

Probabilistische online Wissensgraphkonstruktion aus natürlicher Sprache

Clemens Damke

6. September 2017
Version: Entwurf 1



PADERBORN UNIVERSITY
The University for the Information Society

Department of Electrical Engineering,
Computer Science and Mathematics
Warburger Straße 100
33098 Paderborn



INTELLIGENT
SYSTEMS

Intelligent Systems Group (ISG)

Bachelorarbeit

Probabilistische online Wissensgraphkonstruktion aus natürlicher Sprache

Clemens Damke

- | | |
|--------------|---|
| 1. Korrektor | Prof. Dr. Eyke Hüllermeier
Institut für Informatik
Universität Paderborn |
| 2. Korrektor | Prof. Dr. Axel-Cyrille Ngonga Ngomo
Institut für Informatik
Universität Paderborn |
| Betreuer | Dr. Theodor Lettmann und Prof. Dr. Eyke Hüllermeier |

6. September 2017

Clemens Damke

Probabilistische online Wissensgraphkonstruktion aus natürlicher Sprache

Bachelorarbeit, 6. September 2017

Korrektoren: Prof. Dr. Eyke Hüllermeier und Prof. Dr. Axel-Cyrille Ngonga Ngomo

Betreuer: Dr. Theodor Lettmann und Prof. Dr. Eyke Hüllermeier

Universität Paderborn

Intelligente Systeme

Institut für Informatik

Pohlweg 51

33098 Paderborn

Abstract

Hallo Welt. Test5.

Inhaltsverzeichnis

1	Einleitung	3
1.1	Motivation	3
1.2	Ziele der Arbeit	4
1.3	Aufbau der Arbeit	5
2	Verwandte Arbeiten	7
2.1	Ansätze zur Wissensrepräsentation	7
2.1.1	Logische Grundlagen	7
2.1.2	Entwicklung maschineller Wissensrepräsentation	9
2.1.3	Aktuelle Wissensrepräsentationsprojekte	11
2.2	NLP-Werkzeuge	12
2.3	Wissensgraphkonstruktionsverfahren	14
3	Theoretische Grundlagen	17
3.1	Wissensmodellierung mit Konzeptgraphen	17
3.2	Dependency Parsing und Coreference Resolution	17
3.3	Modellierung von Hinge-Loss-MRFs mit PSL	17
4	Vorgeschlagenes Wissensgraphkonstruktionsverfahren	19
4.1	Wissensgraphontologie	19
4.2	Graph-Persistenzschicht	19
4.3	NLP-Phase	19
4.4	Graphkonstruktionsphase	19
5	Auswertung	21
5.1	Testmethode	21
5.2	Ergebnisse	21
6	Zusammenfassung	23
A	Anhang	27
	Literatur	29

Einleitung

“ *The actual world cannot be distinguished from a world of imagination by any description. Hence the need of pronoun and indices, and the more complicated the subject the greater the need of them.*

— **Charles Sanders Peirce**
Mathematiker und Philosoph

1.1 Motivation

In den letzten Jahren hat die Repräsentation von Wissensbasen durch Graphen, sog. Wissensgraphen, immer mehr an Bedeutung gewonnen. Google, Bing und IBM Watson benutzen solche Wissensgraphen z. B. zum Beantworten von komplexen Suchanfragen.

Die Grundidee dabei ist es, Entitäten durch Knoten und Relationen durch Kanten abzubilden. Entitäten können konkrete Dinge, wie z. B. Personen, aber auch abstrakte Konzepte, wie z. B. historische Epochen, sein. Relationen beschreiben beliebige Beziehungen zwischen den Entitäten, z. B. $person(\text{Da Vinci}) \xrightarrow{\text{lived in}} epoch(\text{Renaissance})$. Die Entität, von der eine solche Relation ausgeht, wird als Subjekt und die Zielentität als Objekt der Relation bezeichnet.

Die Typen von Entitäten bzw. Relationen (z. B. *person* bzw. *lived in*) und deren Bedeutung sind dabei i. d. R. formal in einer sog. Ontologie spezifiziert. Die Ontologie beschränkt also die Menge gültiger Wissensgraphen, was eine effiziente maschinelle Verarbeitung der im Graph enthaltenen Informationen ermöglicht.

