Computerlinguistisches Seminar der Universität Zürich

Semantikrepräsentation für Antwortextraktion SS 2001

Prof. M. Hess, lic. phil. S. Clematide

Question-Answering-Strategien in TREC-8

Abgabedatum: 09.07.2001

Cornelia Steinmann
Kleine Kirchgasse 24
5507 Mellingen
056/ 491'44'53
cornelia_steinmann@gmx.ch

Inhaltsverzeichnis

Einleitung

Question-Answering (Q&A) ist eine Technologie mit der aus grossen Wissensbasen auf konkrete Fragen relevante Antworten ermittelt werden sollen. In der ersten Question-Answering Task innerhalb der TREC-8 wurden Systeme, die Q&A ohne Beschränkung auf ein bestimmtes Thema betreiben, zum ersten Mal miteinander verglichen. Die vorliegende Arbeit stellt in TREC-8 verwendete Strategien zum Question-Answering vor und versucht, Leistungsfähigkeit und Grenzen vielversprechender Methoden aufzuzeigen.

Die Untersuchung stützt sich fast ausschliesslich auf die Track Reports der TREC-8 und die Beiträge der einzelnen Systemhersteller. Die zitierten Seitenzahlen beziehen sich auf die unter http://trec.nist.gov/pubs/trec8/t8_proceedings.html erhältlichen PDF-Dateien (Stand 09.07.01). Da der Grossteil der Systeme speziell für TREC-8 konstruiert wurde, existieren keine weiteren Unterlagen zu den Gesamtsystemen. Zusätzliche Informationen zu einzelnen Systemkomponenten wären verfügbar gewesen, wurden aber nicht verwendet, da für TREC vorgenommene Modifikationen aus ihnen nicht hervorgehen. Nicht technische Details, sondern angewendete Konzepte sollen im Zentrum dieser Arbeit stehen.

Als Grundlage für die Beschreibung der Q&A-Strategien definiere ich den Begriff Question-Answering in der Verwendung der TREC-8 und gebe eine genaue Darstellung von Aufgabenstellung und Rahmenbedingungen der Q&A-Teilaufgabe. Darauf folgt die Beschreibung verwendeter Strategien, geordnet nach Arbeitsschritten, die ein Q&A-System bewältigen muss. Anschliessend sollen Besonderheiten, die sich durch die Fragestellung und verwendete Methoden ergaben, erörtert und die vielversprechendsten Strategien hervorgehoben und auf ihre Weiterverwendbarkeit im Allgemeinen und für LUIS geprüft werden.

Die Fussnoten nennen meistens nicht alle Systeme, die eine bestimmte Teilstrategie verwenden, sondern dienen als Beispielbelege. Eine Liste der verwendeten Abkürzungen befindet sich am Ende dieser Arbeit auf Seite 25. Abgekürzt verwendete Wörter wurden generell mit S-Plural versehen, unabhängig von den Pluralformen der Vollformen.

1. Definition von Question-Answering

Unter Frage-Antwort-Systemen kann man „natürlichsprachliche Zugangssysteme zu Informationssystemen“ mit „natürlichsprachlicher Ein- und Ausgabe“ verstehen.[1] Beispiele für Informationssysteme sind Informationsretrievalsysteme, Datenbanken oder in der TREC-8 Textkorpora, die als Wissensbasis benutzt werden. In dieser Definition werden Q&A-Systeme von Textstellenwiedergewinnungs- und Antwortextraktionssystemen (Passage Retrieval & Answer Extraction) unterschieden, weil bei den letzteren beiden Ansätzen auch unfertige Satzstrukturen als Output zugelassen sind und die Antworten nicht generiert werden.

Das Ziel der Q&A Task in TREC-8 war: „to retrieve small snippets of text that contain the actual answer to a question“.[2] Damit ist die Aufgabestellung auch für Systeme geeignet, die nach der vorherigen Definition als Textstellenwiedergewinnungs- oder Antwortextraktions-systeme gelten. Die Generierung von Antworten war aber ausdrücklich erlaubt.[3]

In den TREC-Konferenzen gelten also nicht nur generierte Sätze als gültige Outputs von Q&A-Systemen, sondern auch Textpassagen, einzelne Wörter oder extrahierte Teil- oder Ganzsätze. In diesem Sinn wird „Question-Answering“ in der vorliegenden Arbeit verwendet.

2. Trec-8 Q&A-Task

2.1 Ziele und Aufgabenstellung[4]

Die Q&A-Task wurde 1999 in TREC-8 zum ersten Mal durchgeführt. Sie hatte zum Ziel, Question-Answering-Systeme miteinander zu vergleichen und allgemein anwendbare Evaluationskriterien für solche Systeme zu ermitteln.[5]

Die Systeme sollten auf 200 gegebene Fragen Antworten vordefinierter Maximallänge aus einem 2 Gigabyte grossen Datenkorpus extrahieren oder auf dieser Basis beruhend generieren. Das Korpus bestand aus Zeitungstexten, transkribierten Radiosendungen und Budgetberichten der Regierung.

Die insgesamt 20 Forschungsgruppen von Universitäten und aus der Privatwirtschaft konnten sich in zwei durch die Antwortlänge unterschiedenen Kategorien beteiligen. In der einen Kategorie wurden mögliche Antworten auf eine Länge von 50 Bytes, das heisst 50 Zeichen, beschränkt, in der anderen standen 250 Bytes zur Verfügung. Jede Organisation konnte pro Kategorie maximal zwei Antwortsets übermitteln. Die Systeme sollten in der ersten Kategorie

---

[1] Hess, Einführung in die Computerlinguistik I, S.16f.
[2] National Institute of Standards and Technology. (NIST). The TREC-8 Question Answering Track Report, E. Voorhees. S. 1.
[3] Ebenda. S. 1.
[4] Dieser Abschnitt stützt sich, falls nicht anders vermerkt, auf: NIST, Track Report, S. 1f.
[5] NIST. The TREC-8 Question Answering Track Evaluation. E. Voorhees, D. Tice. S. 1.

relevante Phrasen, in der zweiten relevante Sätze zurückliefern, die jeweils die Antwort enthalten sollten. In der Regel nützten die Systeme die erlaubte Bytezahl aber voll aus. Sie fügten zusätzliche Satz- oder sogar Wortteile bis zur erlaubten Maximallänge hinzu, um die Chance zu erhöhen, dass die Antwort tatsächlich im String vorhanden war. Eine Ausnahme bildeten AT&T, die mit ihrem Named-Entity-Ansatz in der 50 Byte Kategorie nur die Named Entities (NEs) als Antworten zurücklieferten, sowie CL-Research und die New Mexico State University, die als Antworten ganze Sätze vorschlugen.[6]

2.2 Fragen

2.2.1 Art der Fragen

Die für TREC-8 verwendeten Fragen sollten kurze, maximal 50 Bytes lange Antworten haben und Fakten basiert sein.[7] Wie die Fragetypanalyse der Southern Methodist University (SMU) ergab, konnten deshalb bereits mehr als die Hälfte der Fragen durch Namen von Personen und Organisationen (58 mal), Ortsangaben (41 mal) oder Daten (23 mal) beantwortet werden. Eine Begründung wurde hingegen nur 2 mal verlangt. Der Rest der Antworten bestand zu mehr als der Hälfte aus Zahlangaben mit ihren Einheiten.[8] MITRE gibt die Anzahl der Antworten, die Zahlen verlangen, mit 63 an.[9] Nach einer Analyse von Cymfony verlangten 80% der Fragen eine Named Entity als Antwort.[10]

2.2.2 Herkunft und Auswahl[11]

Um möglichst verschiedene Fragearten und Themenbereiche abzudecken, wählten die Organisatoren von TREC-8 die Testfragen aus einem Pool von 1837 Fragen aus vier verschiedenen Quellen aus. Von den 1500 Fragen aus dem FAQ-Finder der Cornell University wurden nur 24 ins Testset aufgenommen. Weitere 337 Fragen wurden von TREC-Organisatoren, Jurymitgliedern und Systemherstellern beigesteuert. In der Regel wurden die Dokumente von Hand oder mit einer Suchmaschine nach Textstellen abgesucht, zu welchen eine Frage formuliert werden konnte. Dadurch ähnelten viele Fragen dem Antworttext stark. 38 der

---

[6] AT&T. AT&T at TREC-8. A. Singhal, S. Abney, M. Bacchiani, M. Collins, D. Hindle, F. Pereira. S. 6. CL-Reasearch. Question-Answering Using Semantic Relation Triples. K.C. Litkowski. S. 5; New Mexico State University. CRL's TREC-8 Systems Cross-Lingual IR, and Q&A. B. Ogden, J. Cowie, E. Ludovik, H. Molina-Salgado, S. Nirenburg, N. Sharples, S. Sheremtyeva. S. 7.
[7] NIST. Evaluation Report. S. 1.
[8] Nach einer Tabelle aus: Southern Methodist University (SMU). LASSO: A Tool for Surfing the Answer Net. D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Gîrju, V. Rus. S. 3.
[9] The MITRE Corporation. A Sys Called Qanda. E. Breck, J. Burger, L.Ferro, D. House, M. Light, I. Mani. S. 5. Hier zählen allerdings auch Daten und Zeitangaben dazu.
[10] Cymfony Inc. Information Extraction Supported Question Answering. R. Srihari, W. Li. S. 5.
[11] Dieses Unterkapitel bezieht sich auf: NIST. Evaluation Report. S. 2f.

