Structured Data and Inference

Fritjof Wolf

11. Januar 2013

Überblick

- Motivation
- Quellen für strukturierte Daten
- Anwendungen strukturierter Daten
- Zukünftige Verbesserungen

Motivation

- Strukturierte Daten vs. unstrukturierte Daten
- Überprüfung von Nebenbedingungen
  => zusätzliche Bewertung der Antwortkandidaten
- Beispiel:

THE HOLE TRUTH (1200): Asian location where a notoriously horrible event took place on the night of June 20, 1756. (Answer: „Black Hole of Calcutta“)

Quellen strukturierter Daten

- Standard Online-Datenbanken
- Automatisch extrahierte Daten
- Manuell extrahierte Daten

Standard Online-Datenbanken

- Unterschied zu WolframAlpha und CYC
- Standard-Datenbanken, z.B. für Filme, Bücher etc.
- DBpedia und Freebase

  → Vereinheitlichung der Datenformate
- Verbindung zu Ontologien, wie z.B. YAGO

Automatisch extrahierte Daten

- Wissensdatenbank aus Tupeln (Entität, Datum, Anzahl)

  → Anzahl des gemeinsamen Auftretens

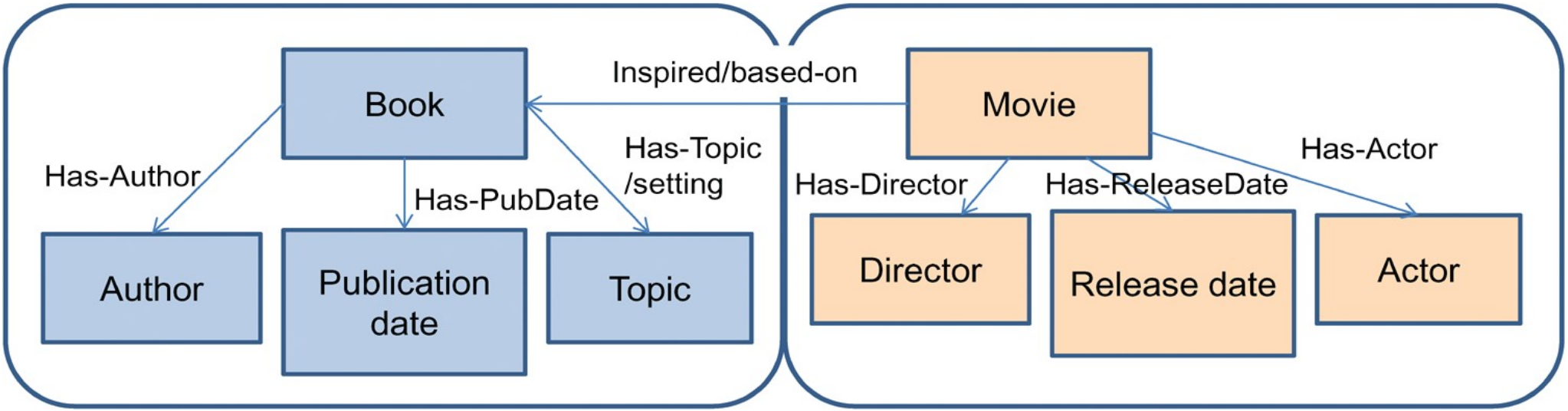
- PRISMATIC

Manuell extrahierte Daten

- Typen-Unverträglichkeit von Oberklassen in YAGO

- Frames für wichtige Themen und Kategorien, wie z.B. U.S. Präsidenten, Länder und Hauptstädte, Preise etc.
- Frames repräsentieren zusammenhängende Gruppe von Begriffen (drücken Erwartung aus)
- Gut zur Überprüfung von Nebenbedingungen geeignet

Frames für Bücher und Filme



Kalyanpur.A, u.a.: „Structured Data and Inference in DeepQA“, S. 4

Anwendungen strukturierter Daten

- Zeitliche und geografische Nebenbedingungen
- Taxonomisches Schlussfolgern
- Semantische Frames
- Verschiedenes

