parsing, Übersetzung, Spracherkennung, Abstracts, automatische Modifikation, Dialog und so weiter. Abhängig von der zu erfüllenden Aufgabe kann es notwendig sein, eine große Anzahl verschiedener Technologien zu integrieren. Angenommen: die Zielarbeit ist, aus einer großen Menge von Textkorpora einen Kommentar zu einem Produkt zu finden, benötigen wir die Informationserfassungstechnologie. Wenn wir weiter beurteilen müssen, ob die Bewertung eines Produkts positiv oder negativ ist, brauchen wir die Sentiment-Analyse-Technologie. Viele natürliche Sprachtechnologien haben maschinelle Lernumgebungen (und maschinelles Lernen funktioniert am besten), aber es bedeutet nicht, dass es beim maschinellen Lernen ausschließlich um die Verarbeitung natürlicher Sprache geht (da traditionelle Suchtechniken oft kein maschinelles Lernen verwenden).

2.2 Geschichte

Die Verarbeitung natürlicher Sprache begann allgemein in den 1950er Jahren, obwohl die Überlegungen zu dieser Thematik schon davor begannen. Turing veröffentlichte 1950 den Aufsatz Computer and Intelligenceünd schlug den sogenannten Turing-Testäls Voraussetzung für die Beurteilung der Intelligenz vor. Das Georgetown-Experiment von 1954 beinhaltete eine automatische Übersetzung von mehr als 60 russischen Sätzen ins Englische. Damaliger Forscher behaupten, die Probleme der maschinellen Übersetzung innerhalb von drei bis fünf Jahren zu lösen. Der tatsächliche Fortschritt war jedoch viel geringer als erwartet: Der ALPAC-Bericht aus dem Jahr 1966 fasst sie Ergebnisse der 10-Jahres-Studie zusammen und legt offen, dass das erwartete Ziel nicht erreicht wurde und

somit die Forschungsfinanzierung für maschinelle Übersetzung drastisch reduziert wurde. Bis in die späten 1980er Jahre wurden statistische Maschinenübersetzungssysteme entwickelt und das Studium der maschinellen Übersetzung konnte vorangetrieben werden. Zu den besonders erfolgreichen NLP-Systemen, die in den 1960er Jahren entwickelt wurden, zählten SHRDLU - ein natürliches Sprachsystem mit einer Vokabelgrenze, ein begrenztes Bildungssystem wie "Building Block Worldünd das von Joseph Wiesenbaum in den Jahren 1964-1966 designte ELIZA. Das Design von ELIZA - das kaum die Botschaft menschlicher Gedanken und Gefühle verwendet - kann manchmal überraschend ähnlich der Interaktion zwischen Menschen sein. Ziel dieses Programms war die Heilung mental erkrankter Menschen durch die Interaktion mit diesem Programm. Aber es gab auch eine Grenze: Wenn die Fragen von Patienten"das minimale Wissen von ELIZA überschritten, erhielten sie möglicherweise eine leere Antwort. Ein Beispiel: ist die Äußerung Ïch habe Heimweh."Die Antwort ELIZAs lautete: "Was bedeutet Heimweh?Ïn den 1970er Jahren begannen Programmierer mit dem Entwurf von "konzeptioneller OntologieProgrammen, die die Informationen in der realen Welt zu Daten machten, die der Computer verstehen konnte. Beispiele sind MARGIE, SAM, PAM, TaleSpin, QUAL, Politik und Plot-Einheiten. Viele Chatbots wurden während dieser Zeit geschrieben, darunter PARRY, Racter und Jabberwacky.

3 Komplexität bei NLP

NLP basiert auf dem Grundgedanken, dass jegliche Form von Sprache, gesprochen oder geschrieben, zunächst erkannt werden muss. Sprache ist jedoch ein sehr komplexes System von Zeichen. Wichtig ist dabei nicht nur das einzelne Wort, sondern sein Zusammenhang mit anderen Wörtern, ganzen Sätzen oder Sachverhalten. Was Menschen natürlicherweise von Geburt an lernen, müssen Computer mit Hilfe von Algorithmen erreichen. Während der Mensch auf seine Lebenserfahrung zurückgreifen kann, muss der Computer auf künstlich erzeugte Erfahrungen zurückgreifen können. Die Herausforderung für die maschinelle Verarbeitung natürlicher Sprache besteht folglich weniger im Produzieren von Sprache, sondern darin, sie zu verstehen. Der Mechanismus der natürlichen Sprachverarbeitung umfasst zwei Prozesse:
• Natürliches Verständnis der Sprache
• Sprachgenerierung