337 handgemachten Fragen wurden den Teilnehmern als Übungsset zur Verfügung gestellt. Weitere 176 Fragen wurden für das Testset verwendet.

Bei der Selektion der Fragen wurden solche, die als ambig oder als allzu deutliche Textumformungen empfunden wurden ebenso eliminiert wie Fragen, die als Antwort eine Liste von mehr als drei Elementen verlangten. Verlangte eine Frage implizit eine zusammengesetzte Antwort, wurde sie entweder weggelassen oder umformuliert. Dementsprechend wurde das Fragewort aus „ Who won the Nobel Prize in medicine in 1992?“ durch „What two US-biochemists“ ersetzt.

2.3 Ablauf

Zur Vorbereitung erhielten die Teilnehmer das erwähnte Übungsset von 38 Fragen. Die Verteilung der Fragetypen im Übungsset war nicht auf diejenige im Testset abgestimmt.[12]

Nach Bekanntgabe der 200 Fragen mussten die jeweiligen Systeme „eingefroren“ werden und durften bis zum Ende der einwöchigen Beantwortungszeit nicht mehr verändert werden.

In beiden Kategorien mussten pro Frage 5 nach ihrer Relevanz geordnete Paare aus einem Textstring und der Identifikationsnummer (Doc-Id) des Referenzdokuments übermittelt werden. Alle diese Strings galten als mögliche Antworten. [13]

2.4 Bewertung

2.4.1 Beurteilungskriterien[14]

Die Korrektheit der Antwortstrings wurde von Menschen als jeweils falsch oder richtig beurteilt. Pro Frage bewerteten drei von insgesamt 15 Jurymitglieder unabhängig voneinander die Korrektheit der Antworten.

Enthielt der String die Antwort, war er als richtig zu bewerten. Da die Korrektheit der Antworten unter der Voraussetzung beurteilt werden sollte, dass der Benutzer dem System voll vertraue, brauchte der Antwortstring keine Begründung zu enthalten.

Enthielt der String die Antwort auf die Frage nicht, wurde er als falsch bewertet, auch wenn das Herkunftsdokument die Antwort beinhaltete.

Als falsch mussten auch Strings beurteilt werden, die zwar die korrekte Antwort enthielten, aber zudem auch verwirrende Zusatzinformationen, wie zum Beispiel Wörter von derselben semantischen Kategorie, welche die Antwort verundeutlichten. So wurde auf die Frage nach der Hauptstadt des Kosovo ein String, der den Ausschnitt „At Vucitrn, 20 miles northwest of

---

[12] NIST. Evaluation Report. S. 3.
[13] NIST. Track Report. S. 1.

Pristina," enthielt, als falsch bewertet, weil er neben der korrekten Antwort „Pristina" zusätzlich den Städtenamen „Vucitrn" beinhaltete, ohne deutlich zu machen, welcher der gesuchte sei.[15]

Falls ein Dokument falsche Fakten enthielt, das System diese Informationen jedoch korrekt extrahiert hatte, wurde der entsprechende String unter Berücksichtigung dieses falschen Kontexts als richtig beurteilt. Der Kontext wurde zudem bei Fragen in Betracht gezogen, deren Antworten vom Datum des Dokuments abhängen. Auf die Übungsfrage: „Who is the president of the United States" sollte „Bush" eine akzeptable Antwort sein, wenn der String aus einem Dokument aus der Bush-Aera, „Clinton", wenn der Bericht aus dessen Administrationsperiode stammte. Handelten die Dokumente von Nicht-Präsidenten gleichen Namens, galten die Antworten nicht als korrekt. So wurde der String „Bush" als Antwort auf die obige Frage aus einem Dokument über Pflanzen extrahiert und war somit inkorrekt. In solchen Fällen spielte der Kontext also die entscheidende Rolle.

Aber auch Antworten, die den Tatsachen entsprachen, aber nicht aus dem ihnen durch die Doc-Id zugeordneten Dokument stammen konnten, wurden als richtig bewertet, weil das Generieren von Antworten erlaubt war und dennoch eine Doc-Id als Referenz für den generierten Satz mitgeliefert werden musste.

Die fallabhängige Einbeziehung des Kontexts störte viele Jurymitglieder und führte zu einigen Fehlern. Solche Erörterungen gehen auch an den Bedürfnissen echter Benutzer vorbei, die in Fragen ohne Zeitangaben in der Regel an den Zuständen der Jetztzeit interessiert sind und sich im Zweifelsfall über die Korrektheit der Antwort im Referenzdokument vergewissern können wollen.

Welchen Genauigkeitsgrad eine Antwort haben musste, lag im Ermessen der Jurymitglieder. Je nach Frage konnten sie auch Teilantworten akzeptieren, also zum Beispiel nur den Nachnamen bei Fragen nach Personen. Bei Zahlen sollten die Einheiten mitgeliefert werden und die Dezimalpunkte durften nicht fehlen.

Durch den Freiraum bei der Beurteilung der nötigen Genauigkeit entstanden einige Unterschiede in der Bewertung. Die Organisatoren der TREC bezeichneten dies aber nicht als Makel der Evaluation, sondern als realistische Nachbildung des Verhaltens von Benutzern.[16]

---

[14] Falls nicht anders vermerkt, stammen die Informationen im diesem Unterkapitel aus: NIST. Evaluation Report. S. 5f. und S.22, Appendix B.

[15] Beispiel aus: NIST. Track Report. S. 4.

[16] NIST. Evaluation Report. S.8. Punkt 2.32.

2.4.2 Wertung

Der Rang der jeweils ersten richtigen Antwort wurde zur Bewertung der Systemleistung verwendet, indem die Antwort den Kehrwert des Ranges als Punktzahl erhielt. Hatte das System keine korrekte Antwort gefunden, erhielt es 0 Punkte. Die Kehrwerte der Ränge wurden aufaddiert und durch die maximal mögliche Antwortzahl geteilt.[17]

Nach Auffassung der Jurymitglieder kamen im Korpus für zwei Fragen keine Antworten vor. Deshalb wurden nur 198 der ursprünglich 200 ausgewählten Fragen bewertet.[18]

Systeme, die erkannten, dass sie keine Antwort finden konnten, wurden gleich behandelt wie solche, die nur falsche Antworten lieferten. Lieferten sie die Antwort nicht an erster Stelle, erhielten sie weniger Punkte. Dadurch stehen Systeme, die viele Antworten richtig beantwortet und hoch eingestuft hatten, auf den ersten Plätzen. Allerdings haben die meisten Systeme richtige Antworten, sofern sie überhaupt welche fanden, auf den ersten Plätzen untergebracht.[19]

3. Grundprobleme beim Q&A und Aufbau der Systeme

Eine Frage lässt sich am einfachsten beantworten, wenn man die Antwort darauf bereits kennt. Da dies beim Question-Answering in der Regel nicht der Fall ist, muss das System Hinweise bekommen, wo oder wonach es suchen soll. Dazu müssen aus der Frage, die gewissermassen die Suchanweisung darstellt, Informationen gewonnen werden, die das Lokalisieren der adäquaten Antwort erlauben.

Zweitens muss die riesige Datenmenge auf vielversprechende Texte bzw. Textstellen reduziert werden, um die aufwendige Detailanalyse effizienter zu machen.

Drittens kann diese verringerte Dokumentmenge dann auf die aus der Frage gewonnen Antwortindikatoren untersucht werden. Die ähnlichsten Textstellen werden auf die verlangte Länge gebracht und als Antwort zurückgeliefert.

Analog dazu lassen sich die meisten der verwendeten Q&A-Systeme in drei Grundbausteine gliedern. Nach der Art ihres Analysegegenstandes werden sie oft als Question-, Text-, oder Answer-Processing-Module bezeichnet. Die einzelnen Systeme gestalten die einzelnen Komponenten sehr unterschiedlich aus und geben ihnen auch unterschiedliches Gewicht. Im folgenden beschreibe ich Aufgaben und Strategien der einzelnen Module.

---

[17] NIST. Track Report. S. 1.
[18] Nr. 131 und 184. NIST. Evaluation Report. S. 3.
[19] NIST. Track Report. S. 1f.

4. Verarbeitung der Fragen

Wie oben erwähnt, muss die Question-Processing-Komponente Informationen verfügbar machen, die das Finden der gesuchten Antwort erlauben. Schlüsselwörter ermöglichen es anderen Komponenten, durch lexikalische Übereinstimmung ähnliche Kontexte zu finden. Tiefergehende semantische Ansätze versuchen auch die Abhängigkeiten zwischen Wörtern in geeignete Anfragen umzuwandeln oder den Typ der gesuchten Antwort zu ermitteln.