Zeitliche Nebenbedingungen

- Extraktion in Analysephase
- TLinks: Tupel aus Entität und Zeitangabe
- Überprüfen der Kompatibilität der Daten
- Abgleich mit einer Liste mit den wichtigen Daten einer Entität

Geografische Nebenbedingungen

- Mögliche Kriterien: Himmelsrichtung, Grenzen, nah und fern
- Suche geografische Informationen in Dbpedia und Freebase
- Direkte Suche in Freebase
- „Nah und fern“-Einordnung mithilfe maschinellen Lernens
- Implizites „Nah“, Beispiel:

„River that connects Lake

Powell in Utah with the Gulf of California“

Evaluation

Table 1 Temporal questions scored; 662 temporal questions out of 3,508-question test set.

	Baseline	*Plus temporal*
Accuracy	66.62%	67.37% (+0.75%)
Precision@70	80.82%	81.9% (+1.08%)

Table 2 Spatial questions scored; 374 spatial questions out of the 3,508-question test set.

	Baseline	*Plus spatial*
Accuracy	65.78%	67.65% (+1.87%)
Precision @ 70	78.63%	80.15% (+1.52%)

Kalyanpur.A, u.a.: „Structured Data and Inference in DeepQA," S. 6

Taxonomisches Schlussfolgern

- Verschiedene TyCor-Algorithmen, Beispiel YAGO TyCor
- Bestimme Typen der möglichen Antworten und der Frage

  → Verbesserung durch Heuristiken
- Typenabgleich in YAGO WordNet
- Verschiedene Regeln, zum Beispiel Synonyme, Unterklassen, Disjunkte Mengen

  => Zahl, die die Ähnlichkeit der beiden Typen angibt

Evaluation

Table 3 Evaluation of accuracy and precision.

Baseline	*Baseline accuracy (Precision@70)*	*Plus YAGO (An)TyCor*	*Plus all TyCors except YAGO (An)TyCor*	*Plus all TyCors including YAGO (An)TyCor*
1. DeepQA system with no answer scorers	50.03% (Precision@70: 63.44%)	54.28% (Precision@70: 67.63%)	58.55% (Precision@70: 75.37%)	58.64% (Precision@70: 75.24%)
2. DeepQA system with all answer scorers except TyCors	65.48% (Precision@70: 81.43%)	68.39% (Precision@70: 83.84%)	69.38% (Precision@70: 86.93%)	70.35% (Precision@70: 87.42%)

Kalyanpur.A, u.a.: „Structured Data and Inference in DeepQA“, S. 6

Semantische Frames I

- Sonderstellung: Unabhängige Pipeline
- Frame-Erkennungsalgorithmen
- Implizite Annahmen über Beziehungen zwischen Entitäten
- Beispiel:

LANGUAGE: The lead singer of the band Dengue Fever is from this country & often sings in Khmer.

(Answer: „Cambodia“)

Semantische Frames II

- Besonders geeignet, wenn Fokus der Frage ein Frame ist
- Erleichtert die Suche nach mehreren Antworten
- Beispiel:

TRICKY QUESTIONS: Of current U.N. member countries with 4-letter names, the one that is first alphabetically. (Answer: „Chad“)

Evaluation

- Kleiner Anwendungsbereich, aber hohe Trefferquote darin
- 90.771 Fragen, bei 223 genau ein Treffer, dabei 87 % richtig
- Gute Ergänzung zur Hauptpipeline
- Bevorzugung bei speziellen Fragen

Andere Anwendungen

- Viele verschiedene Einsatzgebiete
- Answer-in-clue Komponente
- Inferenz
- Beispiel:

($200) WHO SENT ME TO THE SC: Ruth Bader Ginsberg. (Answer: „Bill Clinton“)

