Automatic Knowledge Extraction from large text corpora

Jakob Höper[1]

November 16, 2012

[1] hoeper@zedat.fu-berlin.de

PRISMATIC

Lexikographisches Verzeichnis semantischer Beziehungsinformationen / Knowledge Base

- Annotatierte und klassifizierte Textressourcen
- Extraktion semantischer Beziehungen, Klassifikationen etc. von Satzbestandteilen in *Frames*
- Anwendung semantischer Beziehungsdaten als strukturierte inhaltliche Aussagen
- Anhand gewonnenen Wissens u.a. Beurteilung von Anwortkandidaten, Antworttypen, Ableiten von Anwortmöglichkeiten etc.

Übersicht

Gliederung

1. Aufbau des Textkorpus
   - Auswahl von Quellen
   - Verarbeitung geeigneter Ressourcen
   - Einpflegen in den Korpus
2. PRISMATIC Wissensdatenbank
   - Modellierung der Wissensbasis
   - Import der Rohdaten
   - Anwendungen

PRISMATIC: Rohdaten

- Datenbasis: $\sim$ 30 GB unstrukturierter Textquellen (bereinigt, html-detagged)
- Open Domain: aufgrund der thematische Vielfältigkeit in Jeopardy! müssen Informationen aus möglichst vielen Wissensgebieten zur Verfügung stehen

Ziel: nach Möglichkeit inhaltliche Redundanzen und Paraphrasierungen im Korpus anhäufen, um relevantes Wissen ermitteln zu können.

- Repräsentation in PRISMATIC: ca. 995 Millionen Frames, 1.3 Frames/Satz

Datenaquise - Wikipedia

Wikipedia (Dump von 2010 von 13 GB mit 3.5 Mio Artikeln)

- *Annahme*: besonders geeignet aufgrund von Konzentration auf populäre Themen (freiwillige Mitarbeit/persönliches Interesse)
- Nur knapp fünf Prozent aller richtigen Jeopardy!-Antworten haben keine eigenen Wikipedia-Einträge
- Tatsächlich liegt der recall für Suche und candidate generation auf den Wikipedia-Texten bereits bei über 75%

Datenaquise - Wikipedia (II)

Allerdings: Mangelhafte Ergebnisse bei Fragen nach:

- zu gegebenen Definitionen passendem Wort, Quelle wörtlicher Zitate, ...
- Bibel-Trivia, Roman- und Filmhandlungen, ...

Datenaquise - weitere Quellen

Um Wissenslücken füllen zu können, Nachrüsten mit u.a.:

- zusätzlichen Enzyklopaedien
- Wiktionary, Wikiquote, weiteren Wörterbüchern und Zitatesammlungen
- die Bibel (diverse Editionen), Pressearchive (New York Times etc.) und -artikel
- Buchtexte (Projekt Gutenberg), Werksverzeichnisse, Songtexte

Unpraktisch: Dokumente unterscheiden sich stark in Genre, Aufbau, Länge ...

Vorgehen

Das Anlegen des Textkorpus unterteilt sich in drei Phasen:

1. **Aquise**
   unzureichend abgedeckte Themen werden ermittelt und geeignete Quellensammlungen zugefügt, solange diese die vorhandene Qualität nicht durch Rauschen beeinträchtigen
2. **Transformation**
   aquirierte Quellendokumente werden in eine allgemein vergleichbare Form gebracht
3. **Expansion**
   für die vorraussichtlich brauchbarsten (populärsten) Themen werden zusätzliche Referenzen beschafft, um die Menge an Redundanzen und Paraphrasierungen zu erhöhen

Die Schritte werden wiederholt bis keine Verbesserung mehr zu erreichen ist

Transformation

Ziel I: Titel-bezogene Dokumente

Titel-Bezogenheit spielt in der Ermittlung von Kandidatur-Anworten eine zentrale Rolle für zwei wesentliche Suchansätze:

1. Liefert eine Frage mehrere Eigenschaften der Antwort, steht diese mit hoher Wahrscheinlichkeit im Titel eines passenden Dokuments (document search)
2. Sind Titel-bezogene Dokumente zu in der Frage erwähnten Begriffen vorhanden, ist die Anwort wahrscheinlich unter den Titeln dieser Dokumente (title-in-clue/TIC passage search)

Transformation (II)

Ziel I: Titel-bezogene Dokumente

- **Grundsätzlich Titel-bezogen**: Einträge aus Enzyclopaedien und Wörterbüchern . . .
- **Nicht Titel-bezogen**: Presseartikel, umfangreiche Textkörper wie die Bibel, Aggregationen gesammelter Werke . . .