4.1 Schlüsselwörter

4.1.1 Grundannahmen und Basisset

Der einfachste Ansatz besteht darin, den Kontext der Frage mit Hilfe von Schlüsselwörtern, die aus der Frage gewonnen wurden, zu modellieren. Dahinter steckt die Vorstellung, dass die Konzepte aus der Frage im Bereich der Antwort ebenfalls vorkommen müssen. Die Frage wird getaggt, Stoppwörter, häufige Wörter, die inhaltlich nicht viel hergeben, werden entfernt und die übriggebliebenen Wörter werden als Keywords verwendet. Um sprachlichen Variationen besser gerecht zu werden, erweitern viele System dieses Basisset und greifen dabei häufig auf die Wortarten zurück. Keine Erweiterung vorgenommen haben zum Beispiel die SMU, AT&T und MultiText.[20]

4.1.2 Erweiterungen

Schlüsselwortsets können um morphologische und semantische Varianten oder sogar um bekannte Antworten erweitert werden.

In flektierenden Sprachen ändert sich eine Wortform je nach Wortart und syntaktischer Position. Einem Set von Schlüsselwörtern können entweder alle morphologischen Varianten hinzugefügt werden oder die Terme werden gestemmt und verlieren ihre Endungen.[21] Dadurch matchen auch Wörter aus derselben Wortfamilie.

Vollständiges Generieren von allen morphologischen Varianten einer Kategorie ist für eine relativ wenig flektierende Sprache wie das Englische machbar. Die Formen von Verben, Adjektiven und Nomen können mit einfachen Algorithmen regulär gebildet werden. Unregelmässige Formen werden entweder in Listen aufgeführt oder können in einem Thesaurus wie WordNet nachgeschlagen werden.[22]

---

[20] SMU, S. 4. AT&T, S. 2. MultiText/University of Waterloo/Carnegie Mellon University. Fast Automatic Passage Ranking (MultiText Experiments for TREC-8). G.V. Cormack, C.L.A. Clarke, D.I.E. Kisman,C.R. Palmer. S. 6.

[21] Z.B. CL-Research, S. 4. RMIT/Sharp Laboratories of Europe Ltd./CSIRO. The RMIT/CSIRO Ad Hoc, Q&A, Web, Interactive, and Speech Experiments at TREC 8. M. Fuller, M. Kaszkiel, S. Kimberley, J. Zobel, C. Ng, R. Wilkinson, M. Wu, S. 4. MITRE, S. 3.

[22] Z.B.: National Taiwan University. Description of Preliminary Results to TREC-8 QA Task. C-J Lin, H-H Chen. S. 2.

Durch die Erweiterung des Schlüsselwortsets um Synonyme versuchen die Systeme, die Eigenschaft der natürlichen Sprache, selbst eng verwandte Sachverhalte verschieden auszudrücken, nachzubilden. Diese Art der Erweiterung kann sich auf bestimmte Wortarten wie Nomen oder Verben beschränken und geschieht mit Hilfe von Thesauren wie WordNet oder Wortlisten.[23]

Einige der Fragen könnten besser durch Online-Enzyklopädien beantwortet werden als durch Zeitungstexte. Verschiedene Systeme haben Fragen nach Hauptstädten und ähnlichem aus Enzyklopädien extrahiert und der Anfrage als Schlüsselworte zugefügt. [24]

Wohl aus diesem Grund konnte ein System „Pristina“ als Hauptstadt des Kosovo extrahieren, obwohl im Kontext diese Tatsache nicht erwähnt wurde.

4.1.3 Extraktion nach syntaktischen Kriterien

QALC taggt die Wörter einer Frage. Durch Wortartkategorien vorgegebene syntaktische Muster werden zum Extrahieren der grösstmöglichen Konstituenten verwendet. Zusatzterme entstehen durch die Zerlegung dieser Phrasen in kleinere Einheiten, sofern diese nicht mit demselben Wort beginnen. Dadurch bleibt der Bezug der modifizierenden Elemente auf den Kern der Phrase erhalten. Aus dem String „name of the US helicopter pilot“ wird so zuerst „US helicopter pilot“. Diese Phrase wird anschliessend in „helicopter pilot“ und „pilot“ zerlegt.[25] Das macht Sinn, da keine Helikopter gesucht werden, sondern deren Piloten. Dieses Set von Schlüsseltermen wird in Grammatikregeln umgewandelt. Zu den einzelnen Wörtern werden alle in einer Datenbank enthaltenen Wörter mit dem gleichen Wurzelmorphem und die in WordNet aufgeführten Synonyme ermittelt. Die Grammatikregeln werden so verändert, dass sie auch mit den morphologischen und semantischen Varianten funktionieren. Eine solche Regel kann dann beim Suchen der Antwort sowohl „car maker“ als auch „making many automobiles“ als Formulierungen für dasselbe Konzept erkennen. Allerdings werden auch unpassende Phrasen wie „making cuts in auto“ als Varianten der obigen Sätze akzeptiert.[26]

---

[23] Ebenda.

[24] Z.B. University of Ottawa/ National Research Council. Ask Me Tomorrow: The NRC and University of Ottawa Question Answering System. J. Martin, C. Lankester.

[25] LIMSI-CNRS. QALC - the Question-Answering program of the Language and Cognition group at LIMSI-CNRS. O. Ferret, B. Grau, G. Illouz, C. Jacquemin, N. Masson. S. 4f.

[26] Ebenda, S. 5f.

4.1.4 Einsatz der Schlüsselwörter

Die Schlüsselwörter dienen als Input für die Text- und die Antwortverarbeitungskomponente. Mit ihrer Hilfe werden Kontexte modelliert, in denen eine richtige Antwort vermutet wird und Übereinstimmungen zwischen Anfrage und Dokument, bzw. Antwort, getestet. Die genaue Rolle und die benötigte Anzahl von Schlüsselwörtern hängt von den nachfolgenden Komponenten und ihren Methoden ab.

Hat das System eine eigene IR-Komponente, dienen die Schlüsselwörter als Selektionsbeschränkung. Besonders effizient wird die resultierende Textmenge von der SMU reguliert. Zu den Schlüsselwörtern werden keine Varianten gebildet, sondern nur die in der Frage vorkommenden Nicht-Stoppwörter verwendet. Diese werden anhand von acht gewichteten Regeln in Untergruppen eingeteilt. Zu Beginn werden die Schlüsselwörter aus den ersten sechs Gruppen an die Text-Processing-Komponente weitergegeben. Wenn diese Anfrage zu viele Textstücke zurückliefert, werden die beiden letzten Untergruppen der Anfrage als zusätzliche Beschränkung hinzugefügt. Liefert die erste Anfrage eine zu kleine Textmenge, werden nach und nach Schlüsselwörter entfernt. Die Schlüsselwörter aus den zuletzt angefügten Sets werden zuerst entfernt. Innerhalb eines Sets werden Schlüsselwörter, die vom Satzende stammen, zuerst entfernt. Die Sets werden, sofern ausreichend Schlüsselwörter vorhanden sind, angepasst, bis eine Anfrage eine Textmenge im vorgegebenen Rahmen liefert.[27]

Die Anzahl benötigter Schlüsselwörter kann auch davon abhängen, ob die Frage eine erkennbare Named Entity als Typ verlangt. Cymfony sucht normalerweise Synonyme zu den in der Frage vorkommenden Verben und bildet als morphologische Varianten die Vergangenheitsformen dazu. Wenn der Fragetyp nicht auf eine Named Entity abbildbar ist, werden zusätzliche Synonyme zum Bezeichner des Fragetyps gebildet. Dem Keywordset einer mit „why“ beginnenden Frage werden als zusätzliche Keywords z.B. noch because, because of, due to, for the reason etc. hinzugefügt.[28] Wenn keine Named Entity gesucht werden kann, soll diese fehlende Information durch mehr kontextstiftende Wörter ausgeglichen werden. Der Erfolg dieser Methode wurde leider nicht ausgewertet. Wenn keine Antworttypen verwendet bzw. gefunden werden, werden im Allgemeinen die Stellen mit den grössten Schlüsselworthäufungen extrahiert.[29]

---

[27] SMU, S. 2-4.
[28] Cymfony, S. 8.
[29] Der Track Report von NIST, S. 4 beschreibt dies als generelle Strategie vieler Systeme.