($400) WHO SENT ME TO THE SC: Clarence Thomas. (Answer: „George H W Bush“)

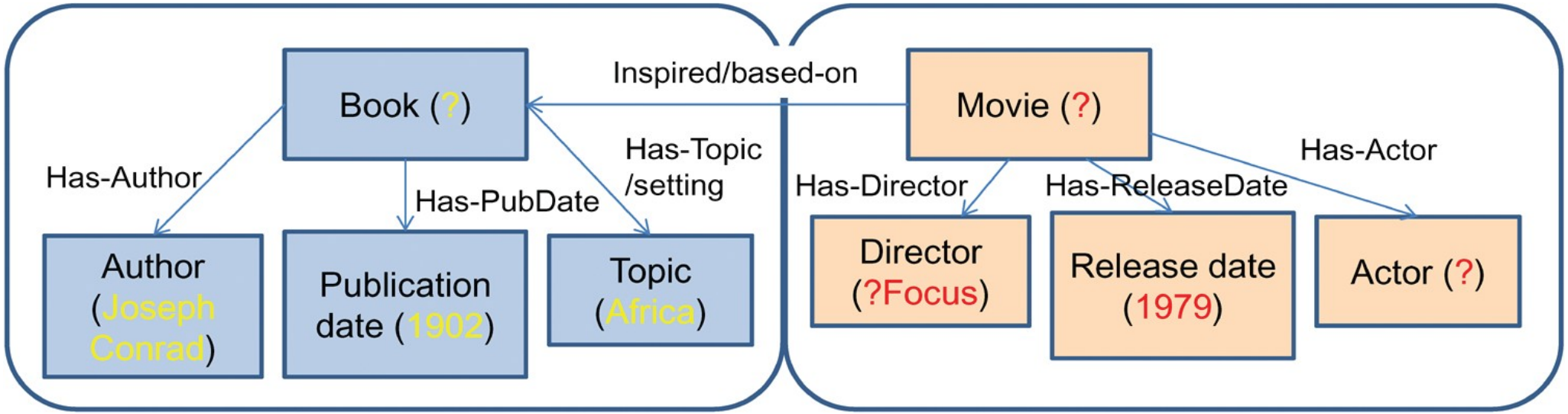
($600) WHO SENT ME TO THE SC: Thurgood Marshall. (Answer: „Lyndon Johnson“)

Zukünftige Entwicklungen

- Mehr Ontologien aus der Linked Open Data Cloud
- Künstliche Intelligenz zur Interpretation der Frames
- Iterative Verfahren zur Lösung einer Frage
    - Starte mit den Anfangs bekannten Daten
    - Formuliere Teilfragen und löse diese
    - Löse eigentliche Frage
- Automatische Suche nach typischen Relationen in einem Frame mithilfe von Data Mining

Beispiel

WAR MOVIES: A 1902 Joseph Conrad work set in Africa inspired this director to create a controversial 1979 war film.



Kalyanpur.A, u.a.: „Structured Data and Inference in DeepQA“, S. 11

WAR MOVIES: A 1902 Joseph Conrad work set in Africa inspired this director to create a controversial 1979 war film.

1. Solve for the Book by calling DeepQA with the generated question: „This 1902 book by Joseph Conrad is about Africa.“ (Answer: „Heart of Darkness“).
2. Use our structured data to verify that the publication date of the book is 1902.

WAR MOVIES: A 1902 Joseph Conrad work set in Africa inspired this director to create a controversial 1979 war film.

3. Invoke DeepQA to solve for the Movie given the Book, using the question „Heart of Darkness inspired this controversial 1979 war film.“ (Answer: „Apocalypse Now“).
4. Use our structured data to verify that the release date of the movie is 1979.
5. Use our structured data to look up the director of the movie (answer: „Francis Ford Coppola“).

Fragen

Quellenverzeichnis

- Kalyanpur. A, u.a.: „Structured Data and Inference in DeepQA“,