Da Wissensgraphen in zahlreichen Domänen einsetzbar sind, wird deren automatisierte Konstruktion bereits seit Jahren erforscht. Manuelles Konstruieren und vor allem anschließendes Warten und Aktualisieren von Wissensgraphen, ist aufgrund der abzubildenden Datenmengen nicht praktikabel. Bei einer maschinellen automatisierten Konstruktion sind insbesondere zwei Anforderungen problematisch:

1. Das Verarbeiten von unstrukturierten Eingaben, wie z. B. natürlichsprachlichen Texten.
2. Effizientes Eingliedern neuer Informationen in einen bestehenden Wissensgraphen. Dieses Eingliedern von Informationen umfasst im Speziellen:
 - **Entity Resolution:** Hinzukommende Entitäten, die bereits im Graphen enthalten sind, müssen als Duplikate erkannt werden. Dies ist i. d. R. nicht trivial, da die selbe Entität durch viele verschiedene, oftmals vom Kontext abhängige, Token repräsentiert werden kann; z. B. *Bob* vs. *Robert* oder *Der Papst* vs. *Franziskus*.
 - **Link Prediction:** Hinzukommende Entitäten müssen mit bereits bestehenden Entitäten in Relation gesetzt werden. Hinzukommende Relationen können zudem benutzt werden um andere Relationen zu inferieren; z. B.

$$female(A) \wedge B \xrightarrow{\text{son of}} A \implies A \xrightarrow{\text{mother of}} B$$

Die Kombination dieser beiden Anforderungen ist interessant, da das meiste verfügbare Wissen in natürlichsprachlicher Textform vorliegt und zudem permanent neues Wissen entsteht. Ein automatisiertes Wissensgraphkonstruktionsverfahren, welches beide Anforderungen berücksichtigt, ist daher in diversen Domänen von Nutzen. Ein Beispiel hierfür ist die Auswertung von Kommunikationsdaten aus E-Mails oder Chat-Nachrichten mit dem Ziel die sozialen Beziehungen und Intentionen der Kommunikationspartner zu ermitteln.

1.2 Ziele der Arbeit

Das übergeordnete Ziel dieser Arbeit ist es, ein Verfahren zu finden, welches das soeben beschriebene Problem der automatisierten Wissensgraphkonstruktion für E-Mail-Daten löst. Konkret sei ein Stream von E-Mails gegeben, denen jeweils ein Inhalt, ein Absender, eine Menge von Empfängern und ggf. weitere Metadaten, wie z. B. Absendezeit, Absendeort oder IP-Adresse, zugeordnet ist. Die Nachrichteninhalte werden der Einfachheit halber als ausschließlich englischsprachig angenommen. Außerdem wird eine, für E-Mails und andere Kurznachrichten typische, eingeschränkte Sprachkomplexität angenommen. Die Nachrichten sollen nacheinander in das zu konstruierende System eingefügt werden, welches sukzessive einen Wissensgraphen daraus erzeugt.

Für diese Erzeugung muss eine Reihe von Teilproblemen gelöst werden:

1. **Onotologie:** Spezifikation einer Wissensgraphontologie, die mächtig genug ist, um die Diversität natürlichsprachlich beschriebener Informationen abzubilden.
2. **Repräsentation:** Spezifikation der maschinellen Repräsentation des Wissensgraphen.
3. **Sprachverarbeitung:** Finden eines Verfahrens, welches die natürlichsprachlichen Inhalte der Nachrichten in eine für die Wissensgraphkonstruktion geeignete Form bringt.
4. **Grapherweiterung:** Finden eines Verfahrens, um eine eintreffende Nachricht in den bestehenden Wissensgraphen einzufügen.

Das aus den Teillösungen zusammengesetzte Verfahren muss, neben der offensichtlichen Anforderung einen Wissensgraphen zu konstruieren, zudem folgende technische Anforderungen erfüllen:

1. **Erweiterbarkeit:** Es sollen Schnittstellen eingeplant sein, um neben der Sprachverarbeitung auch andere Verarbeitungsverfahren, z. B. für Bilder, hinzufügen zu können. Die Graphontologie, Graphrepräsentation und das Grapherweiterungsverfahren dürfen also nicht zu sehr auf die Struktur natürlicher Sprache zugeschnitten sein.
2. **Parallelisierbarkeit:** Das Verfahren soll in der Lage sein die Rechenleistung mehrerer Prozessorkerne zu nutzen. Diese Anforderung betrifft insbesondere das Grapherweiterungsverfahren.

1.3 Aufbau der Arbeit

Kapitel 2

Kapitel 3

Kapitel 4

Kapitel 5

Kapitel 6

Verwandte Arbeiten

Die in 1.2 beschriebenen Ziele werden bereits seit langem erforscht. Der Begriff *Wissensgraph* wurde 2012 durch Google popularisiert, die Ideen dahinter lassen sich allerdings bis ins Ende des 19. Jahrhunderts zurückverfolgen. Dieses Kapitel zeigt auf, wie sich die Themen dieser Arbeit in die bisherige Forschung einfügen. 2.1 ordnet hierfür das Konzept des Wissensgraphen in die Entwicklungsgeschichte der Wissensrepräsentation ein. In 2.2 wird ein Überblick über die momentan verbreiteten *natural language processing* (NLP) Werkzeuge gegeben. Das Wissensgraphkonzept und NLP wird schließlich in 2.3 kombiniert und es werden die aktuell verwendeten Verfahren zur Wissensgraphkonstruktion beschrieben.