3.1 Natürliches Verständnis der Sprache

NLU(Natural Language Understanding)versucht die Bedeutung von gegebenem Text zu verstehen. Die Art und Struktur jedes Wortes im Text muss für NLU bekannt sein. Zum Verständnis der Struktur versucht das NLU, die folgende Mehrdeutigkeit in natürlicher Sprache aufzulösen.
Lexikalische Ambiguität: Wörter haben mehrere Bedeutungen. Ein Beispiel: Anna hat ihren Rucksack neben der Bank verloren. Bank zum Sitzen oder wo man Geld abheben kann.
Syntaktische Ambiguität: Satz hat mehrere Parse Bäume. Ein Beispiel: junge Männer und Kinder: [[junge Männer] und [Kinder]] und [junge [Männer und Kinder]]
Anaphoric Mehrdeutigkeit: Phrase oder Wort, das zuvor erwähnt wird, aber eine andere Bedeutung hat. Als nächstes wird der Sinn jedes Wortes durch die Verwendung von Lexika (Vokabular) und Satz von grammatikalischen Regeln verstanden.

3.2 Sprachgenerierung

Bestimmte Wörter haben jedoch eine ähnliche Bedeutung (Synonyme) und Wörter mit mehr als einer Bedeutung (Polysemie). Es ist der Prozess der automatischen Erzeugung von Text aus strukturierten Daten in einem lesbaren Format mit aussagekräftigen Sätzen und Sätzen. Das Problem der Erzeugung natürlicher Sprache ist schwer zu lösen. Es ist

eine Teilmenge von NLP. Natürliche Sprache Generation in drei vorgeschlagenen Stufen unterteilt:
Textplanung: Damit die Bereitstellung des primären Inhalts in strukturierten Daten erfolgen kann.
Satzplanung: Die Sätze werden mit strukturierten Daten kombiniert, um den Informationsfluss darzustellen.
Realisierung: Grammatisch korrekte Sätze werden schließlich erzeugt, um Text darzustellen.

Bei dem Gebiet von Dialogsystem haben die Technologie Giganten auf dem Markt haben alle ihre eigenen intelligenten Sprachassistenten entwickelt, zum Beispiel Google Assistant, Apple Siri, Amazon Alexa und Microsoft Cortana usw. Aber die sind leider noch nicht so ideal/intelligent, wie man sich wünscht. Gemäß der aktuellen Entwicklungsstufe der künstlichen Intelligenz gibt es zwei wesentliche Elemente, um KI-Assistenten besser zu entwickeln:
• Angemessene führende NLP-Technologie
• eine massive, qualitativ hochwertige Sprache Kommunikationsdatensatz

4 Verfahren bei NLP

4.1 N-Gramm

N-Gramm sind das Ergebenis der Zerlegung eines Textes in Fragmente. Der Text wird dabei zerlegt, und jeweils aufeinanderfolgende Fragmente werden als N-Gramm zusammengefasst. Die Fragmente können Buchstaben, Phoneme, Wörter und Ähnliches sein. Normalerweise werden N-Gramm von Text oder Korpus genommen. N= 1, genannt unigram; N= 2, genannt bigram; N= 3, genannt trigram, u.s.w. Ein Beispiel: „Informationsverarbeitung“ als ein Text. So sind die Item vom Text unter 5-Gramm wie folgende: Infor, nform, forma, ormat, rmati, matio, ation, tions, ionsv, onsve, nsver, svera, verar, erarb, rarbe, arbei,rbeit, beitu, eitun, itung.
Manchmal kommt auch „ “ am Anfang ergänzt, um einfach zu suchen. Noch ein Beispiel: „Das ist nur ein Beispiel.“ Bigram: Das ist, ist nur, nur ein, ein Beispiel. Der Grund für dieses Sprachmodell basiert auf der Idee: In der gesamten Sprachumgebung, die Wahrscheinlichkeit des Auftretens von Satz T, besteht aus der Auftrittswahrscheinlichkeit von N Elementen, die T bilden, wie in der folgenden Formel gezeigt:
P(T)=P($W_1W_2W_3W_n$) $= P(W_1)P(W_2|W_1)P(W_3|W_1W_2)\dots P(W_n|W_1W_2\dots W_{n-1})$

Die obige Formel ist schwierig anzuwenden. An dieser Stelle erscheint das Markov-Modell, Das Modell basiert darauf, das Erscheinen eines Wortes hängt nur von den wenigen Wörtern ab, die davor erscheinen. Dies vereinfacht die obige Formel stark.
P(W_1)$P(W_2|W_1)P(W_3|W_1W_2)\dots P(W_n|W_1W_2\dots W_{n-1})$
$\approx P(W_1)P(W_2|W_1)P(W_3|W_2)\dots P(W_n|W_{n-1})$

4.2 Word2Vec

Die Elemente in NLP sind in Form von Symbolen, die abstrakte der Menschheit zusammengefasst. (wie Chinesisch, Englisch, Latein, etc.), so dass sie in numerische Form umgewandelt werden müssen oder in einen mathematischen Raum eingebettet werden müssen, nämlich Word-Einbettung. Word2vec, ist eine Methode vom Worteinbettung. Word2Vec lernt tatsächlich den Text, um die semantische Information des Wortes durch Wortvektor auszudrücken. Das heißt, durch einen eingebetteten Raum, so dass semantisch ähnliche Wörter im Raum innerhalb einer kurzen Entfernung. Einbetten ist eigentlich ein Mapping, das Wort vom ursprünglichen Raum zu einem neuen multidimensionalen Raum-Mapping, das heißt, der Raum, wo die ursprünglichen Wörter in einen neuen Raum eingebettet werden. Die häufige verwandete Methoe in Word2Vec

sind Skip-Gramm (Continuous Skip-gram Model) und CBOW (Continuous Bag-Of-Words Model). Skip-Gram ist das Eingabewort , um den Kontext vorherzusagen. CBOW ist ein gegebener Kontext, um ein eingegebenes Wort vorherzusagen. Natürliche Sprachverständlichkeitsprobleme führen zu maschinellen Lernproblemen, und der erste Schritt besteht definitiv darin, einen Weg zu finden, diese Symbole zu berechnen. Bei NLP ist das intuitivste und am häufigsten verwendete Wortschreibweise ist die One-hot-Darstellung, die jedes Wort als langen Vektor darstellt. Die Dimension dieses Vektors ist die Größe des Vokabulars, wobei die meisten Elemente 0 sind und nur eine Dimension den Wert 1 hat. Diese Dimension repräsentiert das aktuelle Wort. Zum Beispiel: "Mikrofon"wird dargestellt als [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ...] "Mike"wird ausgedrückt als [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 ...] Jedes Wort ist eines der großen 0 im Meer. Diese One-Hot-Darstellung wäre, wenn sie spärlich gespeichert wäre, sehr einfach: Weisen Sie jedem Wort eine numerische ID zu. Beispiel: Im Beispiel ist das Mikrofon als 3 markiert, Mike als 8 (angenommen, dass mit 0 beginnt). Wenn Sie programmgesteuert eine Hash-Tabelle verwenden möchten, weisen Sie jedem Wort eine Zahl zu. Eine solche prägnante Ausdrucksmethode arbeitet mit Algorithmen wie maximaler Entropie, SVM, CRF usw. zusammen und hat bereits verschiedene Hauptaufgaben in NLP gut erfüllt. Natürlich gibt es auch bei dieser Ausdrucksweise ein wichtiges Problem: das "WortlückenPhänomen: Zwei beliebige Wörter sind isoliert. Licht kann aus diesen beiden Vektoren nicht sehen, ob die zwei Wörter miteinander verwandt sind oder nicht, selbst Synonyme wie Mikrofone und Mikrofone können nicht verschont bleiben.