4.2 Semantische Ansätze

4.2.1 Bestimmung von Fragetypen

Häufig gibt eine Frage nicht nur Hinweise auf den Kontext, in dem sie zu finden ist, sondern bestimmt auch den semantischen Typ der Antwort. Einige Fragewörter lassen sich immer demselben Typ von Antwort zuordnen. So verlangen When- und Where-Fragen Zeit- bzw. Ortsangaben. Bei der Beantwortung solcher Fragen kann explizit nach einer Phrase vom entsprechenden Typ gesucht werden. Zum Teil werden auch fixe Abbildungsmuster für komplexe Fragewörter wie „how old“ oder „how far“ festgelegt.[30]

Viele Fragewörter sind für sich allein genommen aber hochgradig ambig. Eine What-Frage kann nach nahezu jedem Objekt oder Ereignis fragen. In der Regel gibt es in der Frage jedoch Phrasen, welche sie für einen menschlichen Leser eindeutig machen. Die Antwort auf die Frage „What is the largest city in Germany?“ muss der Name einer Stadt sein. Die gesuchte Antwort soll also vom selben semantischen Typ sein wie „city“.[31]

Um solche desambiguierenden Phrasen zu finden, werden meistens syntaxbasierte Regeln mit nicht instanzierten Variablen vom Typ : What NP oder what AUX NP verwendet. Solche Muster werden für möglichst viele ambige Fragewörter erstellt. Die Kerne der so extrahierten NPs werden durch Nachschlagen in Listen semantischen Kategorien zugewiesen.[32] Wird kein solches Muster erkannt, oder hat die Frage kein Fragewort, so verwendet beispielsweise der NE-Ansatz von AT&T den Typ der ersten NP als Antworttyp.[33] Qalc sucht Strukturen wie die oben beschriebenen auch im Inneren von Sätzen.[34]

Die verwendeten syntaktischen Informationen werden meist von flachen Parsern geliefert. Im Unterschied dazu generiert die University of Maryland ganze Dependenzbäume. Der Regent des Fragewortes erhält anhand einer handgemachten Liste einen semantischen Typ zugewiesen.[35]

Trotz dieser Bemühungen ist es nicht immer möglich, einer Frage einen eindeutigen Antworttyp zuzuweisen. Die Ambiguität bleibt entweder erhalten, oder wird durch unterschiedliche Gewichtung auf eine bestimmte Lesart hin beeinflusst. In beiden Fällen entscheidet der Kontext darüber, welcher Typ schliesslich in der Antwort vorkommt.[36]

---

[30] Cymfony, S. 8.
[31] Beispiel von SMU, S. 2.
[32] Z.B. LIMSI-CNRS, S. 3 und Cymfony, S. 6f. Xerox arbeitet mit Wortartmustern. Xerox Research Centre Europe. Xerox TREC-8 Question Answering Track Report. D. Hull, S. 2.
[33] AT&T, S. 4.
[34] LIMSI-CNRS, S. 3.
[35] University of Maryland, College Park, MD/ University of Manitoba/ University of Maryland, Baltimore County, MD. TREC-8 Experiments at Maryland: CLIR, QA and Routing. D.W. Oard, J. Wang, D. Lin, I. Soboroff. S. 6.
[36] Beispiele bei IBM, S. 3, AT&T, S. 5, Cymfony, S. 8. Xerox, S. 4-6.

Antworttypen können auch in hierarchische Beziehung zueinander gesetzt werden. Falls in einem besonders guten Kontext keiner der gesuchten Typen vorkommt, kann so nachgeprüft werden, ob ein allgemeinerer Typ vorhanden ist.[37]

Die New Mexico State University hat Templates erstellt, die neben dem Typ der Antwort auch Angaben über NEs und ausgesuchte Wörter oder Wortkombinationen enthielten, die in der Nähe der Antwort vorkommen sollten. Dazu wurden die Fragen mit fixen Mustern verglichen. Fragevarianten wurden auf dasselbe Muster abgebildet.[38]

Die verschiedenen Systeme, die semantische Kategorien einsetzten, unterscheiden sich vor allem in der Anzahl vergebener Kategorien. Je feinere Kategorien vergeben und erkannt werden können, desto genauer sind die Resultate. Sammelkategorien schneiden schlechter ab. Oft fehlte die Bestimmung häufiger und eigentlich banaler Kategorien wie Währungen.[39]

4.2.2 Umwandlung syntaktischer Strukturen

Nur zwei der teilnehmenden Systeme haben versucht, aus den Fragen und Texten Wissensrepräsentationen zu generieren, welche die durch die syntaktischen Strukturen vorgegebenen Beziehungen zwischen den einzelnen Wörtern in den Termen der Wissensbasis bewahren.

CL-Research parste die Fragen vollständig und erzeugte durch die Analyse des Syntax-Baumes dreiteilige Terme, bestehend aus einem Wort oder einer Phrase mit bestimmten kategoriellen Eigenschaften, einer semantischen Rolle, die das Wort oder die Phrase im Satz einnimmt (vergleichbar mit thematischen Rollen) und dem Bezugswort der ersten Phrase. Fragen und Sätze aus den Dokumenten wurden gleich behandelt. Pro Frage enthält eine der Formeln eine ungebundene, mit dem gesuchten Typ versehene Variabel.[40]

Die University of Sheffield hat eine spezielle Grammatik geschrieben, die syntaktische Strukturen und Informationen zu Named Entities in Quasi-Logische-Formeln (QLF) umsetzt. Anstelle des Fragewortes wird eine mit „qvar“ markierte Variabel verwendet.[41]

---

[37] MITRE, S. 3.
[38] New Mexico State University, S. 6f. Dieser Ansatz war nicht komplett implementiert und wird deshalb im Folgenden ausser Acht gelassen.
[39] Z.B. bei AT&T. S. 6. Sie vergaben zwar ein Tag für Währungen, in den Texten wurden aber keine Entsprechungen gesucht.
[40] CL-Research, S. 3.
[41] University of Sheffield. University of Sheffield TREC-8 Q & A System. K. Humphreys, R. Gaizauskas, M. Hepple, M. Sanderson. S. 4.

5. Verarbeitung der Dokumente

Das Text-Processing hat die Aufgabe, die Dokumente so vorzuverarbeiten, dass anschliessend die Extraktion der Antwort möglich wird. Dazu gehört das Sichtbarmachen von linguistischem Wissen wie semantischen Typen von Nominalphrasen, Erkennung von Named Entities, das Auflösen von Koreferenzen und das Finden besonders treffender Dokumente oder Textstellen.

5.1 Anreicherung der Texte mit linguistischem Wissen

5.1.1 Erkennung und Markierung semantischer Typen

Die semantischen Typen, die in den Dokumenten gesucht werden, entsprechen der Typologie des jeweiligen Question-Processing-Moduls. Die Typen beziehen sich auf bestimmte Kategorien von Named Entities. Named Entities sind nach der Definition der MUC 6 Named Entity Recognition Task[42] „unique identifiers of entities“, [43] also Namen für individuelle Einheiten und keine Gattungsbezeichnungen. Das heisst, dass der String „Disneyland“ als NE bestimmt werden müsste, der String „Vergnügungspark“ hingegen nicht. Neben Personen, Organisationen und Orten zählt die MUC-6 auch Daten, Zeitangaben und Mengen wie mit Währungen oder Prozentzeichen kombinierte Zahlangaben zu den Named Entities.[44]

Die in TREC-8 verwendeten Verfahren zum Ermitteln semantischer Typen analysierten zum Teil wesentlich mehr Kategorien und legten in der Regel mehr Gewicht darauf, einer Wortgruppe das richtige semantische Tag zu verleihen, als sicherzustellen, dass es sich wirklich um einen Namen und nicht um eine Gattungsbezeichnung handelte. Deshalb setze ich in dieser Arbeit eine Named Entity einer semantischen Kategorie mit einer NP derselben semantischen Kategorie gleich.

Die Strategien, die angewendet werden, um Named Entities zu finden, reichen von ausgeklügelten NE-Parsern über Nachschlagen in Wortlisten bis zu simplen Heuristiken die etwa alle auf „in“ folgenden Nominalphrasen als Ortsbezeichnungen taggen.[45] Der Hauptunterschied besteht in der Art und Anzahl erkannter Kategorien und der Genauigkeit der Erkennung.

Die semantischen Typen können vor dem Information-Retrieval oder erst für die reduzierte Textmenge bestimmt werden. Die vorgängige Bestimmung erlaubt es, die verlangten Typen in die Anfrage einzubeziehen und somit bei der Selektion der Dokumente zu berücksichtigen. Die

---

[42] MUC 6 Task Definition. http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html. Stand vom 06.07.01.
[43] Named Entiy Recognition Task. http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_2.html. Stand vom 06.07.01.
[44] Ebenda.
[45] Parser: Cymfony, S. 5. Heuristiken: CL-Reasearch. S. 4.

nachträgliche Bestimmung ist weniger zeitaufwändig.[46] In zweistufigen Systemen können semantische Typen z.B. als zusätzliche Beschränkung zur Einengung auf Paragraphen verwendet werden.

5.1.2 Auflösen von Koreferenzen

Sätze in Texten stehen in der Regel nicht für sich allein, sondern sind durch anaphorische und kataphorische Bezüge miteinander verbunden. Konzepte aus vorhergehenden Sätzen werden wiederholt, ohne dass dasselbe Wortmaterial verwendet wird. Einmal erwähnte Nominalphrasen werden häufig nicht materiell, sondern durch Referenz von Pronomen und Teil-NPs wiederaufgenommen. Deshalb kann die Stelle mit dem bestmöglichen Kontext für die Antwort den Antworttyp scheinbar nicht enthalten, obwohl eigentlich ein Pronomen auf den gesuchten Typ referiert.

Die Frage „Who built the Hancock Building“ macht am besten mit dem Satz „He designed the Hancock Building.“ Das Pronomen „he“ kann nur durch Auflösen von Koreferenzen mit den Sätzen „I. M. Pei is a well known architect.“ und „Pei studied at MIT.“ in Verbindung gebracht werden. Das System Qanda ergänzt den besten Kontext mit der längsten vorkommenden Form der referenzierten Nominalphrase zu „I. M. Pei: He designed the Hancock Building“[47].

Die *University of Pennsylvania* teilt die Dokumente in Abschnitte und Sätze auf, führt Wortarttagging durch und ermittelt anschliessend Nominalphrasen und Named Entities. Für die Koreferenzauflösung in Betracht gezogen werden Eigennamen, definite Nominalphrasen oder Personalpronomen in der dritten Person. Eigennamen und Nominalphrasen gelten als mit einer NP koreferent, wenn bestimmte Teile mit dieser übereinstimmen. Bei Pronomen wird die Wahrscheinlichkeit anaphorischer Koreferenz für die 20 vorausgehenden NPs untersucht. Einbezogen werden unter anderem die Zahl von Nominalphrasen und Sätzen die zwischen der NP und dem Personalpronomen stehen, die Position der NP im Satz und die Kongruenz zwischen Nomen und Pronomen. Koreferente Nominalphrasen vererben alle ihre Bestandteile an die anderen koreferierenden Phrasen, was beim Informationretrieval höhere Punktzahlen ergibt und bei der Detailanalyse den Ersatz von Pronomen erlaubt.[48]

Die University of Sheffield löst Koreferenzen zwischen Nominalphrasen auf, indem sie ausser der grammatischen Kompatibilität auch die semantische überprüft. Die Kerne der in Frage kommenden Nominalphrasen werden anhand der Ontologie des Systems semantischen Typen

---

[46] IBM annotiert den gesamten Korpus im Voraus. IBM T.J. Watson Research Center/Columbia University. The Use of Predictive Annotation for Question Answering in TREC8. J. Prager, D. Radev, E. Brown, A. Coden,V. Samn. S. 1. Als Sekundärkriterium werden die NEs z.B. von NTT Data eingesetzt. NTT Data Corporation. NTT DATA: Overview of system approach at TREC-8 ad-hoc and question answering. T. Takaki. S. 6.

[47] MITRE, S. 3f.

[48] University of Pennsylvania. Using Coreference in Question Answering. T. Morton, S. 1f.

zugeordnet. Überprüft wird, ob sich die Typen in der Hierarchie der Ontologie ausschliessen, und durch wie viele Hierarchiestufen sie von einander getrennt werden. Zusätzlich berücksichtigt wird die Anzahl übereinstimmender Attribute.[49] Ebenfalls Koreferenzen aufgelöst hat die University of Maryland.[50]

5.1.3 Logische Formen und verwandte Ansätze

CL-Research und die University of Sheffield erstellten für die ausgewählten Texte Wissensbasen aus QLFs bzw. Termen, welche die Relationen eines Satzes speicherten. Diese wurden auf dieselbe Weise erstellt wie die entsprechenden Repräsentationen der Fragen, enthielten aber keine Variablen. Für die Details verweise ich deshalb auf den Abschnitt 4.2.2 Umwandlung Syntaktischer Strukturen auf Seite 10.

5.2 Eingrenzen von Textstellen: IR-verfahren

5.2.1 Eignung für Q&A

Traditionelles Information-Retrieval beurteilt die thematische Relevanz eines Dokuments, indem die Häufigkeit von bestimmten Schlüsselwörtern ermittelt wird. Da einzelne Terme als Anfrage verwendet werden, werden homographe Wörter mit anderer Bedeutung nicht ausgeschlossen. Abhängigkeitsbeziehungen zwischen den einzelnen Anfragetermen, die einen Teil der Semantik ausmachen, gehen verloren.[51] Die Dichte der Schlüsselwörter wird meist nicht beachtet. Sie können über den gesamten Text verteilt sein, so dass sie oft gar nichts miteinander zu tun haben. Andererseits besteht die Möglichkeit, dass Dokumente, die in einem kurzen, aber entscheidenden Abschnitt die Antwort enthalten, schlecht eingestuft werden, weil sie insgesamt wenig Punkte erhalten.

Trotz dieser Nachteile ist stichwortbasiertes Information-Retrieval im Gegensatz zu NLP-Methoden schnell und effizient. Deshalb filtern alle verwendeten Systeme den riesigen Textkorpus von TREC-8 mit irgendeiner Art IR. Die besten Resultate wurden mit Passageretrieval ereicht, weshalb dieser Ansatz unten genauer beschrieben wird.

5.1.2 Allgemeines Vorgehen

Im einfachsten Fall wird die Frage direkt an ein IR-System weitergeleitet, welches sie als Query benutzt.[52] Andernfalls erhält das System ein speziell erstelltes Schlüsselwortset als Anfrage.

---

[49] University of Sheffield. S. 5f.
[50] University of Maryland, S. 6.
[51] Eine Ausnahme ist Qualc von LIMSI. Das System taggt Terme, die den Konstituenten einer Phrase entsprechen.
[52] AT&T S. 2, MulitText S. 6, SMU, S 4. Einige Hersteller hatten zudem keine eigene IR-Komponente. Ihnen stellte AT&T pro Frage die 200 Top Resultate ihres IR-Systems zur Verfügung.

Das Kernstück vieler IR-Systeme ist die Indexierung. Die Systeme unterscheiden sich dadurch, ob sie nur einzelne Wörter, zusätzlich Varianten dazu oder auch Termkompositionen indexieren. Eigennamen und Mehrwortterme erhalten in der Regel höhere Punktzahlen.[53]

IBM hat zusätzlich zu den üblichen Termen auch semantische Typen indexiert und verwendet diese bereits in der Anfrage. Zuerst müssen die Schlüsselwörter matchen, damit nicht ein Fragewort die Stelle der Antwort besetzen kann, dann werden Passagen mit den am besten passenden Fragetypen ausgewählt[54].

Unterschiede bestehen vor allem in der Anzahl der verwendeten Top-Dokumente. Die Fehleranalyse vieler Hersteller ergab, dass relevante Dokumente oft nicht innerhalb der Top 10 zu finden waren und die Anzahl verwendeter Dokumente die Leistung des Systems beeinflusst. Um mehr relevante Stellen mit erträglichem Zeitaufwand analysieren zu können, wird häufig Passageretrieval eingesetzt.

5.2.3 Passageretrieval

Passageretrieval liefert kurze Textpassagen statt langen Dokumenten. Die Dichte der Frageterme ist dabei eines der wichtigsten Kriterien. Die Terme sollen möglichst nahe beieinander, in einer möglichst kurzen Sequenz vorkommen. Die Häufigkeit eines einzelnen Terms in dieser Sequenz oder im gesamten Dokument spielt hingegen keine Rolle.[55] Zudem erhalten Textstellen, die einige oder alle Schlüsselwörter in der durch die Frage vorgegebenen Reihenfolge enthalten Bonuspunkte.[56] Hinter beidem steckt die Vorstellung, das gedankliche Konzepte in Frage und Antwort gleich formuliert sein sollten. Übereinstimmende Abfolge kann ein Zeichen dafür sein, dass die syntaktische Struktur zwischen den Wörtern, und damit evtl. auch die semantische Abhängigkeit, dieselbe ist.

Ein zweistufiges IR-Verfahren gewählt haben z.B. AT&T, NTT Data, Scai und die SMU.[57]

Bei diesem Ansatz wird das Textkorpus zuerst mit einem Dokument-Retrieval-System reduziert. Dabei ist möglichst hoher Recall gewünscht. Ein Passagenretrievalsystem filtert anschliessend Absätze heraus, die eine möglichst hohe Schlüsselwortdichte aufweisen. Die Dokumente werden dazu in Sätze und Paragraphen aufgeteilt. Für jede Texteinheit, d.h. für einen Satz, einen Paragraphen oder ein durch eine Anzahl Bytes definiertes Fenster wird eine Punktzahl errechnet. Bewertet wird die Anzahl verschiedener Schlüsselwörter in der Texteinheit, gegebenenfalls, ob es sich um Varianten oder ursprüngliche Terme handelt, der Abstand zwischen den

---

[53] Ein Beispiel dafür ist LIMSI-CNRS. S. 5f.
[54] IBM S. 3f.
[55] Ebenda, S. 1.
[56] Z.B. SMU, S. 5.
[57] AT&T, S. 2f. SMU, S. 4f. Seoul National University. SCAI TREC-8 Experiments. D-H Shin, Y-H Kim, S. Kim, J-H Eom, H-J Shin, B-T Zhang, S. 6. Beispiele für den folgenden Absatz finden sich u.a. bei diesen Herstellern.

verschiedenen Schlüsselwörtern und die Übereinstimmung der Reihenfolge der Schlüsselwörter mit der ursprünglichen Fragereihenfolge. Bei der Berechnung der Punktzahlen werden die Schlüsselwörter oft gewichtet. In der Regel erhalten Eigennamen und Mehrwortterme mehr Gewicht als Nomen.[58]

Ein Fenster kann auch durch die kleinst mögliche Textmenge definiert werden, in der alle oder mindestens *n* Schlüsselworte vorkommen. In diesem Fall werden für die Bewertung auch die Grösse des Fensters und die Anzahl fehlender Terme in Betracht gezogen.[59]

Schlüsselwörter können auch als Zentrum einer Passage eingesetzt werden. Je weiter neue Anfrageterme von diesem Zentrum entfernt sind, desto weniger tragen sie zur Punktzahl dieser Passage bei. Mit einem ähnlichen Ansatz ist AT&T in der 250-er Kategorie mit 135 richtig beantworteten Fragen auf den 2. Platz gekommen. In der 50 Byte Kategorie reichte es mit demselben Ansatz nur für 77 richtige Antworten und Platz 10.[60]

Die Passagen werden nach ihren Punktzahlen geordnet und nach den Gegebenheiten der einzelnen System weiterverarbeitet.

6. Verarbeitung der Antwort

Die letzte Komponente hat im Allgemeinen den Auftrag, geordnete und eventuell linguistisch angereicherte Texte oder Textstellen mit den aus der Frage gewonnenen Informationen zu vergleichen und eventuell neu zu ordnen, die am besten passenden Stellen auf die richtige Länge zu bringen und fünf nach ihrer Relevanz geordnete Strings als Antwort auf die Frage zu übermitteln. Auf welche Art Frage und Textstellen miteinander verglichen werden, hängt von der Arbeitsweise der vorhergehenden Komponenten ab.

Anzahl und Reihenfolge von Schlüsselwörtern werden von den meisten Systemen auf Satzebene ermittelt, aber auch die anderen im Abschnitt Passageretrieval erwähnten Fenstervarianten kommen vor. Die Terme werden je nach System gewichtet und daraus wird die Relevanz der Textstelle, eventuell in Abhängigkeit zur NE-Bestimmung, errechnet. Passageretrieval-Systeme nehmen die Gewichtung einzelner Sätze meist vorweg.

Verlangt die Frage einen bestimmten Antworttyp, tragen die entsprechenden Named Entities entweder wie Schlüsselwörter zur Punktzahl des Fensters bei oder eliminieren solche, die keine NE vom verlangten Typ enthalten. Wird die NE, welche die Frage im Wesentlichen beantwortet, in den Kandidatsätzen erst bestimmt, wenn die Schlüsselwörter bereits gematcht wurden, verhindert man, dass als mögliche Antworten Wortgruppen angesehen werden, die aus

---

[58] Ausnahmen sind z.B. RMIT/CSIRO, S. 4 und Qanda. MITRE, S. 3.
[59] Z.B. SMU, S. 5.
[60] At&T, S. 2 und Track Report S. 5.

der Frage stammen und vom verlangten Typ sind.[61] Dasselbe Ziel kann erreicht werden, wenn man die Textstellen, die als Antwortkandidaten ermittelt wurden, auf Frageterme filtert. Der Nachteil der letzteren Methode ist, dass Frageterme als Substrings durchaus in der Frage erwünscht sein können oder dass die Sätze durch ihre Entfernung ungrammatisch werden wie bei Xerox.[62]

Werden logische Formen verwendet, werden beim Matching zusätzlich zu der Übereinstimmung von Typen und Schlüsselwörtern auch syntaktische Abhängigkeiten berücksichtigt.[63] Ähnlich vorgegangen ist die University of Maryland. Stimmen mehrere Wörter in einem Satz mit Fragetermen überein, die zur selben Konstituente gehören, wird überprüft, ob das auch in der Textstelle der Fall ist. Andernfalls wird der Satz aus der Liste der Antwortkandidaten gestrichen.[64]

Nach diesen Auswahlverfahren werden die Textstellen eventuell neu geordnet. Falsche Einordnung von Antwortkandidaten kann viele Fehler verursachen. Qanda und CL-Research haben in 18 beziehungsweise 23 Fällen zwar einen Satz mit der richtigen Antwort gefunden, ihn aber ausserhalb der ersten fünf eingeordnet.[65]

Die geordneten Textstellen müssen an die Antwortformate angepasst werden. Die Mehrheit der Systeme hat sich auf Trunkierung beziehungsweise Expansion beschränkt.

AT&T reduziert in den Passage-Retrieval-Ansätzen den Satz mit der höchsten Punktzahl bis auf die Antwortgrösse. Bleiben zu viele Bytes übrig, werden Funktionswörter und in der Frage vorkommende Terme ausgelassen. Ist dies zu wenig, wird die Anzahl erlaubter Bytes vom Anfang des Satzes her ausgegeben.[66] Dadurch kann die Lösung weggeschnitten werden.[67] Bleiben Bytes übrig, werden die benachbarten Sätze, Wörter und Wortteile bis zum Erreichen der Limite angefügt.[68]

Generiert wurden Antworten hauptsächlich im Zusammenhang mit Koreferenzauflösung. Die besten Kontexte werden mit den informativsten, das heisst mit den längsten NEs kombiniert.[69]

---

[61] IBM, S. 3.
[62] Xerox. S. 4-6.
[63] CL-Reserach, S. und University of Sheffield, S. 6f.
[64] University of Maryland, S. 5.
[65] CL-Reasearch, S. 6. MITRE, S. 4f.
[66] AT&T, S. 5f.
[67] LIMSI-CNRS, S. 9. Geschieht 3 mal bei CL-Research. S. 6.
[68] Z.B. AT&T, S.5f.
[69] MITRE S. 3 für Qanda und University of Sheffield, S.6.

7. Bewertung einzelner Strategien

Nach dem die verwendeten Strategien vorgestellt wurden, soll ihre Tauglichkeit für Q&A festgestellt werden. Die Leistung einzelner Komponenten wurde nicht evaluiert. Man kann aber die Bewertung der Gesamtleistung durch TREC-8 und die Anzahl korrekt beantworteter Fragen, sowie die Fehleranalysen der einzelnen Hersteller heranziehen.

Fraglich ist, ob die Erkenntnisse von TREC-8 problemlos auf andere Question-Answering Projekte übertragbar sind. Der folgende Abschnitt befasst sich deshalb mit Schwachstellen in der Fragestellung der TREC-8, damit auf diesem Hintergrund lohnenswerte Entwicklungen aufgezeigt werden können.

7.1 Schwachstellen der TREC-8-Fragestellung

Die Fragestellung klammert einige Probleme aus, die ein System im Kontakt mit realen Benutzern bewältigen können sollte.

Die zur Extraktion benötigte Zeit wurde nicht bewertet. Es bestand kein Anreiz, Abläufe zeitlich zu optimieren, was aber vielleicht ein Anreiz war, zeitaufwändigere Methoden zu testen.

Ein System, das mehrere richtige Antworten fand, wurde nicht belohnt. Auch Systeme, die den Benutzer vor höchstwahrscheinlich falschen Antworten verschonen, werden durch die Aufgabestellung nicht gefördert, da die Organisatoren eine solche Einsicht nicht mit Punkten belohnten und sicher stellten, dass sich die Antwort im Textkorpus befindet. Aus demselben Grund bestanden die in TREC-8 verwendeten Fragen hauptsächlich aus z.T. sehr deutlichen Umformulierungen von Textstellen. Damit garantierten die Organisatoren den Teilnehmern indirekt eine im Vergleich mit realen Benutzerfragen hohe Zahl von zutreffenden Schlüsselwörtern. Dadurch steigt auch der Nutzen von Heuristiken, welche die Ähnlichkeit zwischen Anfrage und Zielsatz testen. Einen alternativen Ansatz dazu hat nur LIMSI-CNRS mit Qalc angeboten.[70]

Die Beschränkung auf kurze faktenbasierte Fragen scheint den gegenwärtigen Leistungsmöglichkeiten von Q&A-Systemen angemessen zu sein, führt aber dazu, dass Systeme mit NP oder NE-Erkennung klar bevorzugt werden. Die Verteilung der Fragetypen wirkte ebenfalls in diese Richtung.[71] Typen die keine NP oder eine NE verlangen, waren schwach vertreten. Erkenntnisse bezüglich des Nutzens von NP und NE-Erkennung können sich also nur auf die beschränkte Fragestellung der TREC beziehen.

Weil pro Antwort eine Dokumentnummer als Referenz angegeben werden muss, selbst wenn sie aus einem Lexikon stammte, enthielten die Strings häufig eine an sich richtige Antwort, die aber

---

[70] Vergleiche Abschnitt 4.1.3.

im falschen Kontext stand oder zumindest in diesem nicht verifiziert werden konnte. Dies war allerdings auch nicht verlangt.

Zu untersuchen wäre, ob die Gattung „Newstexte“ Eigenschaften hat, die gewisse Strategien überdurchschnittlich gut abschneiden lassen.