2.1 Ansätze zur Wissensrepräsentation

2.1.1 Logische Grundlagen

Die Entwicklung der Wissensrepräsentation hängt eng mit der Entwicklung der Logik zusammen. Während in der formalen Logik und Mathematik die Prädikatenlogik das allgemein verwendete Kalkül ist und alternative Formalismen kaum verbreitet sind, finden im Bereich der Wissensrepräsentation bis heute diverse andere Kalküle Verwendung. Diese werden im Folgenden kurz vorgestellt.

Begriffsschrift (1879) Gottlob Freges Buch über die *Begriffsschrift* gilt als eines der bedeutsamsten Werke der Logik. Sie ist der erste Formalismus mit der Mächtigkeit der modernen Prädikatenlogik zweiter Ordnung mit Identität. Frege benutzt hierfür eine zweidimensionale Notation, die sich stark von der heute gebräuchlichen, linearen, an die Algebra angelehnte Notation unterscheidet.

$$\begin{array}{|c} \vdash^a \\ \hline \begin{array}{|c} \vdash \mathfrak{A}(a) \\ \vdash \mathfrak{B}(a) \end{array} \end{array} \Leftrightarrow \exists a : P(a) \vee R(a)$$

Im Gegensatz zur Prädikatenlogik gibt es keine eigene Syntax für *UND* und *ODER*; diese Operatoren müssen durch die Kombination von Negation und Implikation abgebildet werden. Zudem gibt es ausschließlich den Allquantor; eine existenzquantisierte

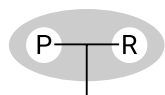
Aussage muss durch Negation der negierten allquantisierten Aussage ausgedrückt werden.

Prädikatenlogik: Peirce Notation (1885) Unabhängig von Frege entwickelte der amerikanische Mathematiker Charles Sanders Peirce ebenfalls ein prädikatenlogisches Kalkül. Peirces Notation hatte starke Ähnlichkeiten mit der heute benutzten linearen Schreibweise. Statt den modernen Symbolen hat Peirce allerdings die algebraischen Operatoren benutzt, um die Analogien zwischen Logik und Algebra auszudrücken.

$$\Sigma_a P_a + R_a \Leftrightarrow \exists a : P(a) \vee R(a)$$

Existential Graphs (1897) Neben seiner zuvor entwickelten linearen prädikatenlogischen Notation, hat Peirce zudem viele Jahre an einem alternativen graphischen Kalkül gearbeitet, welches er *Existential Graphs* (zu dt. Existenzgraphen) nannte. Ähnlich wie die Begriffsschrift werden Existenzgraphen zweidimensional dargestellt. Von dieser Gemeinsamkeit abgesehen, funktionieren sie allerdings fundamental verschieden. Ein logischer Ausdruck wird hier durch einen ungerichteten Graphen beschrieben. Die konkrete räumliche Anordnung der Knoten und Kanten hat dabei keine semantische Relevanz.

Peirce hat Existenzgraphen als ein dreistufiges aufeinander aufbauendes System konzipiert. Die erste Stufe, die sog. α -Graphen, umfasst alle notwendigen syntaktischen Elemente, um ein Kalkül mit der Mächtigkeit der Aussagenlogik zu erhalten. Die β -Graphen bilden die zweite Stufe und erweitern die Syntax der α -Graphen, sodass die Mächtigkeit der Prädikatenlogik erster Ordnung erreicht wird. Sowohl für α -, als auch für β -Graphen, ist die Vollständigkeit und Korrektheit bewiesen. Die dritte Stufe (γ -Graphen) wurde von Pierce nie vollendet; sie deckt in etwa die Mächtigkeit der heutigen Prädikatenlogik höherer Ordnung sowie der Modallogik ab.



$$\Leftrightarrow \exists a : P(a) \vee R(a)$$

Wie schon die Begriffsschrift, sind Existenzgraphen syntaktisch minimal. Direkt ausdrücken lässt sich lediglich *UND*, der Existenzquantor und die Negation. Ein weiterer Unterschied zur heutigen Prädikatenlogik ist die Beschreibung logischer Inferenzen. Im Gegensatz zu den prädikatenlogischen Ersetzungsaxiomen, die auf der syntaktischen Struktur von logischen Ausdrücken operieren (z. B. für Kommutativität), lassen sich die Ersetzungsaxiome für Existenzgraphen als Graphtransformationsregeln verstehen, die bestimmte Teilmengen der Knoten und Kanten eines Ausdrucks durch andere äquivalente Knoten- und Kantenmengen ersetzen.

Prädikatenlogik: Peano-Russell Notation (1910) Die zweidimensionalen Notationen wurde häufig kritisiert, da sie die lineare, algebraische Notation der symbolischen Logik von Boole und De Morgan verwarfen. Freges Begriffsschrift und Peirces Existenzgraphen konnten sich daher nicht durchsetzen. Peirces algebraische prädikatenlogische Notation hingegen, stieß auf größere Akzeptanz. Giuseppe Peano hat auf deren Basis eine ähnliche Notation entwickelt, welche allerdings nicht die algebraischen Operatoren benutzt, damit sich logische Ausdrücke besser mit mathematischen Ausdrücken kombinieren lassen. Bertrand Russell hat Peanos Notation anschließend in leicht abgewandelter Form in den *Principia Mathematica* (1910) benutzt. Diese sog. Peano-Russell-Notation ist im Wesentlichen identisch mit der modernen Schreibweise.

Trotz des Verschwindens der zweidimensionalen Notationen, finden sich noch heute Anlehnungen daran. So ist z. B. die Negation $\neg A$ auf Freges negierten Inhaltssstrich $\neg A$ und der Ableitungsoperator \vdash auf Freges Urteilsstrich mit angefügtem Inhaltsstrich \vdash zurückzuführen.

2.1.2 Entwicklung maschineller Wissensrepräsentation

Die Idee Computer zur Lösung beliebiger Probleme zu benutzen ist nicht neu. Da ein solches maschinelles Problemlösen immer die Verfügbarkeit von Hintergrundwissen über die Problemdomäne erfordert, wurden Methoden zur Wissensrepräsentation immer im Zusammenhang mit Problemlösern erforscht. So wie effiziente Datenstrukturen die Implementation effizienter Algorithmen ermöglichen, ermöglichen gute Wissensrepräsentationen die Implementation guter Problemlöser. Was genau nun als ein guter Problemlöser verstanden wird, hat sich im Laufe der Jahre allerdings immer wieder verändert.

Universelle Problemlöser Einer der ersten maschinellen Problemlöser war der von Simon, Shaw und Newell 1955 entwickelte *Logic Theorist* (LT). LT war in der Lage logische Aussagen zu beweisen, indem er systematisch Ersetzungsaxiome auf eine gegebene Aussage angewandt hat, bis die gesuchte Lösung abgeleitet wurde.

Die Grundidee des LT haben Simon, Shaw und Newell 1959 im *General Problem Solver* (GPS) erweitert. Es wurden Heuristiken hinzugefügt, um den Suchraum geschickter zu durchlaufen. GPS war ein universeller Problemlöser, konnte also jedes Problem lösen, das sich durch eine Menge von Horn-Klauseln ausdrücken lässt. Zwar war es so theoretisch möglich Probleme aus diversen Domänen zu lösen, aufgrund der kombinatorischen Explosion war GPS allerdings nicht zur Lösung komplexer praktischer Probleme geeignet.

Expertensysteme Aufgrund der Misserfolge universeller Problemlöser für praktische Probleme, hat die Forschung begonnen sich mehr auf die Entwicklung von Expertensystemen zu fokussieren. Expertensysteme besitzen für gewöhnlich eine Wissensbasis in der domänenspezifisches Wissen in Form von Regeln und Fakten kodiert ist. Eine sog. Inferenzmaschine benutzt diese Regeln und Fakten um Probleme zu lösen.

Semantic Networks (1956) Die Idee, Graphen als Datenstruktur für Wissensbasen zu verwenden, taucht erstmal in den sog. *Semantic Networks* (zu dt. semantische Netzwerke) auf. Dieser Ansatz beschreibt Wissen als Menge von *(subject, predicate, object)*-Tripeln. Es gibt darüber hinaus keine klaren Regeln, wie ein semantisches Netz strukturiert sein muss.

Conceptual Graphs (1976) Wie genau mit Graphen komplexes Wissen beschrieben werden kann, das über eine reine Taxonomie hinaus geht, blieb bei semantischen Netzen unklar. John F. Sowa's *Conceptual Graphs* (zu dt. Konzeptgraphen) lösen dieses Problem. Statt Wissen als eine Menge von Beziehungen abzubilden, wird es als prädikatenlogischer Ausdruck verstanden. Hierfür baut Sowa auf Peirces Existenzgraphen auf, die bis dahin weitestgehend unbeachtet waren.



Dieser Ansatz erlaubt es komplexe Wissensbasen zu konstruieren, in denen nicht nur gespeichert werden kann, ob ein Konzept existiert, sondern auch, ob es nicht oder nur möglicherweise existiert. Da Konzeptgraphen, so wie schon die Existenzgraphen, ein vollständiges und korrektes Logikkalkül sind, lassen sich zudem Inferenzregeln für sie definieren. Der Vorteil hierfür einen Graphen statt eines prädikatenlogischen Ausdrucks zu verwenden ist, dass eine Graphstruktur einen deutlich effizienteren Zugriff auf gespeichertes Wissen ermöglicht.

