Was ist Open Domain Question and Answering

- Automatisches Beantworten einer Frage
- Fragen werden meist natürlichsprachlich gestellt
- Knowledge base ist eine große Sammlung von Texten
- Texte und Fragen sind nicht auf eine Domäne eingeschränkt
- Erste Evaluation auf der TREC 8 (1999)
- Die besten Systeme können schon über 70% aller gestellten Fragen beantworten

Historisches, Wurzeln von OD-QA

- Erste closed Domain QA Systeme gibt es schon seit Anfang der 60er. Berühmte Systeme
  - BASEBALL(1961): NLP Frontend zu einer Baseball Datenbank
  - LUNAR(1971): System für Geologen zur Beantwortung von Fragen bezüglich der Ausswertungen der Apollo Mission. Dieses System konnte über 90% der Fragen richtig beantworten.
- Interaktive Dialog Systeme wurden Anfang der 70er entworfen, zum Beispiel GUS, ein künstlicher Reiseberater(1977).
- Seit Ende der 70er beschäftigt man sich mit Text-Verständniss Systemen. Diese Systeme sollen Testfragen über einen Text richtig beantworten.

TREC 8 Regeln

- Es wurden 198 Fragen gestellt bei einer Korpusgröße von 528.000 Dokumenten
- Antwortlänge darf je nach Lauf maximal 50 oder 250 Bytes lang sein
- Es müssen immer 5 Antworten zurückgegeben werden. Je nach Position der richtigen Antwort wird gewertet.

$$MRR := \frac{1}{Q} \sum_{i=1}^{Q'} \frac{1}{\text{Platz der richtigen Antwort zur i-ten Frage}},$$

  wobei Q die Anzahl aller Antworten, Q' die Anzahl vom System richtig beantworteter Fragen ist.
- Fast alle Fragen waren auf Fakten basierende, schnell zu beantwortende Fragen und die Antworten waren in den Texten vorhanden
- Das beste System erreichten eine MRR von 0.66 und konnten ca. 70% aller Fragen beantworten

Änderungen bei TREC 9

- Fragenanzahl auf 682 Fragen und Korpusgröße auf 979.000 Dokumente erhöht
- Mehrere Formulierungen einer Frage
- Aufnahme schwierigerer Fragen ohne eind. Antwort, z.B. *Who is Colin Powell?*
- Das FALCON System war mit Abstand das beste mit einer MRR von 0.58 (zu 0.32) und 66% richtig beantworteten Fragen (zu 43%). Bei einer Antwortlänge von 250 Bytes konnte das System sogar 86% richtig beantworten.

Änderungen bei TREC 10

- Fragenanzahl: 500
- Antwortlänge max. 50 Bytes
- Fragen ohne Antwort im Textkorpus
- ca. 25% der Fragen waren Definitionsfragen
- Neuer Track: Statt nur einer Antwort wurde eine Liste von Antworten verlangt. Zum Beispiel:
  Name 4 countries that can produce synthetic diamonds
- Insight aus Moskau war das beste System mit einer MRR von 0.68 und 69% richtig beantworteten Fragen
- Im List-task war das FALCON System mit einer Genauigkeit von 76% das Beste

Änderungen bei TREC 11

- Komplett neuer Textkorpus mit ca. 1 Millionen Dokumenten (3 Gb)
- Die exakte Antwort war gefordert (keine Textschnipsel)
- Nicht mehr fünf, sondern nur eine Antwort, dafür sollte das System die Güte seiner Antworten selber bewerten
- Neues Bewertungssystem: Confidence-weighted score (Q Antworten als geordnete Liste)

$$\frac{1}{Q}\sum_{i=1}^{Q}\frac{\text{Anzahl Richtiger unter den ersten i Antworten}}{i}$$

- Bestes System mit Abstand war PowerAnswer, welches aus FALCON hervorging mit 83% richtig beantworteten Fragen

Aufbau von OD-QA Systemen

Ein typisches QA System besteht aus mehreren Modulen

- Analyse der Frage
- Präprocessing der Dokumentensammlung
- Auswahl von Kanditaten innerhalb der Dokumentsammlung
- Analyse der Kanditaten
- Extraktion der Antwort
- Anzeige der Antwort

Analyse der Frage

- Manchmal wird nur eine Teilmenge der Sprache zugelassen (Limitiertes Vokabular und Syntax)
- Es kann impliziten Input geben, z.B. Fragen, die sich auf vorherige Fragen beziehen, oder best. Userpreferenzen
- Das System muss die Frage parsen und in eine Form bringen, die für die nächsten Module passend ist. Typischer Weise muss das System folgendes erkennen:
  - Den Antworttyp (Ort (Where), Person (Who), Datum (When),...). Hier kann es Schwierigkeiten geben, wenn der Antworttyp nicht offensichtlich aus der Fragestellung hervorgeht, z.B. „Was war das erste Auto von Opel?“.
  - Identifikation von Keywords für die Textsuche
  - Syntaktischer und semantsicher Zusammenhang der einzelnen Entitäten

Präprocessing der Dokumentensammlung

- Um in dem Korpus geeignet zu suchen muss er vorher verarbeitet werden.
- Hier werden Indexierungen wie beim klassischen Information Retrieval vorgenommen.
- Manche Systeme machen ein shallow Parsing des Korpuses, z.B. Tagging, Chunking, NER
- Manche Systeme versuchen Relationen aus den Texten zu extrahieren z.B. der Form: Subject - Relation - Objekt um genauer Suchen zu können, vorallem bei nicht symmetrischen Relationen.

Auswahl von Dokumentenkanditaten

- Um Dokumente zu finden, welche die Antwort enthalten könnten, wird die Dokumentsammlung mit klassen IR Techniken durchsucht.
- Die meisten QA Systeme benutzen hierbei boolische und nicht rangbasierte Suchsysteme
- Oft werden die Dokumente weiter unterteilt, z.B. in Absätze oder die zu suchenden Wörter müssen innerhalb eines Fensters sein
- Seit der TREC 10 wird allen Teilnehmern eine leistungsfähige Suchmaschine zur Verfügung gestellt
- Viele Systeme wiederholen die Suche mit anderen Suchwörtern bei unzureichender Dokumentenanzahl oder Qualität, z.B. erhöht das LASSO System die Anzahl der Suchwörter bei zu vielen Ergebnissen
- Ist der Korpus schon speziell bearbeitet worden, können weitergehende Suchtechniken benutzt werden

Analyse der Kandidaten

- Typischer Weise werden die gefunden Dokumente oder Textpassagen intensiver analysiert.
- Die meisten Systeme machen hier (wenn nicht schon im Präprocessing geschehen) eine NER. Diese erkennt meistens viele verschiedene Klassen von Entitäten, z.B. Orte, Personen, Firmen, Produkte, Daten,... und kategorisiert auch Unterklassen, z.B. bei Ort: Stadt, Land, Staaten,...
  Dies ist wichtig, um den richtigen Antworttyp zu finden.
- Typischer Weise werden auch andere NLP Techniken eingesetzt wie Tagging und Chunking
- Manche Systeme versuchen auch eine logische Struktur der Dokumente aufzubauen

Extraktion der Antworten

- Bis jetzt hat man eine Darstellung der Frage und der in Frage kommenden Dokumente, hieraus wird eine Liste der in Frage kommenden Antworten erstellt, sortiert nach ihrer Wahrscheinlichkeit.
- Überprüfung des Antworttyps. Hierbei wird häufig WordNet zu Hilfe genommen.
- Berechnung der Wahrscheinlichkeiten durch Linearkombination heuristischer Funktionen , z.B. Abstand der Schlüsselwörter.
- Systeme mit logischer Darstellung der Sätze können weitere Überprüfungen vornehmen.

Darstellung der Antwort

- Bei den frühen TREC's sollte ein 50 bzw. 250 Byte Fenster ausgegeben werden
- Zusätzlich noch das Dokument in dem die Antwort gefunden wurde
- Im praktischen Einsatz möchte man nur die Antwort kein Textschnipsel, dem wurde bei TREC 11 Rechung getragen

Typischer Aufbau (Diagramm)

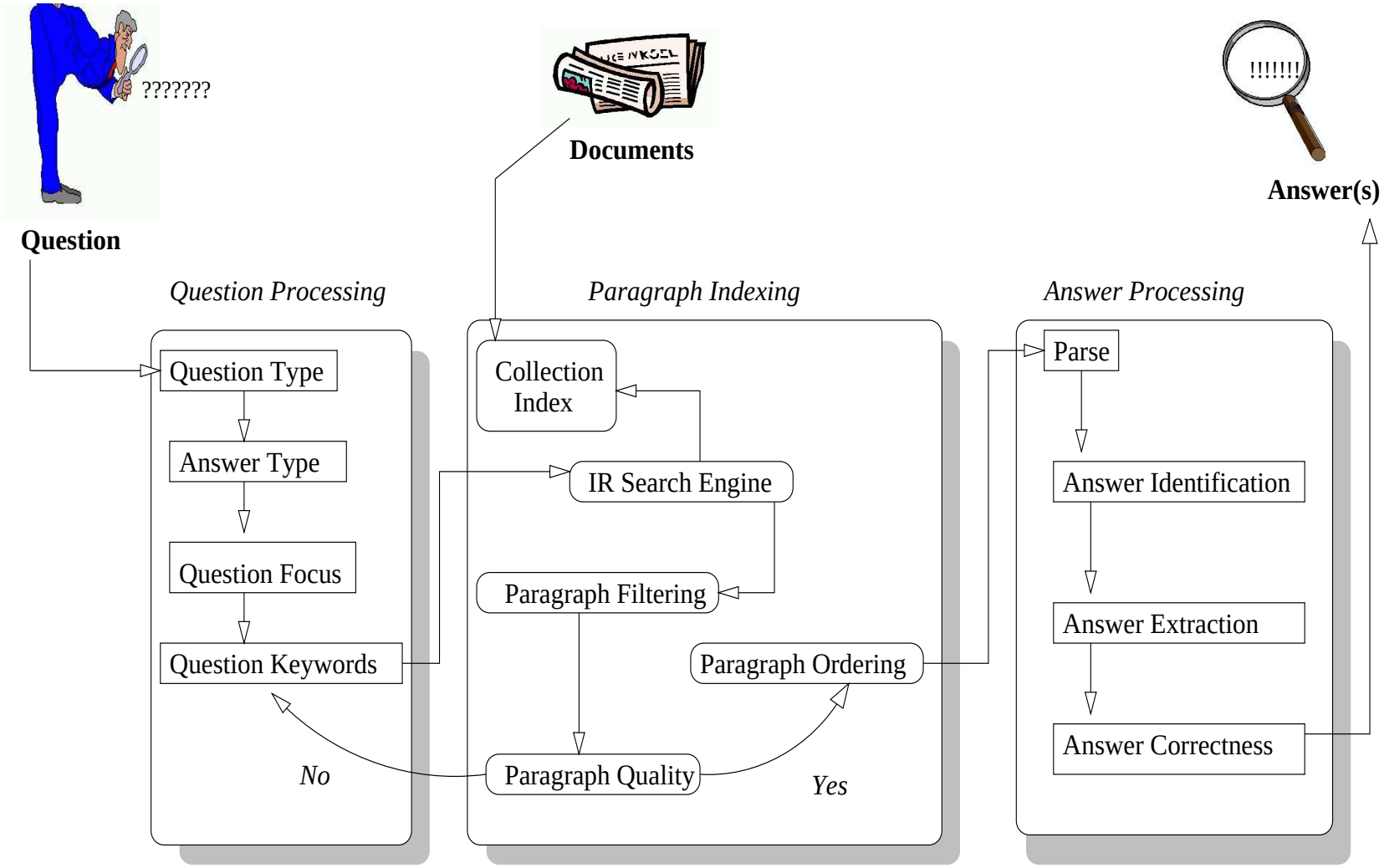
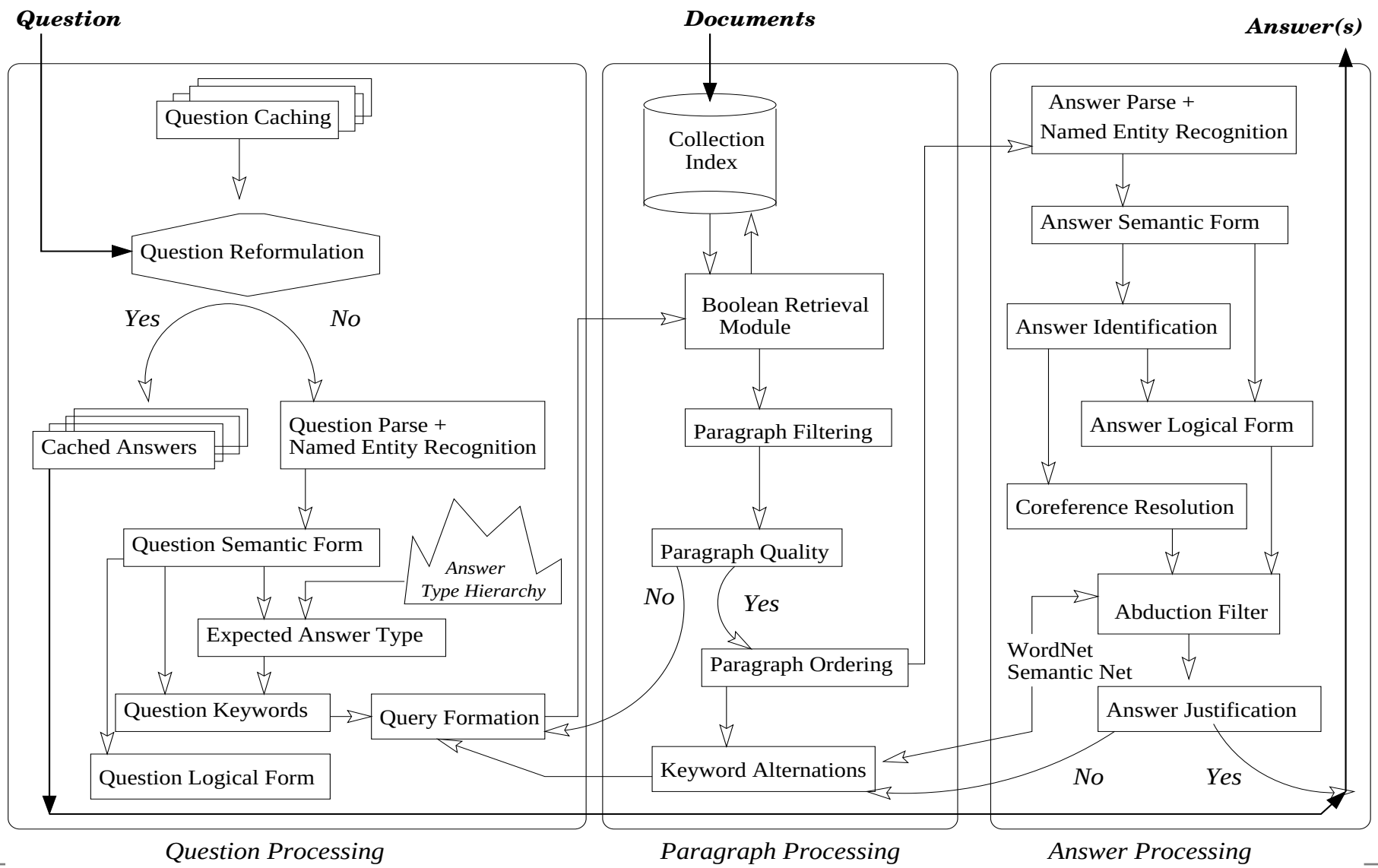


Figure 1: Architecture of the LASSO Q/A System

Highlights des FALCON Systems

- Entwickelt von Sanda Harabagiu, Don Moldovan, Marius Pasca, u.a von der *Southern Methodist University in Dallas, Texas*
- Erfolgreichstes System auf der TREC 9
- Nutzt intensiv WordNet
- Stellt sowohl Fragen als auch in eventuelle Antworten in einer semantischen und logischen Form dar.
- Versucht Reformulierungen von schon gestellten Fragen zu erkennen

Falconaufbau(Diagramm)

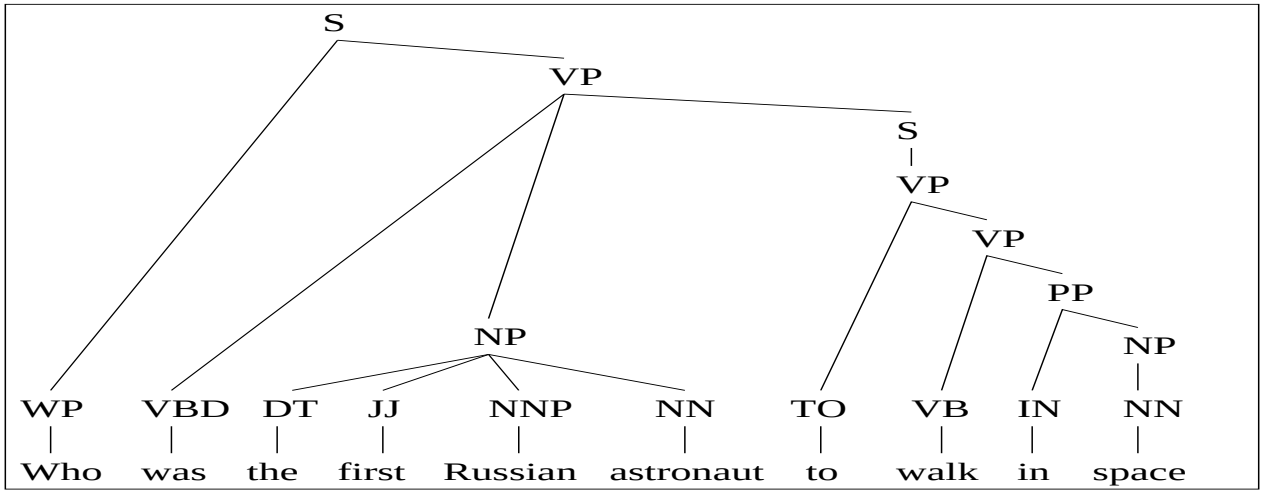


Frageverarbeitung

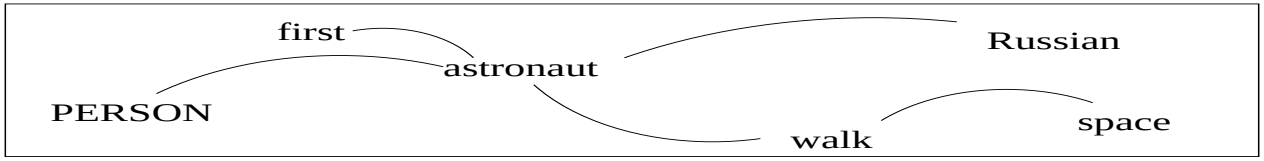
Frage wird mit einem statistischen Parser geparst und dann in eine semantische Form gebracht. Dabei werden auch Schlüsselwörter herausgefunden.

Q733: Who was the first Russian astronaut to walk in space?

Question parse:



Question semantic representation:



Frageverarbeitung 2

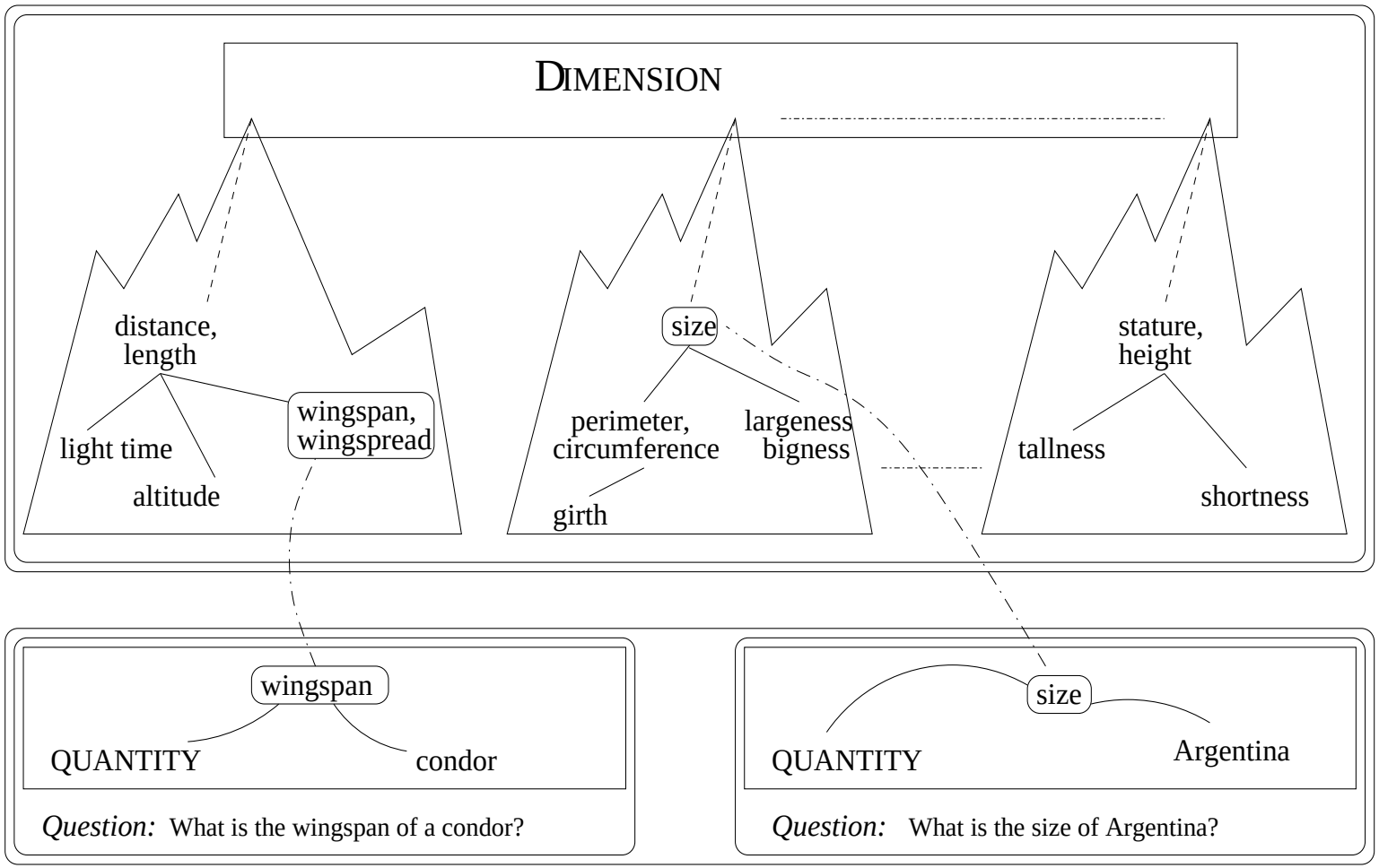
- Der erwartete Antworttyp wird aus einer Antworttaxonomy gewählt, diese ist verzweigt. FALCON benutzt 18 Top Level Klassen:

Date	Time	Organization
Reason	Manner	Nationality
Product	Money	Language
Mammal	Game	Dog breed
Location	Reptile	Numerical value
Quotation	Alphabet	Percentage

- Jedem Antworttyp ist ein NER Typ zugeordnet. FALCON's NER Modul erkennt 27 verschiedene Kategorien.
- Um den zu erwarteten Antworttyp herauszufinden wird die Frage erst nach klassischen Fragemustern dursucht, z.B. *Who (is|was|are|were) <person name(s)>?* .

WordNet und Antwort Taxonomy

Der Wort, nach dem gefragt wird, wird aus der semantischen Form der Frage gewonnen. Dann wird mittels WordNet der Antworttyp bestimmt.

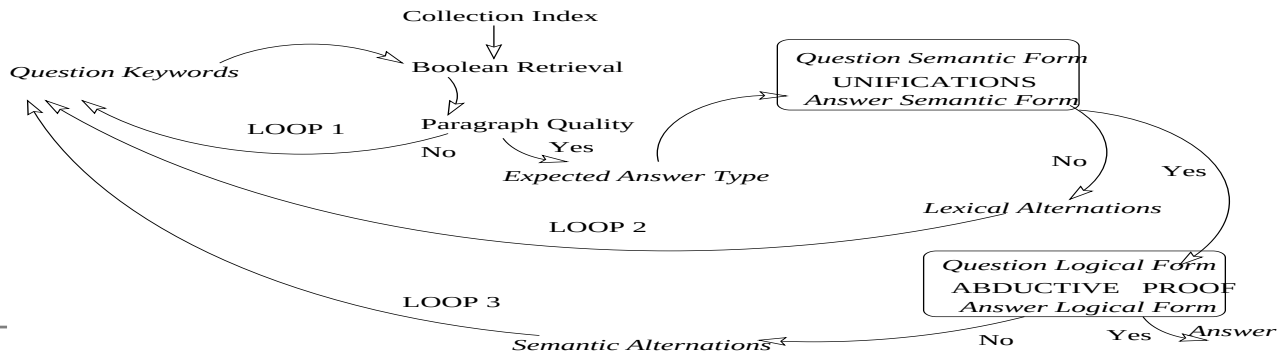


Keyword Alternativen

- FALCON benutzt für die Suche auch Varianten zu den Schlüsselwörtern
  - Morphologische Alternativen nach WordNet, z.B. für erfinden Erfinder
  - Lexikalische Alternativen, z.B. für töten ermorden
  - Semantische Alternativen
- Je nach Anzahl und Güte der gefundenen Seiten, werden die Schlüsselwörter variiert

Antwort Verifikation

- Um die Güte gefundenen Antwort zu berechnen, benutzt FALCON keine heuristischen Funktionen, sondern versucht die Antwort auf die Frage zu verifizieren.
- Dafür wird sowohl die Frage als auch die Antwort in logische Formen konvertiert, die dann durch einen Rückwärtsalgorithmus miteinander abgeglichen werden.
- Dadurch ist das System in der Lage auf eine Frage wie z.B. *Who killed Lee Harvey Oswald?* die Antwort *L. H. Oswald killed Kennedy* als falsch zu interpretieren.
- Sollten alle Antworten falsch sein, wird die Suche mit variierten Schlüsselwörtern wieder gestartet.



Zusammenfassung und Ausblick

- QA Systeme können heutzutage faktische Fragen sehr gut und präzise beantworten.
- Weitere Forschung wird betrieben zur Beantwortung komplexerer Fragen, z.B.:
  - Definitionsfragen: *Wer war Boris von Loesch?*
  - Fragen, mit mehreren Antworten: *Nenne die fünf größten Flüsse der Welt?*
  - Crosslanguage
  - Zusammenfassungen: *Worum geht es in dem Buch A?*
  - Meinungen: *Warum sollte man die Steuern erhöhen?*
  - Komplexe Zusammenhänge: *Was passiert, wenn man ...*

Zusammenfassung und Ausblick 2

- Antworten werden bis jetzt meist aus einem Dokument extrahiert
- In Zukunft sollen komplexe Zusammenhänge zwischen den Dokumenten erkannt werden, z.B. Verweise, Vergleich von Aussagen
- Antworten sollen auch in Form von Zusammenfassungen erzeugt werden
- Bessere Interaktion mit dem Nutzer, z.B. Fragen, die sich auf vorige Fragen beziehen, z.B.
  Wer war Abraham Lincoln?
  Wie alt wurde er?
- Nachfragen vom System zu unklaren Fragen
  Wie groß ist Paris?
- Aussagen zur Richtigkeit der vom System gegebenen Antwort
- Antworten sollen in Echtzeit gegeben werden

Am Rande

Auf manche Themen konnte leider nicht eingegangen werden, da dies den Umfang sprengen würde. Diese Fragen sind aber trotzdem interessant.

- Probleme bei der Evaluation von QA Systemen
- Erstellung eines Beispiel Frage und Antwort Katalogs
- QA Systeme und Internet
- Genauere Funktionsprinzipien von QA Systemen
- Zusammenhang zwischen QA und anderen NLP Disziplinen

Noch Fragen?