7.2 Beurteilung einzelner Komponenten

Die Beurteilung einzelner Komponenten gestaltet sich schwierig, da die Systeme als Ganze und nicht für einzelne Bereiche evaluiert wurden. Hier sollen nur möglichst vielversprechende Teilansätze behandelt werden, deren Leistung allerdings von diversen vor- und nachverarbeitenden Modulen beeinflusst wird. Gut aufeinander abgestimmte Komponenten verbessern die Resultate, wie das Beispiel der SMU zeigt. Andererseits können fehlerhaftes Wortarttagging und Parsen, Fehler bei der Aufteilung in Sätze und Abschnitte, vollständiges Entfernen von Satzzeichen und ungeschickte Trunkiermethoden die Leistung anderer Komponenten beeinträchtigen. In den folgenden Abschnitten sollen die eben erwähnten Methoden und Probleme nur am Rande angesprochen werden. Das Hauptziel soll sein, umfassendere Strategien auf ihre Tauglichkeit zu beurteilen.

7.2.1 IR-Komponente

Die IR-Komponente ist ein wichtiger Engpass, da sie die zu verarbeiteten Texte zwar reduzieren soll, zum Teil aber alle relevanten Dokumente wegfiltert. Viele Systeme haben nur etwa die 5 bzw. 10 bestklassierten Dokumente[72] weiterverarbeitet, weil sie die Genauigkeit von IR-Methoden gerade für Question-Answering überschätzten. Dokument-Retrieval berücksichtigt häufig nicht, dass auch kurze Textstellen eine Antwort enthalten können, ohne das der Score des Gesamtdokuments hoch ist.

Da einige Systeme kein eigenes Dokumentretrievalsystem besassen und deshalb die 200 von AT&T bestklassierten Dokumente nutzten, liegen Vergleichszahlen vor. Qanda und CL-Research benutzen nur die besten 10 Dokumente, die laut ihren Angaben für 24% bzw. 33% der Fragen keine Antwort enthielten.[73] Die Passageretrievalkomponente von AT&T sichtete die besten 50 Dokumente und lieferte nur für 16% der Fragen keinen relevanten Paragraphen.[74] Die 50 besten Dokumente müssen also deutlich mehr Antworten enthalten haben, da die Resultate von AT&T trotz zweiter Einschränkung deutlich besser sind.

---

[71] Vergleiche Abschnitt 2.2.1.

[72] Die Ersten 5 Dokumente hat z.B. New Mexico State University, S. 7 verwendet, die ersten 10 Qanda von MITRE, S. 2. und CL-Research, S. 7. Die University of Sheffield hat je zweimal die ersten 5 der University of Massachusetts und die ersten 10 Dokumente von AT&T verwendet. S. 1f.

[73] MITRE, S. 5 und CL-Research, S. 6.

[74] AT&T, S. 7.

7.2.2 Bestimmung von semantischen Kategorien und NE-Erkennung

Der Nutzen von Named Entities für kürzeste Antworten wiederspiegelt sich in der Systemrangliste in der 50-Byte-Kategorie. Die Systeme auf den ersten neun Plätzen haben alle in irgendeiner Form Fragetypen auf Named Entities abgebildet.[75] Cymfony, das System mit der differenziertesten Typenkategorie, belegt dabei den ersten Platz. Die Fehleranalyse für den viertplazierten NE-Ansatz von AT&T ergibt, das 18% der Fragen fehlerhaft beantwortet wurden, weil das System keine entsprechenden Named Entities bestimmen konnte. Mehr als die Hälfte dieser Fehler wurden durch das Fehlen einer Kategorie für Währungen verursacht. Weitere 8% der Fragen verursachten Probleme, weil der Fragetyp falsch analysiert wurde oder weil er keine Named Entity verlangt.[76] Das bedeutet einerseits, dass Fragen, die durch Named Entities beantwortet werden können, leichter zu beantworten sind und zweitens, dass eine gute Abdeckung möglicher Typen deutliche Leistungsverbesserungen hervorbringen kann. Korrekte Erkennung von Named Entities kann auch Parser unterstützen, da in NEs häufig unbekannte Wörter und schwer analysierbare Strukturen vorkommen, die den Parser belasten.

In beiden Kategorien stammt das bestplatzierte Passagenretrievalsystem von AT&T. Liegt es in der 50 Byte Kategorie erst auf dem neunten Platz, belegt es in der 250-er Kategorie den zweiten und beantwortet 9 Fragen mehr als der auf demselben Passageretrievalsystem basierende NE-Ansatz. Auch LIMSI-CNRS stellt in dieser Kategorie keine Leistungsverbesserung durch Verwendung von Named Entities fest. LIMSI vergab nur die vier Kategorien Zeit, Person, Ort und Zahl und die Named Entities wurden im Unterschied zu anderen Systemen schwächer gewichtet als die Schlüsselwörter.[77] Laut der University of Iowa können Leistungssteigerungen durch höhere Gewichtung von Named Entities erreicht werden.[78] Trotzdem scheint effizientes Passageretrieval gute Dienste zu leisten, solange grössere Outputs erlaubt sind. Das bedeutet, dass die Verwendung von Named Entities einzugrenzen hilft, was zur Antwort gehören muss. Solange eine gewisse Menge an zusätzlichem, nicht zwingend informativem Text erlaubt ist, steigern sie die Leistung eines Systems nicht wesentlich. Sollen in Zukunft Sätze losgelöst von Textstrukturen generiert werden, kann man auf solche Informationen aber sicher nicht verzichten.

---

[75] NIST, Track Report, S. 3. Alle folgenden Informationen zu Platzierungen und der Anzahl korrekt beantworteter Fragen, beziehen sich auf diese Stelle.
[76] AT&T, S. 7f.
[77] LIMSI CNRS, S. 9.
[78] University of Iowa, S. 5.

7.2.3 Auflösen von Koreferenzen

Koreferenzen wurden nur von wenigen Systemen aufgelöst. Ge/Penn, drittplaziert in der 250-Byte-Kategorie ist eines davon. Es hat aus den besten Dokumenten je ein um den am besten Absatz platziertes Fenster von 20 KB verwendet und dem Fragetyp entsprechende Named Entities gesucht.[79] Der direkte Nutzen der Koreferenzauflösung lässt sich aus dem Paper dieser Organisation also nicht ermitteln. Bei der Fehleranalyse haben Qanda und CL-Research festgehalten, wie oft bei ihren Resultaten Koreferenzauflösung zur Beantwortung einer Frage nötig gewesen wäre. Mit einem solchen Verfahren hätte Qanda in 9 Fällen aus dem am höchsten eingestuften Material die richtige Antwort generieren können, in drei weiteren Fällen wäre es nötig gewesen, Zusammenhänge über Satzgrenzen hinaus zu erkennen.[80] CL-Research nimmt sogar an, dass Koreferenzauflösung das Resultat für 17 Fragen verbessert hätte. Zudem verlangten laut dieser Analyse ein Fünftel der Fragen die Auflösung relativer Zeitangaben.[81]

7.2.4 Logische Formen und syntaktische Beschränkungen

Die beiden Systeme, die logische Formen bzw. Relationen speichernde Terme verwendeten, erzielten keine herausragenden Resultate. Die University of Sheffield wurde in beiden Kategorien dritt- und viertletzte. In der 50-Byte-Kategorie konnte sie 16 bzw. 14 Fragen richtig beantworten, in der anderen 19 bzw. 22. CL-Research nahm nur in der 250er teil und erreichte von 24 möglichen den 17. Rang. Dieses System hatte 83 Antworten finden können.[82] Man kann vermuten, dass die Leistung des Systems steigen würde, wenn man die Auflösung von Koreferenzen und Daten sowie die relativ schwache, nur musterbasierte NE-Erkennung verbessern würde.

Beide System haben zu wenige Dokumente verwendet.[83] Die schlechte Leistung der Sheffield University wird dadurch verstärkt, dass nur die erste mögliche Antwort und keine weiteren erstellt wurden. Das System vertraute also vollständig auf die Leistung der externen IR-Komponente und nützte die Anzahl möglicher Antworten in keiner Weise aus.[84]

Bessere Erfolge erreichte die University of Maryland. Ihr System, das syntaktische Strukturen als Selektionsbeschränkungen einsetzte, war in 50-Byte Kategorie das 6. beste und beantwortete 80 Fragen korrekt.[85]

---

[79] University of Pennsylvania, S. 1f.
[80] MITRE, S. 4f.
[81] CL-Research, S. 6.
[82] CL-Research, S. 6f.
[83] Vergleiche Anmerkung 72.
[84] University of Sheffield, S. 6f.
[85] NIST. Track Report, S. 3.

Weil nur wenige Beispiele vorhanden sind, die zudem schlecht optimierte Komponenten verwenden, kann der Nutzen solcher Strategien an den TREC-8-Resultaten nicht abgelesen werden.

7.3 Anpassung von Aufgabenstellung und Systemstrategien in TREC-9[86]

In TREC-9 mussten die Teilnehmer mit Hilfe eines grösseren Korpus von ca. 3 Gigabyte, der aus 6 Quellen stammte, 693 Fragen (bewertet wurden 682) beantworten. Davon waren 193 syntaktische Varianten zu den 500 Originalfragen. Die Fragen waren auf realistischere Weise gewonnen worden als in TREC-8. Echte Fragen und Frage-Ideen aus Logfiles wurden verwendet. Die Frage-Ideen wurden ohne Bezug auf Dokumente zu Fragen formuliert. Erst anschliessend wurde getestet, ob Antworten im Korpus vorhanden waren.

Die Antwort muss in TREC-9 neu auf jeden Fall unterstützt werden. In TREC-8 hatten einige Systeme die Antworten in Lexika ermittelt und Strings gesucht, die diese Wörter enthielten, während das Dokument keine Begründung für die Antwort enthielt. Diese Ausnahmeregelung hätte eigentlich die Generierung unterstützen sollen. Weiterhin richtig ist eine Antwort, wenn sie aus einem irrenden Dokument stammt.

Die Aufgabestellung war insgesamt schwerer und die Systeme erreichten schlechtere Punktzahlen, waren aber dennoch deutlich besser geworden. Auf die veränderten Anforderungen reagierten sie mit der Verfeinerung der Fragetypenklassifikation, besseren Methoden zum Finden der Antwort und häufigerem Einsatz von WordNet.

7.4 Lohnende Ansätze für LUIS[87]

Das Preprocessing von LUIS erkennt einige fixe Muster wie E-Mail-Adressen, Daten oder Währungen. Diese Erkenntnisse werden bisher aber noch nicht zur Beantwortung der Frage verwertet. Sinnvoll wäre die Klassifikation von Fragetypen verbunden mit NE-Erkennung. Die NEs und die bereits erkannten Muster könnten verschiedenen Typen von Fragen zugeordnet werden. Die Desambiguierung von Fragewörtern ist im Deutschen genau wie im Englischen möglich. Die Erkennung der NEs dürfte allerdings wegen der Gross- und Kleinschreibung etwas mehr Schwierigkeiten bereiten.

Sinnvoll ist das Einbeziehen von Named Entities und Fragetypen schon allein deshalb, weil viele einfache, knapp formulierte Fragen, für die normalerweise zu wenige Schlüsselwörter zur Verfügung stehen, durch diese zusätzliche Suchanweisung beantwortet werden könnten.

---

[86] Dieses Unterkapitel stützt sich auf: NIST. Overview of the TREC-9 Question Answering Track. E. Voorhees.
[87] Die Angaben zu LUIS stützen sich auf das Paper von Roberto Nespeca vom 25.04.01.

Andererseits könnten die NEs auch zur Ausdünnung der gefundenen Paragraphen und Textstellen dienen. Durch die mehrstufige Suchstrategien die Anzahl in Frage kommender Textstellen dynamisch angepasst werden.

Zusammenfassung

Trotz offenem Ansatz behandelt die TREC bedingt durch die Auswahl der Fragen ein Teilproblem des Question-Answering. Für die Beantwortung kurzer, Fakten suchender Fragen hat sich die Bestimmung eines Antworttyps, verbunden mit der Suche nach einem entsprechenden Element, als fruchtbar erwiesen. Für diese Art Fragen ist es zudem sinnvoller, nicht relevante Dokumente, sondern relevante Passagen zu suchen. Der Nutzen verschiedener NLP-Strategien, wie Auflösen von Koreferenzen oder Erzeugen einer semantischen Repräsentation mittels logischer Formen, kann nicht schlüssig ermittelt werden. Fehleranalysen verschiedener Systeme lassen den Nutzen von Koreferenzauflösung vermuten. Logische Formen scheinen von anderen Komponenten massiv in ihrer Leistung eingeschränkt zu werden. Dies legen Analysen zu ähnlichen Komponenten von anderen Systemen nahe. Da aber nur die Sheffield University mit solchen Formen gearbeitet hat, lassen sich keine definitiven Schlüsse ziehen. Die Verwendung logischer Formen sollte weiterhin geprüft werden. Das legen auch die besseren, auf einem ähnlichen Ansatz beruhenden, Ergebnisse von CL-Research nahe. Deshalb sollten die Entwicklungen in den Q&A-Tasks der TREC weiterverfolgt werden. Von den in TREC-8 verwendeten Strategien bieten sich als Ergänzung zu LUIS Fragetypologien und NE-Erkennung an, weil sie keine grossen Systemanpassungen erfordern, sondern als Einzelkomponenten eingebaut werden können.

Bibliographie

Literatur

AT&T. AT&T at TREC-8. A. Singhal, S. Abney, M. Bacchiani, M. Collins, D. Hindle, F. Pereira.

CL Research. Question-Answering Using Semantic Relation Triples. K.C. Litkowski.

Cymfony Inc. Information Extraction Supported Question Answering. R. Srihari, W. Li.

GE Research & Development/ Rutgers University/ Swedish Institute of Computer Science/ Stockholm University/ Conexor OY, Helsinki.

M. Hess. Einführung in die Computerlinguistik I. Wintersemester 1998/99.

IBM T.J. Watson Research Center/Columbia University. The Use of Predictive Annotation for Question Answering in TREC8. J. Prager, D. Radev, E. Brown, A. Coden (IBM T.J. Watson Research Center), V. Samn (Columbia University).

LIMSI-CNRS. QALC - the Question-Answering program of the Language and Cognition group at LIMSI-CNRS. O. Ferret, B. Grau, G. Illouz, C. Jacquemin, N. Masson.

The MITRE Corporation. A Sys Called Qanda. E. Breck, J. Burger, L.Ferro, D. House, M. Light, I. Mani.

MUC 6 Task Definition. http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html. Stand vom 06.07.01.

MultiText/University of Waterloo/Carnegie Mellon University. Fast Automatic Passage Ranking (MultiText Experiments for TREC-8). G.V. Cormack, C.L.A. Clarke, D.I.E. Kisman (University of Waterloo) C.R. Palmer (Carnegie Mellon University).

National Institute of Standards and Technology. Overview of the Eighth Text REtrieval Conference (TREC-8). E. Voorhees, D. Harman.

National Institute of Standards and Technology. Overview of the TREC-9 Question Answering Track. E. Voorhees.

National Institute of Standards and Technology. (NIST). The TREC-8 Question Answering Track Evaluation. E. Voorhees, D. Tice.

National Institute of Standards and Technology. (NIST). The TREC-8 Question Answering Track Report, E. Voorhees.

National Taiwan University. Description of Preliminary Results to TREC-8 QA Task. C-J Lin, H-H Chen.

New Mexico State University CRL's TREC-8 Systems Cross-Lingual IR, and Q&A. B. Ogden, J. Cowie, E. Ludovik, H. Molina-Salgado, S. Nirenburg, N. Sharples, S. Sheremtyeva

NTT Data Corporation. NTT DATA: Overview of system approach at TREC-8 ad-hoc and question answering. T. Takaki.

RMIT/Sharp Laboratories of Europe Ltd./CSIRO. The RMIT/CSIRO Ad Hoc, Q&A, Web, Interactive, and Speech Experiments at TREC 8. M. Fuller, M. Kaszkiel, S. Kimberley, J. Zobel (RMIT), C. Ng (RMIT and Sharp Laboratories of Europe Ltd.), R. Wilkinson (CSIRO), M. Wu (RMIT and CSIRO).

Seoul National University. SCAI TREC-8 Experiments. D-H Shin, Y-H Kim, S. Kim, J-H Eom, H-J Shin, B-T Zhang.

Southern Methodist University. LASSO: A Tool for Surfing the Answer Net. D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Gîrju, V. Rus.

University of Iowa. Filters, Webs and Answers: The University of Iowa TREC-8 Results. D. Eichmann, P. Srinivasan.

University of Maryland, College Park, MD/ University of Manitoba/ University of Maryland, Baltimore County, MD. TREC-8 Experiments at Maryland: CLIR, QA and Routing. D.W. Oard, J. Wang (University of Maryland, College Park, MD, D. Lin (University of Manitoba), I. Soboroff (University of Maryland, Balitmore County, MD).

University of Massachusetts. INQUERY and TREC-8. J. Allan, J. Callan, F-F Feng, D. Malin.

University of Ottawa/ National Research Council. Ask Me Tomorrow: The NRC and University of Ottawa Question Answering System. J. Martin (National Research Council), C. Lankester (University of Ottawa).
University of Pennsylvania. Using Coreference in Question Answering. T. Morton.
University of Sheffield. University of Sheffield TREC-8 Q & A System. K. Humphreys, R. Gaizauskas, M. Hepple, M. Sanderson.
Xerox Research Centre Europe. Xerox TREC-8 Question Answering Track Report. D. Hull.

Abkürzungsverzeichnis

IR	Information Retrieval
NE	Named Entity
NLP	Natural Language Processing
NP	Nominalphrase
Q&A	Question-Answering
SMU	Southern Methodist University