Aus Dokumenten mit bestimmter Struktur lassen sich Titel-bezogene Pseudo-Dokumente extrahieren:

- ggf. Zerlegen des Dokuments
- Betitelung eines Fragments oder mehrerer seiner Kopien mit Bezeichnern für Entitäten, die als wahrscheinliche Antworten im Zusammenhang mit den im Pseudo-Dokument zu findenden Hinweisen infrage kommen

Transformation (III)

Ziel I: Titel-bezogene Dokumente

Ein Verzeichnis über Shakespeare's gesammelte Werke wird zerlegt in mehrere Sätze Pseudo-Dokumente, für welche jeweils jedes Pseudo-Dokument enthält:

- ein Stück Shakespeare's, der Titel ist immer 'Shakespeare'
- ein Stück Shakespeare's, der Titel ist jeweils der Titel des Stücks
- alle Zeilen eines bestimmten Charakters, der Titel ist jeweils der Name des Charakters

Änalog: Songtexte, Bücher der Bibel, Buchtexte von Project Gutenberg

Transformation (IV)

Ziel II: Optimierung von Titel-bezogenen Dokumenten

Transformationen werden teilweise auf bereits Titel-bezogene Dokumente angewandt, um deren Verwertbarkeit für Suchvorgänge und Evidence-Scoring zu verbessern, und in späterem Parsing die Gewinnung von zusätzlichen semantischen Informationen zu ermöglichen

- Wörterbucheinträge, die einander als alternative Schreibweisen querreferenzieren werden zusammengelegt, um die Menge an Definitionen unter einem Titel zu steigern
- Definitionen in Wörterbucheinträgen werden in grammatikalisch vollständige Sätze umgewandelt, in denen der Titel des Dokuments genannt wird

Expansion

Dokumentquellen aus Dictionaries und Enzyclopaedien können sich als Ressource für die Such- und Evaluationssysteme in DeepQA als ungenügend erweisen:

- Information kann unvollständig sein: nicht alle Fakten sind unbedingt in einem Artikel genannt
- Wörterbücher und Lexika neigen nicht dazu, Informationen zu paraphrasieren und Redundanzen zu schaffen

Deshalb: Ergänzen ausgewählter Themen um Informationen aus Online-Quellen.

Auswahl der populärsten Themen anhand von Wikipedia hyperlink metadata, Häufigkeit eines Worts im allgemeinen Gebrauch (Wikipedia: 300,000 Artikel; Wiktionary: 100,000 Einträge)

Expansion II I

Expansion pipeline

Für jedes der ausgewählten, nach Popularität geordneten *Seed*-Dokumente:

1. **Retrieval**: beschaffe bis zu 100 Ergebnisse einer Suchanfrage bei Yahoo!
2. **Extraction**: Zerlegung der Suchergebnisse in Absätze, Nuggets genannt. Bei html etwa in Elemente wie `p`, `li`, `td`
3. **Scoring**: beurteile die Relevanz eines extrahierten Nuggets hinsichtlich des Ausgangsdokuments
   Ein *logistic regression model*, welches zunächst *supervised* auf einer kleinen Untermenge der *Seed*-Auswahl trainiert wird, verwendet dabei ausschlaggebende Predikatoren wie:
   - Ähnlichkeit zwischen Inhalt des Nuggets und dem des Seed-Dokuments: Wortverteilung, language modeling techniques, tf-idf Wortgewichtung ...

Expansion II II

Expansion pipeline

- Attribute des Suchergebnisses: Yahoo! Search Rank ...
- spezifische Attribute des Nuggets: Länge des Nuggets etc.
- Attribute der Nuggets auf den Nachbarpositionen im jeweilgen Retrieval-Ergebnis

4. **Merging**: sortiere Nuggets nach absteigender Relevanz. Wörtlich mit höherrangigen identische oder unter einem bestimmten threshold liegende Nuggets werden aussortiert. Die übrigen werden, zu einem neuen Pseudo-Dokument aggregiert, in den Quellenkorpus aufgenommen werden ausgefiltert.

Expansion III

Beispiel

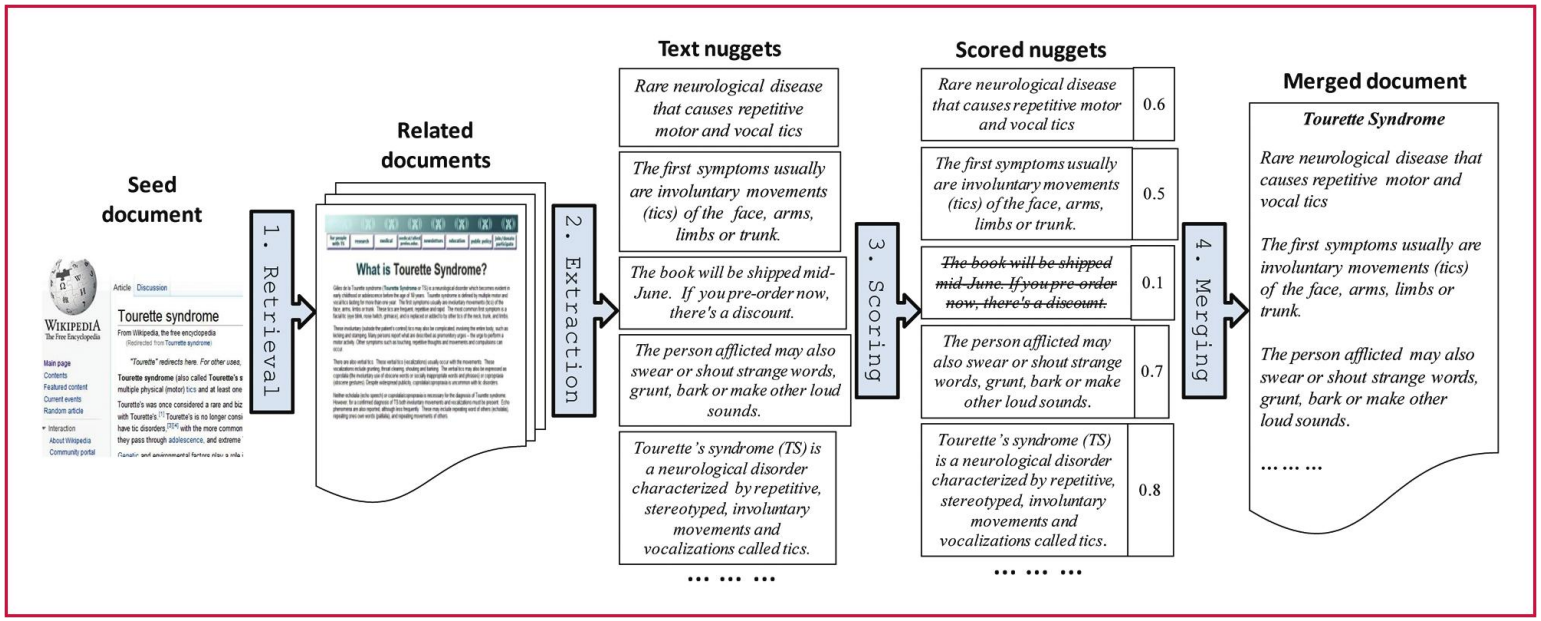


Figure 1: Quellenexpansion des Seed-Dokuments 'Tourette Syndrome' (Wikipedia) (Chu-Carroll et al., 2012)

Wikipedia-Corpus wächst während Expansion von 13 auf 59 GB

PRISMATIC

Ansatz

Zweistufiges Konzept:

1. flaches Wissen wird aus einem großen Textkorpus extrahiert und besteht aus **syntaktischen Beziehungen** innerhalb von Sätzen. Die syntaktischen Informationen sind in Frames und Slots repräsentiert
2. auf Grundlage der ermittelten Strukturen wird mithilfe von aggregierenden Statistiken und ontologischen Kategorisierungen **semantisches Wissen** gewonnen

PRISMATIC

Knowledge Base

Eine verifizierbare Wissensdatenbank (*knowledge base*) läszt sich unter Verwendung von Kategorien der Desktiptiven Logik (DL) als System aus folgenden Komponenten darstellen:

- Terminologische Axiome (ABox): Kategorien- und Rollenkonzepte definieren Entitätsklassen und binäre Beziehungen (intensionale Aussagen) Beispiel: $\textsc{Scientist} \sqsubseteq \textsc{Person}$, $\textsc{win}(\textsc{Scientist}, \textsc{Award})$
- Assertionale Axiome (TBox): Aussagen über Zugehörigkeiten von Entitäten/Instanzen zu Klassen und von Instanzpaaren zu Beziehungskonzepten: $\textsc{win}(\textsc{Einstein}, \textsc{NobelPrize})$ (extensional)

PRISMATIC

Knowledge Base II

Eine zur Wissensdatenbank erweiterte maschinenlesbare Ontologie erlaubt es, automatische Ergänzungen der vorhandenen Wissensbasis abzuleiten. So können aus extensionalen Aussagen ausreichender Menge allgemeine Axiome induziert und umgekehrt aus solchen wiederum assertionale Axiome über zuzuordnende Entitäten deduziert werden.

PRISMATIC

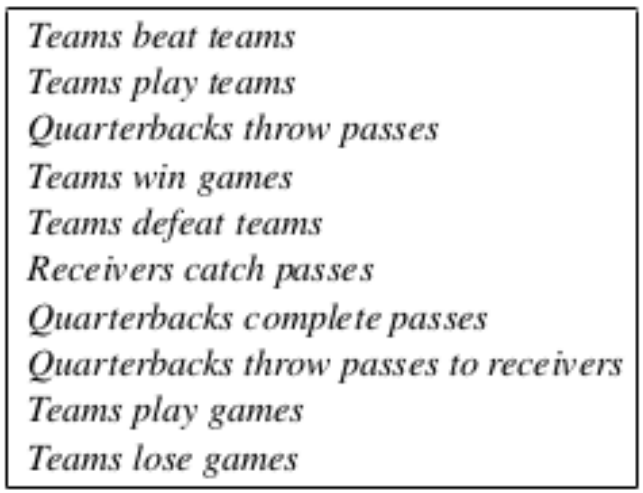
Knowledge Base IIb

PRISMATIC ist daran interessiert, gleichfalls über intensionale wie extensionale Prädikate verfügen zu können. So kann die Komponente einerseits durch Ableitung und Zufuhr von Texten ständig ihr Wissen vertiefen (*learning by reading*) als auch mithilfe konkreter extensionaler Aussagen über spezielle Individuen ("Einstein") zum Beispiel Antwortmöglichkeiten bewerten, sozusagen auf deren Wahrheitsgehalt überprüfen.

- Die nötigen Axiome gewinnt PRISMATIC allerdings durch die automatische Extraktion aus unstrukturiertem Text, bedarf also keinerlei vorgefertigter, domain-spezifischer Ontologien.

PRISMATIC

Automatische Extraktion von Wissen: Beispiel



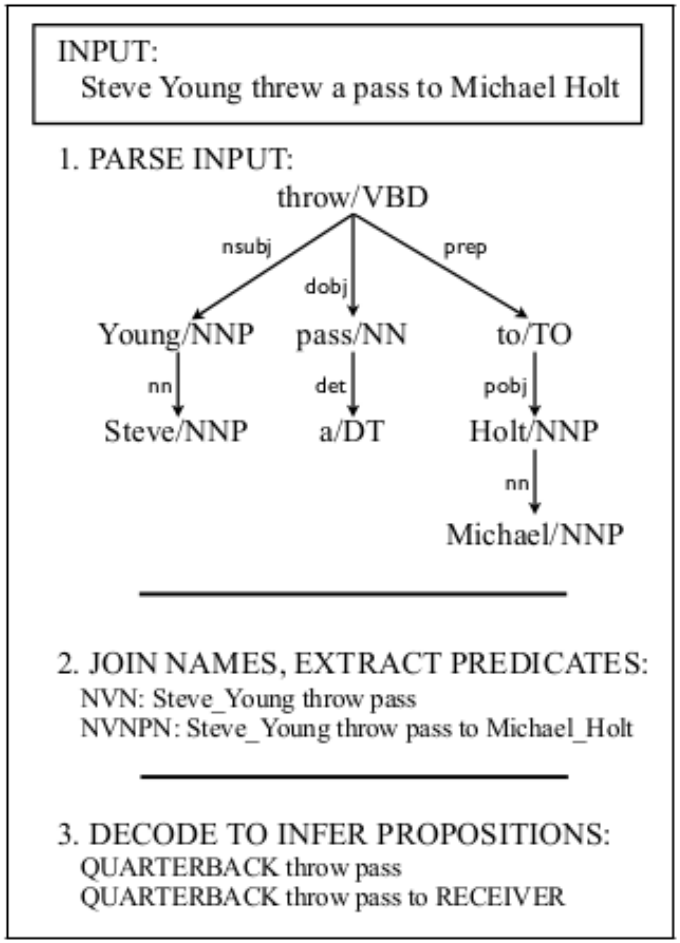


Figure 2: Aus Textressourcen gewonnene Beziehungsaxiome (Hovy et al., 2011)

PRISMATIC

Entwurf

Für das Verständnis des Aufbaus von PRISMATIC sind folgende Konzepte zentral:

- **Frame**: die grundlegende semantische Einheit. Frames beinhalten in Slots dargestellte Beziehungen von Textsegmenten
- **Slot**: binäre Relation, meistens entsprechend Abhängigkeitsbeziehungen aus dem dependency parser
- **Slot value**: entweder Satzsegment wie in Abängigkeitsbeziehung als Lemma oder zu einem solchen gehöriger Entitätentyp oder Querverweis auf andere Frames
- **Frame projection**: Funktion zur Filterung einer bestimmten Auswahl an Slots über mehrere oder alle Frames, die angewendet wird, um sich wiederholende Muster in Aussagen oder semantische Ableitungen zu ermitteln

PRISMATIC

Konstruktion

Die Initialisierung der Struktur von PRISMATIC durchläuft drei Phasen:

1. **Corpus processing**: Anwendung mehrerer NLP-Tools auf dem Textkorpus (dependency parser u.a.)
2. **Frame extraction**: Einträge und Ergänzungen aus den vom Parser erstellten dependency trees werden in Slot-Frame-Format gebracht
3. **Frame projection**: Frameprojektionen mit Aussicht auf hohe Aussagekraft werden ermittelt, Häufigkeit und Abhängigkeit von anderen Variablen wird errechnet

PRISMATIC

Überblick

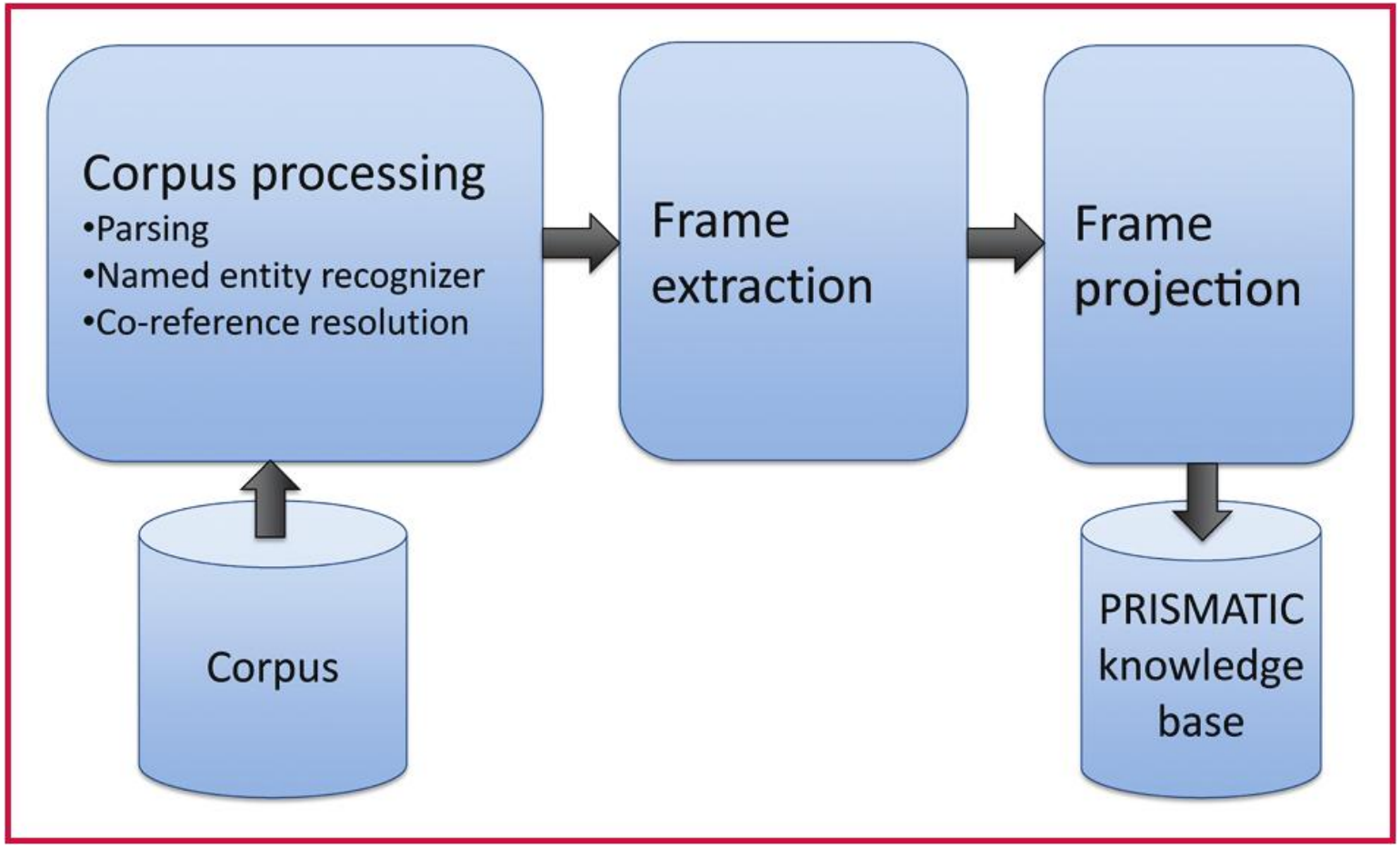


Figure 3: Pipeline zur Kostruktionsphase der Wissensdatenbank PRISMATIC (Fan et al., 2012)

PRISMATIC

Corpus Processing

Das Ziel des Corpus Processing ist, für die Weiterverwendung in PRISMATIC-Frames semantisch brauchbare Informationen zu gewinnen

- English Slot Grammar (ESG) als dependency parser
- regelbasierte Ermittlung von Koreferenzen auf Entitäten
- relation detection erlaubt Belegung von Slots wie `is_a` (Instanz)
- regelbasierter name entity recognizer (NER) klassifiziert Slot-Inhalte, um Aussagen auf begrifflicher (intensionaler) Ebene tätigen zu können

English Slot Grammar

Dependency Parser

Slot Grammar McCord (1990) verarbeitet unstrukturierten Eingabetext und gibt satzweise Abhängigkeitsbäume zurück. Nach der Segmentierung in und lexikalischer Analyse der Einzelworte beginnt die Verarbeitung der Grammatik, deren wesentliche Regeln beinhalten:

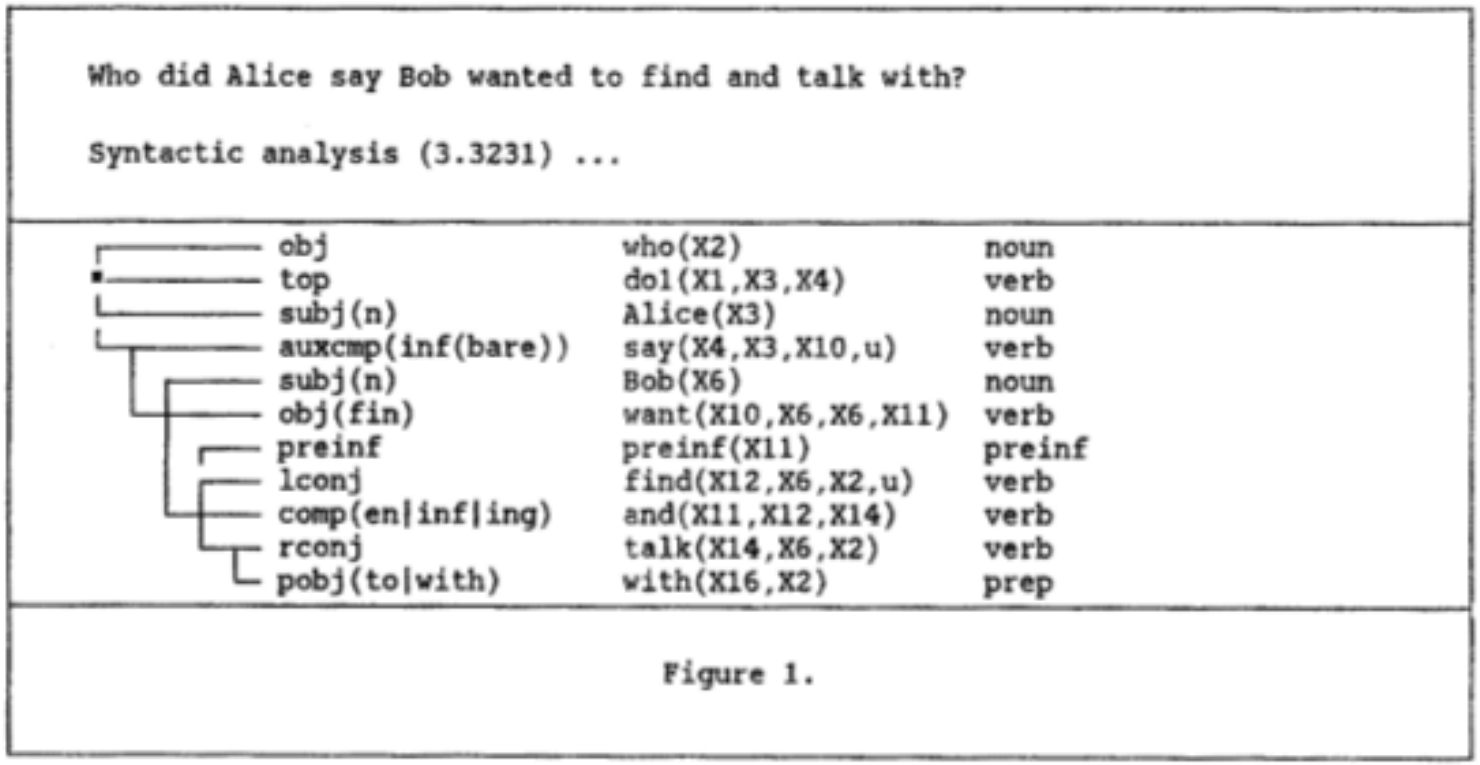
- für jeden POS: Deklarationen von anhängenden Slots
- die möglichen Belegungen für jeden Slot
- Zuordnung eines Slot entweder zum headword eines Knotens oder zu einem anderen Slot
- Definition von zum Slot gehörigen obligatorisch zu füllenden komplementären Slots (Objekt bei Präpositionen o.ä)

Der Parsing-Prozess wird bottom-up und von links nach rechts durchgeführt. Ein gelesenes Wort wird als valide angesehen, wenn alle verpflichtenden Slotzuordnungen erfüllt sind.

English Slot Grammar

Dependency Parser

Slot Grammars wie ESG geben Repräsentationen verarbeiteter Sätze in Baumstrukturen aus, denen Slotbelegungen einzelner Knoten ebenso wie die tiefe logische Struktur der Eingabe zu entnehmen sind.



Who did Alice say Bob wanted to find and talk with?

Syntactic analysis (3.3231) ...

obj	who(X2)	noun
top	do1(X1,X3,X4)	verb
subj(n)	Alice(X3)	noun
auxcmp(inf(bare))	say(X4,X3,X10,u)	verb
subj(n)	Bob(X6)	noun
obj(fin)	want(X10,X6,X6,X11)	verb
preinf	preinf(X11)	preinf
lconj	find(X12,X6,X2,u)	verb
comp(en|inf|ing)	and(X11,X12,X14)	verb
rconj	talk(X14,X6,X2)	verb
pobj(to|with)	with(X16,X2)	prep

Figure 1.

Figure 4: In ESG erzeugter Abhängigkeitsbaum (McCord, 1990)

English Slot Grammar

Dependency Parser II

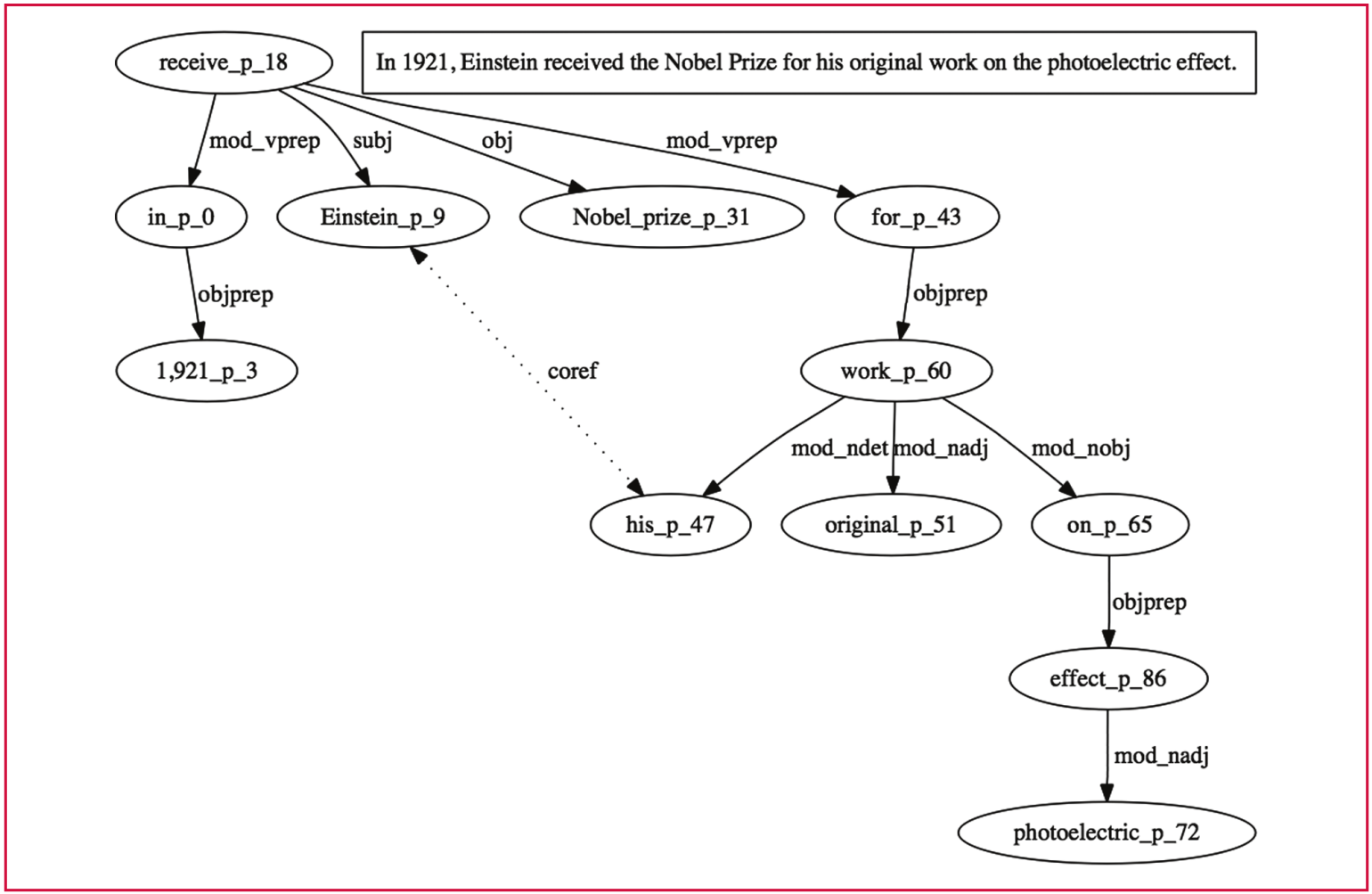


Figure 5: Im dependency parser erzeugte Ausgabe für den Satz 'In 1921, Einstein received the Nobel Prize for his original work on the photoelectric effect' (Fan et al., 2012)

PRISMATIC

Frame Extraction

Die in der ersten Phase erhobenen Beziehungen und Annotationen der Satzsegmente werden aus den Strukturen der NLP-Tools in die angestrebte Frame-Struktur überführt. Bei der im prinzip direkten Übernahme der Slotbelegungen der ESG-Struktur werden zwei Restriktionen angewandt:

- um sich auf die wesentlichen Beziehungen zu konzentrieren, werden nur Slots beachtet, die partizipierende semantische Rollen betreffen und die zum begrenzten Slot-Arsenal von PRISMATIC gehören
- ein in einem PRISMATIC-Frame gespeicherter Dependency-Teilbaum darf eine Höhe von 2 nicht überschreiten. So kann die Anzahl fehlerhafter Frames niedrig gehalten werden

Neben Beziehungen aus dem Parser können auch semantische Relationen wie `is_a` (ist ein/eine) oder Informationen wie aus dem NER oder andere Zusatzinformationen zu Wortkategorien gespeichert werden.

PRISMATIC

Frame Extraction II

Relation and description	*Example*
subj	Subject
obj	Direct object
iobj	Indirect object
comp	Complement
pred	Predicate complement
objprep	Object of the preposition
mod_nprep	*Bat Cave in Toronto is a tourist attraction*
mod_vprep	*He made it to Broadway*
mod_nobj	The object of a nomialized verb
mod_ndet	*City's budget was passed*
mod_ncomp	*Tweet is a word for microblogging*
mod_nsubj	*A poem by Byron*
mod_aobj	*John is similar to Steve*
isa	Subsumption relation
subtypeOf	Subsumption relation

Figure 6: In einem PRISMATIC-Frame mögliche Beziehungen/Slots (Fan et al., 2012)

PRISMATIC

Frame Extraction III

Frame01	
verb	receive
subj	Einstein
type	PERSON/SCIENTIST
obj	Nobel prize
mod_vprep	in
objprep	1921
type	YEAR
mod_vprep	for
objprep	Frame02

Frame02	
noun	work
mod_ndet	his/Einstein
mod_nobj	on
objprep	effect

Figure 7: Zwei aus einem ESG-Abhängigkeitsbaum erzeugte PRISMATIC-Frames (Fan et al., 2012)

PRISMATIC

Frame Projection

Durch den Entwurf geeigneter Projektionen auf den nun in PRISMATIC vorhandenen, syntaktische Beziehungen beinhaltenden Frames werden inhaltliche/semantische Aussagen abgeleitet. Eine Projektion kann ebenso lediglich eine Teilmenge an Slots aus den verfügbaren Frames filtern als auch Wertüberprüfungen von Slot values oder Attributen veranlassen. Nützliche Projektionen wären etwa:

- `S='Einstein', V, O` liefert alle Frames, in denen 'Einstein' als Subjekt vorkommt
- `S='Einstein', V='win', OT` gibt Hinweise darauf, von welchem Typ die Objekte in Frames, in denen 'Einstein' etwas gewinnt, in der Regel sind
- `N, IsA` beantwortet, von welchen Oberbegriffen die Substantive in ihren jeweiligen Frames Instanzen bilden

PRISMATIC

Frame Projection: Strategien

Die genannten Projektionen sind hinsichtlich des Aufbaus der Wissensdatenbank von besonderem Interesse:

- Durch Projektion auf Verben (`V`) und Objekttypnen (`OT`) können terminologische Axiome extrahiert werden (win(Scientist, Award)), wenn `IsA`-Beziehungen ebenfalls berücksichtig werden
- Projektionen, die als Nominative erkannte Subjekte und Objekte extrahieren, liefern assertionale Axiome, die von Watson als Beleg für Anwortkandidaten herangezogen werden können

PRISMATIC

Aggregate Statistics

Um die in Frameprojektionen gewonnenen Aussagen bewerten zu können, brauchen wir Angaben zur Häufigkeit, Verhältnissen zu anderen Größen und andere Hinweise auf Relevanz, z.B:

- Häufigkeit: die absolute Anzahl an Frames, die eine Projektion erfüllen
- Bedingte Wahrscheinlichkeit: $P(A \mid B)$ für die Projektionen A und B
- Normalized pointwise mutual information: Kookkurenz von zwei Projektionen, normalisiert auf Framegrösze
- Korrelation, Abhängigkeiten etc.

PRISMATIC

Anwendungen

Vorschlagen von Antwortkandidaten:

- Aufgrund ihres Wissens über die Wahrscheinlichkeit des Zusammenhangs zwischen zwei Entitäten oder Konzepten liefert PRISMATIC bessere Antwortkandidaten für gegebene Fragekategorien und -LATs als andere Komponenten in Watson: eine von 57 von PRISMATIC vorschlagenen Antworten ist richtig, die übrigen Komponenten liegen bei 134 Vorschlägen einmal richtig. Insgesamt liefert PRISMATIC gute 40% der richtigen Anworten.

PRISMATIC

Anwendungen

Antwortkandidaten mit dem LAT einer Frage abgleichen:

- Bei gegebenem LAT kann PRISMATIC anhand der Relevanz von Frames, die einen Antwortkandidaten als Instanz dieses LAT ausweisen, die Eignung der Antwort beurteilen. Auch hierbei erzielt PRISMATIC bessere Ergebnisse als die meisten eingesetzten Komponenten

PRISMATIC

Anwendungen

Ermitteln eines LAT auf Grundlage eines vorhandenen Fokus

- