

Lösungsvorschlag für das Übungsblatt 1.

Aufgabe 1.

Zusammengefasst aus Ihren Beiträgen

Wie bewerten sie das System insgesamt?

Das Watson System verdeutlicht den Fortschritt der Künstlichen Intelligenz Forschung/Computerlinguistik/Informatik und zeigt, dass Deep-QA eine vielversprechende Hybrid-Technologie ist, die weiterentwickelt werden sollte. Vor allem die sog. Selbsteinschätzung, ob eine Antwort richtig oder falsch ist, stellt einen Unterschied zu ähnlichen vorhergehenden Systemen dar. Bemerkenswert ist auch, dass die Fragen ohne Hilfe des Internets beantwortet werden, sondern mithilfe einer allerdings sehr großen Datenbank. Ein Nachteil des Systems sind die Kosten der benötigten Hardware. Ein Kritikpunkt war von denn Teilnehmern, dass man u. U. von Beginn an an einem System arbeiten hätte sollen, das wirklich nützlich ist. Zum Beispiel in Gebieten wie der Medizin oder der Wissenschaft ist, statt im US Fernsehen damit für die Firma IBM zu werben. Auf der anderen Seite benötigen muss man schon verstehen, dass aufgrund der Investition von mindestens 60 Mannjahren schon ein gewisser Knalleffekt für IBM herauskommen musste.

Welche Teilgebiete der Computerlinguistik/Linguistik werden im Watson System implementiert?

Das Watson System verwendet Phonetik (Sprachsynthese), Morphologie, Syntax (Deep und Shallow-Parsing), Semantik und auch pragmatische Vorstellungen (Modellierung von Fragentypen) aus der Linguistik. Daneben vor allem Technologien aus dem Bereich Knowledge-Engineering.

Was unterscheidet das Watson System von einem intelligenten System etwa im Sinne des Touring Test?

Der Hauptunterschied zwischen dem Watson System und dem Turing Test ist deren Anwendung. Bei dem Turing Test weiß der Anwender nicht, dass er mit einem Computer interagiert. Bei dem Watson System hingegen, wissen alle, dass es sich um einen Computer handelt und dieser getestet wird. Außerdem muss dieser „nur“ Fragen beantworten, und dies in einer Quizshow üblichen Form, die sonst in Gesprächen zwischen Menschen nicht verwendet wird. – Bedenke: wie würde der Touringtest verlaufen, wenn man einen Strohmännchen als Kandidaten verwendet. Würden die Zuschauer merken, dass hier ein Computer spielt?

von einem Ihrer Kollegen (schön ausführlich)

„Das System wirkt sehr komplex und weit fortgeschritten. Auch wenn es nur auf einen sehr speziellen Aufgabenbereich, das Bestehen in der entsprechenden Fernsehsendung, zugeschnitten ist, wurde es doch so angelegt, dass es leicht in anderen Bereichen eingesetzt werden kann. Fragwürdig ist dennoch, ob die vielen Arbeitsjahre, die für die Entwicklung von Watson eingesetzt wurden, nicht besser für "sinnvollere" Systeme eingesetzt hätten werden sollen; insbesondere Dinge wie die spezielle Anpassung für "Jeopardy" und das technische umsetzen des Regelwerkes erscheinen mir von wenig Nutzen für weitere Projekte. Auf der anderen Seite sollte man IBM durchaus dafür dankbar sein, dass sie Aufmerksamkeit auf computerlinguistische Projekte wie dieses gelenkt haben.

Die Computerlinguistik wird in nahezu allen ihren Teilgebieten verwendet, insbesondere im Bereich „knowledge in speech and language processing“, wie er bei Jurafski/Martin genannt wird.

Anfangen von Phonetik auf einem weit fortgeschrittenen Level (Watson kann auch komplexe, selten benutzte Wörter flüssig und deutlich aussprechen). Dieser Bereich kann hingegen wohl als „Bonus“ verstanden werden und war wohl nicht Hauptfokus der Entwickler.

Morphologie ist sicher auch ein kleiner Teilbereich bei der Entwicklung gewesen (Watson kann erkennen, ob die Frage „Which actor...“ oder „Which actors...“ lautet), allerdings aufgrund der verwendeten englischen Sprache wohl von weniger Bedeutung.

Deutlich wichtiger sind die verbliebenen drei Bereiche: Syntax, Semantik und Pragmatik.

Syntaxerkennung ist wichtig, um die essentiellen Wörter einer Frage zu erkennen – nach „the“ muss wohl im Normalfall nicht in der Datenbank gesucht werden, auch Nomen können je nach Wortstellung üblicherweise mehr oder weniger Bedeutung haben, zum anderen eine Antwort in korrektem Englisch zu formulieren.

Zu erkennen, welche Bedeutung der Satz hat, ist die Aufgabe der Semantik. Pragmatik ist wichtig, um zum Beispiel Wortspiele oder Ironie zu erkennen, die ja ebenfalls Teil von Jeopardy sind.

Zur Analyse dieser Bereiche war sicherlich die Entwicklung von Algorithmen nötig, die bei Jurafski/Martin ebenfalls als eigener Teilbereich bezeichnet werden. Algorithmen waren zum Beispiel nötig, um die Ausgangstexte zu analysieren und in datenbanktaugliche Form zu bringen, zur Gewichtung verschiedener Antwortmöglichkeiten und sicherlich vielen weiteren Bereichen.

Trotzdem ist Watson nicht dazu ausgerichtet, in einem Touring-System zu bestehen. Auch wenn es durchaus Ähnlichkeiten gibt, ist Watson doch nur darauf ausgelegt, Fragen mit kurzen Antworten nach dem immer gleichen Muster (What is.../ Who is...) zu beantworten. In einer Touring Maschine muss ein Computer Antworten auf alle möglichen Fragen individuell beantworten können. Hierbei geht es nicht in erster Linie darum, wissenschaftliche Fragen mit hoher Sicherheit zu beantworten (auch von einem Menschen kann nicht erwartet werden, das zu tun) sondern darum, das "Gegenüber" davon zu überzeugen, kein Computer zu sein. Hier geht es auch darum, Gefühle ausdrücken zu können (bzw. so zu tun) und besonders, individuelle Antworten zu geben, die nicht nach einem speziellen Muster ablaufen. Das Spektrum zu gebender Antworten ist also ungleich weiter.“

Noch ein Beispiel

„Meiner Meinung nach ist Watson ein wichtiger Schritt vorwärts in den Bereichen der Forschung, die sich mit AI beschäftigen, und natürlich in der Computerlinguistik, die eine der wichtigsten Bereiche in die Entwicklung von Watson war. Um erfolgreich in Jeopardy teilzunehmen, muss das System verschiedene Teilgebiete der Computerlinguistik benutzen, da während des Quiz muss Watson nicht nur die Fragen richtig verstehen, sondern auch Gigabytes der Information verarbeiten, um die korrekte Antwort zu finden. Dazu benutzt Watson mehr als 100 computerlinguistische Techniken, um natürliche Sprache zu analysieren, Informationsquellen zu identifizieren, Hypothese zu generieren, dazu Beweise zu finden und endlich aus mehreren Hypothesen eine passendste zu wählen. Zum Beispiel, um eine Frage zu verstehen, analysiert das System ihre syntaktische Struktur, klassifiziert sie, findet Schlüsselwörter und LAT (der lexikalische Typ der Antwort), teilt sie in mehreren Teile, für die es eine und dieselbe Antwort gibt, wenn es nötig ist, usw. Außerdem kann das System selbst entscheiden, welche Fragen zu beantworten, und diese Fähigkeit basiert darauf, dass Watson nur auf die Fragen antwortet, denen Zuverlässigkeitswert höher als ausgemachte Grenze ist. Was die Suche nach der Antwort auf die Frage angeht, sind hier Informationsextraktion und Informationssuche natürlich die wichtigsten Teilgebiete der Computerlinguistik. Watson benutzt mehrere Suchtechniken, die verschiedene Ansätze haben; das System verarbeitet Tausende von Dokumenten, Wörterbücher, Thesauri und Artikel und zieht die passende Variante der Antwort aus. Danach analysiert und vergleicht Watson diese Variante und entscheidet, welche von denen am besten passt. Am Ende, wenn das System die korrekte Lösung schon gewählt hat, benutzt es Software, die für die Sprachsynthese verantwortlich ist, um die Antwort zu sagen. Und man muss nicht vergessen, dass alle beschriebene Prozesse in Laufe von paar Sekunden ablaufen!

Was den Unterschied zwischen Watson und früher erstellten intelligenten Systemen, die am Touring Test teilgenommen haben, angeht, besteht er, meiner Meinung nach, in erster Linie darin, dass das Hauptziel von diesen Systemen die Nachahmung vom Benehmen der Menschen ist. Das bedeutet, dass im Touring Test diese Systemen die Fragen so beantworten mussten, wie Menschen – lange denken, zweifeln, auch Fehler machen und nicht alles wissen. Die Aufgabe von Watson ist dagegen, auf alle Fragen in Jeopardy möglichst schnell und richtig zu antworten.

Zum Schluss kann ich sagen, dass Watson wirklich eine wichtige Erfahrung in Bereichen Computerlinguistik und AI war, weil man viel Neues in diesen Gebieten entwickeln musste damit, dass das System erfolgreich funktionieren würde. Obwohl Watson noch manchmal Fehler macht, ist das System im Großen und Ganzen doch gelungen und ich bewerte es eher positiv.“

ein weiteres Beispiel...