Knowledge Graphs (1987) Der Begriff *Knowledge Graph* (zu dt. Wissensgraph) bezeichnete ursprünglich eine Klasse semantischer Netze, deren Relationsmenge formal spezifiziert ist. Dies schränkt die Menge erlaubter Graphen ein und ermöglicht die Definition von Inferenzregeln, um Schlussfolgerungen aus einem gegebenen Graphen zu ziehen. Im Laufe der Jahre ist die Grenze zwischen semantischen Netzen und Wissensgraphen allerdings immer weiter verschwommen, sodass die Bezeichnungen heute oft synonym verwendet werden. Wissensgraphen und Konzeptgraphen müssen weiterhin unterschieden werden, da erstere oftmals Negation und Modalität nicht unterstützen.

2.1.3 Aktuelle Wissensrepräsentationsprojekte

Manuelle Ansätze

Semantic Web Das sog. *Semantic Web* bezeichnet eine Menge von W3C-Standards, die das bestehende Web um eine formale Wissensbeschreibungssyntax erweitern. Zentral ist dabei das *Resource Description Framework* (RDF), mit dem sich beliebige Konzepte, auch Ressourcen genannt, beschreiben und verknüpfen lassen. Ziel ist es über die unstrukturierte Netzstruktur des bestehenden Webs, eine strukturierte, leicht maschiell verarbeitbare, Netzstruktur zu legen. Durch die Anfragesprache *SPARQL* ist es möglich Wissen aus diesem Netz auszulesen. Das Web würde somit zu einem großen dezentralen Wissensgraphen. Tim Berners-Lee beschreibt diese Idee als das "Web 3.0". Obwohl die Technologien hierfür bereits seit Jahren existieren, sind bislang nur wenige Webseiten mit RDF-Tags annotiert. Häufige Kritik ist, dass das Semantic Web zu viel theoretisches Hintergrundwissen über Wissensrepräsentationsverfahren erfordert, um für die meisten Webseitenbetreiber zugänglich zu sein.

WordNet Das *WordNet* der Universität Princeton ist ein frei verfügbares lexikalisch-semantisches Netz für die englische Sprache, d. h. ein semantisches Netz, welches die Bedeutung von Worten in Relation zueinander setzt. Relationen werden dabei z. B. für Synonyme, Hyperonyme (Oberbegriffe) und Meronyme (Bestandteile) eingefügt. Der Datenbestand des WordNets wird manuell gepflegt und resultiert aus der Kombination der Einträge verschiedener Wörterbücher.

Automatisierte Ansätze

Neben den manuellen Grapherzeugungsansätzen des Semantic Webs und des WordNets, gibt es diverse voll- und semiautomatische Ansätze. Diese bauen die Graphstruktur selbstständig aus gegebenen Datenquellen auf.

NELL Das *Never-Ending Language Learning* (NELL) System traversiert selbstständig das Internet und fügt die gefundenen textuellen Informationen in einen Wissensgraphen ein. Hierfür wird eine Kombination verschiedener Modelle verwendet, die regelmäßig angepasst wird. Menschen können optional Feedback für die extrahierten Fakten geben, um die Inferenzqualität weiter zu verbessern.

Google Knowledge Graph Basierend auf den Ideen der in 2.1.2 vorgestellten Wissensgraphen, stellte Google 2012 eine eigene, ebenfalls *knowledge graph* genannte, Wissensgraphentechnologie vor. Sie wird benutzt, um Suchanfragen semantisch, statt



Abb. 2.1. Popularität des Begriffs “knowledge graph” (Quelle: Google Trends)

per String-Matching, zu beantworten. So können z. B. zum Suchbegriff verwandte Ergebnisse angezeigt werden, selbst wenn es keine textuelle Ähnlichkeit zu jenem gibt. Laut Googles Aussagen stammen die Quelldaten u. a. aus Wikipedia Infoboxen, Wikidata und dem CIA World Factbook. Da es sich hierbei, im Gegensatz zu NELL, primär um strukturierte Daten handelt, ist das automatisierte Einpflegen mit hoher Genauigkeit möglich. Wie genau die Daten im Graph repräsentiert werden, ist nicht öffentlich bekannt.

2.2 NLP-Werkzeuge

Neben der Repräsentation von Wissen, ist auch die Verarbeitung natürlicher Sprache ein Kernbestandteil dieser Arbeit. Hierfür existiert bereits eine Vielzahl von *natural language processing* (NLP) Werkzeugen. Trotz dieser Vielfalt lassen sich Kernverarbeitungsschritte festmachen, die in den meisten Werkzeugen verwendet werden.

1. **Tokenization:** Oftmals einer der ersten Verarbeitungsschritte einer NLP Bibliothek. Eine Eingabezeichenkette wird dabei in eine Liste von Tokens zerlegt. Token sind u. a. Wörter, Satzzeichen und numerische Literale.