Skrip-Gramm

Bevor ich das Skip-Gramm-Modell im Detail erkläre, verstehen wir zunächst das Format der Trainingsdaten. Die Eingabe in das Überspringen-Gramm-Modell ist ein Wort wI, dessen Ausgabe der Kontext von wI w, w1, ..., wO, C ist, und die Fenstergröße des Kontexts ist C. Hier ist zum Beispiel der Satz "Das ist nur ein Test für Experiment". Wenn wir Testäls Trainingseingabedaten verwenden, sind die Wortgruppen: { "Das", "ist", "nur", "ein", "für", "Experiment"} die Ausgabe. Für all diese Wörter werden wir eine Ein-Hot-Codierung durchführen. Das Diagramm für das übersprungene Grammmodell ist wie folgt: Abblidung 4.1 skip gram

Als nächstes betrachten wir das neuronale Netzwerk-Modell des Skip-Gramms.Das neuronale Netzwerk-Modell mit einem Sprung-Gramm wurde aus dem vorwärtsgekoppeltenneuralen Netzwerkmodell verbessert, um es auf der Grundlage des neuronalen Feed-Forward-Netzwerkmodellszu verdeutlichen.Das neuronales Wellengittermodell sieht wie folgend aus: Abblidung 4.2 Wellengittermodell

In der obigen Abbildung repräsentiert der Eingangsvektor x die Ein-Hot-Codierung eines Wortes und des entsprechenden Ausgangsvektors y1, ..., yC. Die i-te Reihe der Gewichtsmatrix W zwischen der Eingabeschicht und der versteckten Schicht repräsentiert das Gewicht des i-ten Wortes in dem Vokabular.

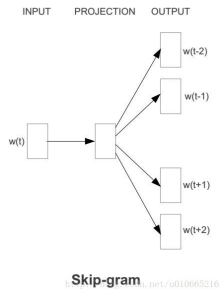

Abbildung 4.1: skip gram

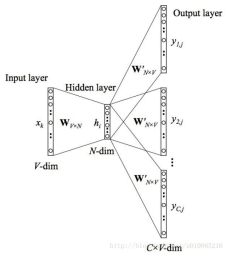

Abbildung 4.2: Wellengittermodell

5 Zusammenfassung und Ausblick

Zusammenfassende lässt sich sagen, dass NLP in der letzten Zeit sich schnell entwickelt. Die Identifizierung von Einzelschallquellen in Innenräumen hat den Menschen überschritten, aber in einer lauten Umgebung sind Gespräche mit mehreren Personen nicht ideal. Es gibt mehr Raum für Verbesserungen. Die wichtigsten grundlegenden Technologie-Tools im Bereich der Verarbeitung natürlicher Sprache haben sich nicht dramatisch verändert. Die wichtigsten technologischen Mittel sind immer noch die folgenden Technologie kombinationen: Word Embedding, LSTM, Sequence-to-Sequence-Framework usw. In dieser Seminararbeit ist nur die Verfahren von Word2Vec beim Bereich Word Embedding bearbeitet.
Mit der kontinuierlichen Entwicklung der Big Data Technologie häufen sich große Korpora Stichprobendaten kontinuierlich in einer erstaunlichen Anzahl und die Verarbeitung natürlicher Sprache vertieft sich weiter in tiefes Lernen. Zehntausende von Stunden Proben wurden für das Modelltraining verwendet Bald wird die Entwicklung der Verarbeitungstechnologie für natürliche Sprache bald in die 100.000-Stunden-Datensample-Trainingsphase eintreten damit eine Vielzahl von Benutzerakzentunterschieden, Mehrdomänen, Mehrdeutigkeitsdaten und komplexe Grammatikregeln abgedeckt werden können. Mit der schnellen Zunahme der Menge an Trainingsdaten wird die Implementierung von LSTM- (Long- und Short-Term-Memory-Model) Modellierung und effektivem Training von CTC (Connection Timing Classification) zu einem technischen Kernproblem.

Literaturverzeichnis

[1] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781, 2013.

[2] Zhinan Dong: NLP Zusammenfassung,
http://www.infoq.com/cn/articles/2015-Review-NLP

[3] licstar: Deep Learning in NLP,
http://licstar.net/archives/328

[4] Deep Learning word2vec,
https://blog.csdn.net/mytestmy/article/details/26961315

[5] https://www.zhihu.com/question/44832436

[6] Chris McCormick: Word2Vec Tutorial,
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/