„Das System wirkt sehr komplex und weit fortgeschritten. Auch wenn es nur auf einen sehr speziellen Aufgabenbereich, das Bestehen in der entsprechenden Fernsehsendung, zugeschnitten ist, wurde es doch so angelegt, dass es leicht in anderen Bereichen eingesetzt werden kann. Fragwürdig ist dennoch, ob die vielen Arbeitsjahre, die für die Entwicklung von Watson eingesetzt wurden, nicht besser für "sinnvollere" Systeme eingesetzt hätten werden sollen; insbesondere Dinge wie die spezielle Anpassung für "Jeopardy" und das technische umsetzen des Regelwerkes erscheinen mir von wenig Nutzen für weitere Projekte. Auf der anderen Seite sollte man IBM durchaus dafür dankbar sein, dass sie Aufmerksamkeit auf computerlinguistische Projekte wie dieses gelenkt haben.

Die Computerlinguistik wird in nahezu allen ihren Teilgebieten verwendet, insbesondere im Bereich „knowledge in speech and language processing“, wie er bei Jurafski/Martin genannt wird.

Angefangen von Phonetik auf einem weit fortgeschrittenen Level (Watson kann auch komplexe, selten benutzte Wörter flüssig und deutlich aussprechen). Dieser Bereich kann hingegen wohl als „Bonus“ verstanden werden und war wohl nicht Hauptfokus der Entwickler.

Morphologie ist sicher auch ein kleiner Teilbereich bei der Entwicklung gewesen (Watson kann erkennen, ob die Frage „Which actor...“ oder „Which actors...“ lautet), allerdings aufgrund der verwendeten englischen Sprache wohl von weniger Bedeutung.

Deutlich wichtiger sind die verbliebenen drei Bereiche: Syntax, Semantik und Pragmatik.

Syntaxerkennung ist wichtig, um zum die essentiellen Wörter einer Frage zu erkennen – nach „the“ muss wohl im Normalfall nicht in der Datenbank gesucht werden, auch Nomen können je nach Wortstellung üblicherweise mehr oder weniger Bedeutung haben -, zum anderen eine Antwort in korrektem ENglish zu formulieren.

Zu erkennen, welche Bedeutung der Satz hat, ist die Aufgabe der Semantik. Pragmatik ist wichtig, um zum Beispiel Wortspiele oder Ironie zu erkennen, die ja ebenfalls Teil von Jeopardy sind.

Zur Analyse dieser Bereiche war sicherlich die Entwicklung von Algorithmen nötig, die bei Jurafski/Martin ebenfalls als eigener Teilbereich bezeichnet werden. Algorithmen waren zum Beispiel nötig, um die Ausgangstexte zu analysieren und in datenbanktaugliche Form zu bringen, zur Gewichtung verschiedener Antwortmöglichkeiten und sicherlich vielen weiteren Bereichen.

Trotzdem ist Watson nicht dazu ausgerichtet, in einem Touring-System zu bestehen. Auch wenn es durchaus Ähnlichkeiten gibt, ist Watson doch nur darauf ausgelegt, Fragen mit kurzen Antworten

nach dem immer gleichen Muster (What is.../ Who is...) zu beantworten. In einer Turing Maschine muss ein Computer Antworten auf alle möglichen Fragen individuell beantworten können. Hierbei geht es nicht in erster Linie darum, wissenschaftliche Fragen mit hoher Sicherheit zu beantworten (auch von einem Menschen kann nicht erwartet werden, das zu tun) sondern darum, das "Gegenüber" davon zu überzeugen, kein Computer zu sein. Hier geht es auch darum, Gefühle ausdrücken zu können (bzw so zu tun) und besonders, individuelle Antworten zu geben, die nicht nach einem speziellen Muster ablaufen. Das Spektrum zu gebender Antworten ist also ungleich weiter.“

Aufgabe 2.

a) Das erste a auf dieser Seite ist das schönste.

Tokenlesart – Das spezielle Vorkommen des Buchstaben „a“.

b) A(1) ist ein schöner Buchstabe(2).

1) Typelesart – Buchstaben dieser Art: der konkrete Zeichentyp.

2) Typelesart – Buchstabe als Begriff.

c) Maria kann mit kyrillischen Buchstaben nichts anfangen.

Tokenlesart – einzelne Buchstaben (in diesem Fall kyrillische, es gibt auch andere).

Typelesart – als Gesamtheit von kyrillischen Buchstaben.

d) Buchstabe ist ein Nomen.

Typelesart – Wortart allgemein.

e) Dieser Satz enthält zweimal das Wort Satz.

Tokenlesart – Dieser Satz konkret.

Typelesart – Satz im Allgemeinen in der Syntax.

Und jetzt die Begriffe Type und Token in der Computerlinguistik:

Types die verschiedenen Wörter (Strings) – auf diese Unterscheidung kommen wir noch. Token die Zahl der Wörter (Strings) in einem Dokument.

Betrachte folgenden Satz:

„Das ist das Haus vom Nikolaus.“

5 Types

6 Token

Wenn man den Punkt nicht mitzählt...