“Today I’m testing myself.” \Rightarrow (Today, I, ’m, testing, myself, .)

2. **Lemmatization:** Abbildung von Tokens auf ihre Lemmata (Grundformen).

(Today, I, ’m, testing, myself, .) \Rightarrow (today, I, be, test, myself, .)

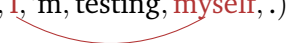
3. **Part-of-speech Tagging:** Abbildung von Tokens auf ihre Wortarten (POS-Tags).

(Today, I, 'm, testing, myself, .) \Rightarrow (adverb,
personal pronoun,
present tense first-person singular verb,
present participle verb,
personal pronoun,
sentence terminator)

4. **Named Entity Recognition:** Klassifizierung von Tokens in Kategorien wie z. B. Person, Ort oder Zeitpunkt.

(Today, I, 'm, testing, myself, .)
date

5. **Coreference Resolution:** Bestimmung von Token-Äquivalenzklassen, die jeweils auf das selbe Konzept verweisen (insbesondere Pronomina und ihr Antezedens).

(Today, I, 'm, testing, myself, .)


6. **Dependency Parsing:** Eine auf den POS-Tags aufbauende syntaktische Analyse, welche die grammatikalischen Abhängigkeiten der Token untereinander ausgibt. Die Menge dieser Abhängigkeiten bildet einen Baum oder baumähnlichen Graphen, der *Treebank* bzw. *Dependency Graph* genannt wird.

(Today, I, 'm, testing, myself, .) \Rightarrow testing $\xrightarrow{\text{subject}}$ I
testing $\xrightarrow{\text{object}}$ myself
testing $\xrightarrow{\text{auxiliary}}$ 'm
testing $\xrightarrow{\text{adverbial modifier}}$ Today

Wie sich erkennen lässt, bauen die Verarbeitungsschritte sukzessive aufeinander auf und bilden eine Art Pipeline. Dieses Pipeline-Modell findet sich auch in vielen NLP-Werkzeugen wieder. Ein solches ist z. B. das Stanford CoreNLP Projekt [Man+14][Cor], welches u. a. Module für alle der soeben vorgestellten Verarbeitungsschritte beinhaltet. Ein weiteres ähnliches Projekt ist Apache OpenNLP [Ope]; es bietet ähnliche Module, wie CoreNLP an.

2.3 Wissensgraphkonstruktionsverfahren

Wie in 2.1.3 gezeigt, gibt es diverse Ansätze um Graphen aus Daten zu konstruieren. Da für das Thema dieser Arbeit insbesondere automatisierte Verfahren relevant sind, die mit unstrukturierten Daten, wie z. B. natürlicher Sprache, umgehen können, werden diese im Folgenden näher beschrieben.

Üblicherweise arbeiten Wissensgraphkonstruktionsverfahren nicht direkt mit den unstrukturierten Eingabedaten, wie z. B. den Inhalten von E-Mails, sondern mit einer Knoten- bzw. Konzeptmenge und ggf. auch einer Kanten- bzw. Relationsmenge, die zuvor, z. B. mittels eines in 2.2 vorgestellten NLP-Verfahrens, aus den Rohdaten extrahiert wurden. Die Wissensgraphkonstruktion ist somit äquivalent zum Problem der *Link Prediction*, also dem Finden von Relationen zwischen den gegebenen Konzepten. Die Link Prediction wiederum lässt sich als ein Problem des *Statistical Relational Learnings* (SRL) auffassen. In der Literatur finden sich im Wesentlichen drei Klassen von SRL-Verfahren [Nic+16], die auf verschiedenen Annahmen über die Korrelation der zu verknüpfenden Informationen basieren:

1. **Latent Feature Models:** Alle Relationen werden als bedingt unabhängig angenommen, sofern bestimmte Eigenschaften über Subjekte und Objekte der Relationen gegeben sind.
2. **Graph Feature Models:** Alle Relationen werden als bedingt unabhängig angenommen, sofern bestimmte Eigenschaften der Struktur des Graphen gegeben sind.
3. **Markov Random Fields:** Es wird angenommen und erlaubt, dass alle Relationen lokale Abhängigkeiten voneinander haben können.

Latent Feature Models Ein Beispiel für ein Latent Feature Modell ist das RESCAL-Verfahren [Nic13], welches auf Tensorfaktorisierung basiert. Die Grundidee dabei ist es, allen Entitäten i einen Feature-Vektor $e_i \in \mathbb{R}^H$ zuzuordnen und für alle Relationen k eine Gewichtsmatrix $W_k \in \mathbb{R}^{H \times H}$ zu finden. Die Konfidenz in die Existenz einer Relation $i \xrightarrow{k} j$ wird durch $e_i^T W_k e_j$ beschrieben. Diese Definition ermöglicht eine sehr schnelle Link Prediction, da lediglich ein Vektor-Matrix-Vektor-Produkt berechnet werden muss. RESCAL liefert gute Ergebnisse, wenn die vorherzusagenden Relationen globale Abhängigkeiten aufweisen. Lokal stark zusammenhängende Teilgraphen werden allerdings schlecht erkannt, da nur der Feature-Vektor und nicht die Nachbarschaft einer Entität berücksichtigt wird; ein Beispiel hierfür sind symmetrische Relationen.

$$A \xrightarrow{\text{married to}} B \implies B \xrightarrow{\text{married to}} A$$

Graph Feature Models Komplementär zu den Latent Feature Modellen sind die Graph Feature Modelle. Statt Entitäten in einen Feature-Raum einzubetten, wird hier die Nachbarschaft der Entitäten betrachtet. Ein Beispiel hierfür ist der Path Ranking Algorithmus (PRA) [Lao+11]. PRA ermittelt Relationen durch zufälliges Durchwandern des Graphen. Um die Stärken der Latent Feature und Graph Feature Modelle zu kombinieren, wurden Hybrid-Modelle, wie z. B. das Additive Relational Effects (ARE) Verfahren [Nic+14], entwickelt, welches die Konfidenzen von RESCAL und PRA addiert.

Markov Random Fields Fundamental verschieden von diesen beiden Verfahren sind Markov Random Fields (MRFs). Hier sind prinzipiell Abhängigkeiten zwischen allen Relationen möglich, was MRFs sehr flexibel macht. Da dies hinsichtlich der Laufzeit schnell impraktikabel wird, wird das Modell um einen Abhängigkeitsgraphen erweitert, der die Anzahl von betrachteten Abhängigkeiten reduziert. Der Abhängigkeitsgraph darf dabei nicht mit dem Wissensgraphen verwechselt werden: Ersterer beschreibt statistische Abhängigkeiten zwischen Klassen von Relationen, während letzterer Abhängigkeiten zwischen spezifischen Fakten beschreibt. Zur Modellierung von Abhängigkeitsgraphen werden i. d. R. Kalküle verwendet, die an eine Prädikatenlogik erster Ordnung angelehnt sind. Das Finden eines Wissensgraphen ist in diesem Modell äquivalent zum Lösen des MAX-SAT-Problems. Wählt man ein Kalkül in dem die Atome stetig, also aus $[0, 1]$ statt aus $\{0, 1\}$ sind, und Formeln zudem ausschließlich Disjunktionen und Negationen gemäß Łukasiewicz S-Norm enthalten, erhält man ein sog. Hinge-Loss-MRF [Bac+13][Bac+15], da die Loss-Funktion der Disjunktion in dieser S-Norm ein Hinge-Loss ist.

Ein konkretes Kalkül, welches sich zur Spezifikation von HL-MRFs eignet, ist die Probabilistic Soft Logic (PSL) [Br10][Bac+15]. MAX-SAT lässt sich für solche HL-MRFs effizient und parallelisierbar mit dem konvexen Alternating Direction Method of Multipliers (ADMM) Optimierungsverfahren [Boy+11] lösen. In dessen ursprünglichen Form ist ADMM allerdings ausschließlich für offline Inferenz geeignet; der Wissensgraph müsste also bei jeder Eingabe neu konstruiert werden. Um dieses Problem zu lösen, wurde das Budgeted Online Collective Inference (BOCI) Verfahren [Puj+15] entwickelt. BOCI nutzt Metadaten, die während der Ausführung von ADMM anfallen, um eine Bewertung für jedes Atom zu berechnen. Die Bewertung eines Atoms beschreibt, wie groß die erwartete Wertveränderung beim Eintreffen neuer Informationen ist. Kommen nun neue Informationen an, müssen diese ausschließlich zusammen mit den m höchstbewerteten bereits existierenden Atomen betrachtet werden, die Werte aller anderen Atome werden fixiert. Je höher das Budget m , desto höher ist die Qualität im Vergleich zu einer Neukonstruktion des Graphen. Es wurde empirisch gezeigt, dass die Inferenzqualität mit BOCI oft nur unwesentlich schlechter ist als bei einer kompletten offline Inferenz.

Mittels ADMM und BOCI lässt sich PSL daher für die online Wissensgraphkonstruktion einsetzen. Der Vorteil dieses Ansatzes gegenüber eines Latent Feature oder Graph Feature Modells ist, dass sich andere domänenspezifische Expertensysteme leicht in eine PSL Inferenz integrieren lassen. PSL erlaubt nämlich die Inklusion von benutzerdefinierten Funktionen und Prädikaten. Diese können benutzt werden, um z. B. die Levenshtein-Distanz zweier Zeichenketten oder domänenspezifisches Hintergrundwissen, wie Distanz zwischen zwei namentlich genannten Orten, mit in die Entity Resolution einfließen zu lassen.

Theoretische Grundlagen

3.1 Wissensmodellierung mit Konzeptgraphen

3.2 Dependency Parsing und Coreference Resolution

3.3 Modellierung von Hinge-Loss-MRFs mit PSL

Vorgeschlagenes Wissensgraph- konstruktionsverfahren

4.1 Wissensgraphontologie

4.2 Graph-Persistenzschicht

4.3 NLP-Phase

4.4 Graphkonstruktionsphase

Auswertung

5.1 Testmethode

5.2 Ergebnisse

Zusammenfassung

6

Anhang

A

Literatur

- [Bac+13] Stephen H. Bach, Bert Huang, Ben London und Lise Getoor. „Hinge-loss Markov Random Fields: Convex Inference for Structured Prediction“. In: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence. UAI'13. Bellevue, WA: AUAI Press, 2013, S. 32–41 (zitiert auf Seite 15).
- [Bac+15] Stephen H. Bach, Matthias Broecheler, Bert Huang und Lise Getoor. „Hinge-Loss Markov Random Fields and Probabilistic Soft Logic“. In: ArXiv:1505.04406 [cs.LG] (2015) (zitiert auf Seite 15).
- [Boy+11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato und Jonathan Eckstein. „Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers“. In: Found. Trends Mach. Learn. 3.1 (Jan. 2011), S. 1–122 (zitiert auf Seite 15).
- [Br10] Matthias Bröcheler, Lilyana Mihalkova und Lise Getoor. „Probabilistic Similarity Logic“. In: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence. UAI'10. Catalina Island, CA: AUAI Press, 2010, S. 73–82 (zitiert auf Seite 15).
- [Lao+11] Ni Lao, Tom Mitchell und William W. Cohen. „Random Walk Inference and Learning in a Large Scale Knowledge Base“. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, S. 529–539 (zitiert auf Seite 15).
- [Man+14] Christopher D. Manning, Mihai Surdeanu, John Bauer et al. „The Stanford CoreNLP Natural Language Processing Toolkit“. In: Association for Computational Linguistics (ACL) System Demonstrations. 2014, S. 55–60 (zitiert auf Seite 13).
- [Nic+14] Maximilian Nickel, Xueyan Jiang und Volker Tresp. „Reducing the Rank in Relational Factorization Models by Including Observable Patterns“. In: Advances in Neural Information Processing Systems 27. Hrsg. von Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence und K. Q. Weinberger. Curran Associates, Inc., 2014, S. 1179–1187 (zitiert auf Seite 15).
- [Nic+16] Maximilian Nickel, Kevin Murphy, Volker Tresp und Evgeniy Gabrilovich. „A Review of Relational Machine Learning for Knowledge Graphs“. In: Proceedings of the IEEE. Bd. 104. 1. Institute of Electrical und Electronics Engineers (IEEE), 2016 (zitiert auf Seite 14).
- [Nic13] Maximilian Nickel. „Tensor Factorization for Relational Learning“. Diss. Ludwig-Maximilians-Universität München, 2013 (zitiert auf Seite 14).

- [Puj+ 15] Jay Pujara, Ben London und Lise Getoor. „Budgeted Online Collective Inference“. In: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence. UAI’15. Amsterdam, Netherlands: AUAI Press, 2015, S. 712–721 (zitiert auf Seite 15).

Webseiten

- [Cor] Stanford CoreNLP. Zuletzt besucht im Juli 2017. URL: <https://stanfordnlp.github.io/CoreNLP> (zitiert auf Seite 13).
- [Ope] Apache OpenNLP. URL: <http://opennlp.apache.org/> (zitiert auf Seite 13).

Abbildungsverzeichnis

2.1	Popularität des Begriffs “knowledge graph” (Quelle: Google Trends) . .	12
-----	--	----

Tabellenverzeichnis

Erklärung zur Bachelorarbeit

Ich, Clemens Damke (Matrikel-Nr. 7011488), versichere, dass ich die Bachelorarbeit mit dem Thema *Probabilistische online Wissensgraphkonstruktion aus natürlicher Sprache* selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Stellen der Arbeit, die ich anderen Werken dem Wortlaut oder dem Sinn nach entnommen habe, wurden in jedem Fall unter Angabe der Quellen der Entlehnung kenntlich gemacht. Das Gleiche gilt auch für Tabellen, Skizzen, Zeichnungen, bildliche Darstellungen usw. Die Bachelorarbeit habe ich nicht, auch nicht auszugsweise, für eine andere abgeschlossene Prüfung angefertigt. Auf § 63 Abs. 5 HZG wird hingewiesen.

Paderborn, 6. September 2017

Clemens Damke