Pseudo-QA-System mit Paraphrasierung

Softwareprojekt von Andriamboavonjy Antsa Harinala, Mirjam Eppinger
Institut für Computerlinguistik, Universität Heidelberg
WiSe 2011/12

1. Ressourcen

- Yahoo!Answers Corpus: Sammlung von geposteten Fragen (4 483 032) und Antworten aus Yahoo!. Für jede Frage sind mehrere Antworten gespeichert.
- Lucene: Java Library zur Implementierung von Methoden aus dem Bereich des Information Retrievals.
- English Gigaword Corpus (5. Edition): über 4 Mio. Tokens.

2. Pseudo-QA-System

- **QA-System** $\triangleq$ Frage-Antwort-System nach engl. „Question Answering“
- **Ziel**: Liefere möglichst gute Antworten auf eine englische Eingabefrage.
- **Methode**:
  - Speichere die notwendigen Informationen des Korpus im Korpus-Index (z. B. Fragen, Antworten, Kategorie, ect.).
  - Suche innerhalb einer bestimten Kategorie mithilfe von Lucene die Fragen, die der Eingabefrage am ähnlichsten sind.
  - Sortiere die gefundenen Fragen absteigend nach ihrem Score.
  - Durch die Sortierung sind auch die Antworten nach dem Score ihrer Frage sortiert.
  - Der Benutzer kann wählen, wie viele Antworten er möchte.
  - Die Antworten werden in sortierter Reihenfolge ausgegeben.
  - Query Expansion möglich duch Paraphrasierung der Eingabefrage-Ngramme.

3. Beispiele zum QA-System

- **Ohne Query Expansion**:

↪ Kategorie: Software

↪ Frage: *What are the advantages of Linux?*

↪ Antwort 1: *The biggest attractions of Linux are it is security, flexibility of customization, configuration and operating systems and that it is mostly free if you download from internet. There are a lot of free software available for linux on the internet. There is also a lot of online support for linux. Network setup is easier as well.*

↪ Antwort 2 : *Use linux and u will know....Linux has lots of freedom in modifications...When u have any windows os,u will be asked sm odr the time to download the codecs or drivers....so its more hectic and also...windows is now a days affected by lots of viruses...so use linux and find out urself.*

- **Mit Query Expansion (Paraphrasen)**:

↪ Kategorie: other – Politics

↪ Frage: *Is oil the reason for war on iraq?*

↪ Paraphrasen: *Japanese war War crisis warfare coalition military action top war*

↪ Antwort 1: *well it is a mix of many many reasons*

↪ Antwort 2: *One of the main reasons which the US government cites was the liberation of the Iraqi people from the dictatorship of Saddam Hussain, though it was a good reason there was no need to amass your troops and those of other allied nations into iraq so now the Iraqis are being opressed by another military dictatorship.*

4. Paraphrasen-Extraktion

- **Grundidee**: Unterschiedliche Wörter bzw. Ngramme, die häufig von den gleichen Wortfenstern umgeben sind, sind einander mit großer Wahrscheinlichkeit ähnlich oder sogar gleichbedeutend.

- **Verfahren nach Pașca & Dienes 2005**
  - Splitte alle Ngramme des Gigaword Corpus in Anker und variables Fragment.
  - Betrachte unterschiedliche variable Fragmente mit dem gleichen Anker als Paraphrasen voneinander.
  - Bestimmte Parameterwerte sind notwendig: Häufigkeit, wie oft unterschiedliche variable Fragmente den gleichen Anker haben, ect.

- **Möglich: Berücksichtung von Zusatzinformationen. Drei Varianten:**
  - ohne Zusatzinformation
  - mit entity-Zusatzinfo: berücksichtige Named Entities in den Sätzen der Ngramme.
  - mit relative-Zusatzinfo: berücksichtige Relativsätze.

- Einmalige Ermittlung der Paraphrasen
- Verarbeitung der riesigen Datenmenge mit Hadoop

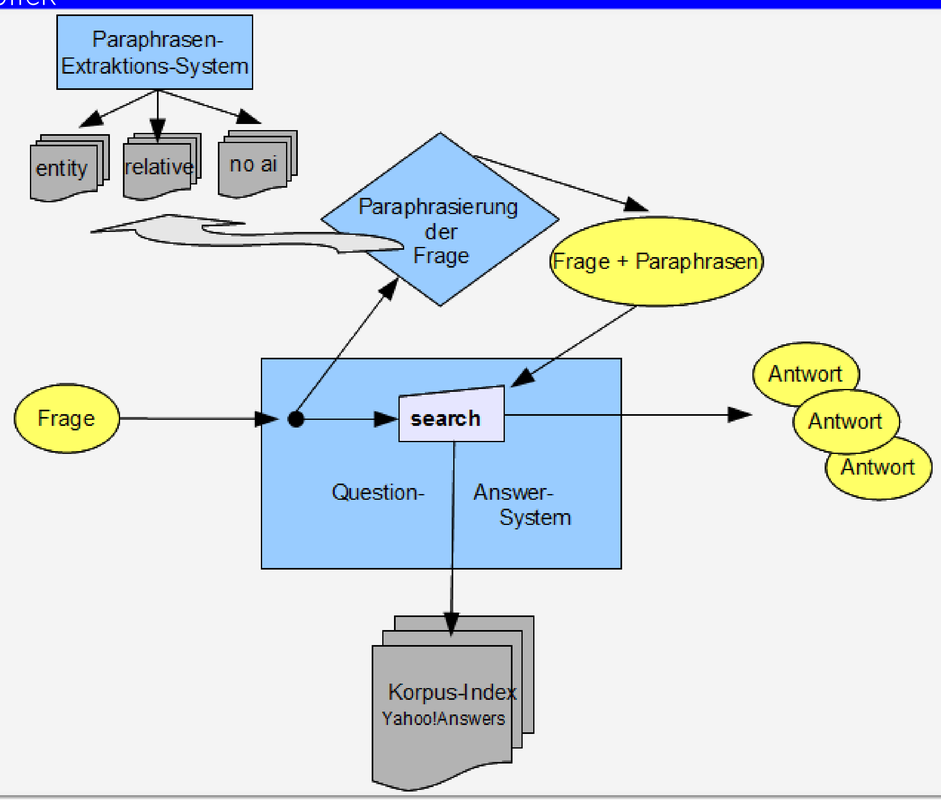
5. Beispiele zur Paraphrasen-Extraktion

- **Version relative**:
  - Frage: *What are the G7 or G8 countries?*
  - Paraphrasen: *states nations*

- **Version no additional info**:
  - Frage: *How do I delete my search the web list?*
  - Paraphrasen: *Web Internet search index*

- **Version entity**:
  - Frage: *who is Israel chief rabbi?*
  - Paraphrasen: *leader commisioner head director chairman chief executive Director General Secretary Generay Commandos Gen. Secretary governor ...*

6. Überblick



7. Evaluation und Ausblick

- **Erstellung des Goldstandards**
  - Betrachte die Antworten auf eine Menge von 2100 Fragen.
  - Betrachte für jede dieser 2100 Fragen die Antworten auf diese Frage.
  - Suche für jede dieser Antworten innerhalb des Training Sets nach den Antworten, die dieser am ähnlichsten sind.
  - Bilde für jede Frage die Schnittmenge aus allen für diese Frage gefundenen Antwort−Mengen.
  - All diese Schnittmengen bilden den Goldstandard.

- **Ergebnisse**

	Mean Precision	Mean Recall
Baseline	4,75%	11,37%
Variante: Paraphrase entity	4,23%	12,56%
Variante: Paraphrase relative	4,29%	14,00%
Variante: Paraphrase no additional information	4,11%	10,69%

- **Ausblick**
  - Parameterwerte für die Paraphrasen-Extraktion verbessern
  - Erstellung des Goldstandards verbessern

Literatur

Pașca, M & Dienes, P. (2005): *Aligning needles in a haystack: Paraphrase acquisition across the Web*. In: Proceedings of the 2nd International Joint Conference on Natural Language Processing, S. 119 – 130.

Jeon, J., Croft, W. B. & Lee, J. H. (2005): *Finding Similar Questions in Large Question and Answer Archives*. In: Proceedings of the ACM Fourteenth Conference on Information and Knowledge Management, S. 76 – 83.

Cooper, R. J. & Rüger, S. M. (2000): *A Simple Question Answering System*. In Proceedings of the 9th Text Retrieval Conference, S. 249 – 255.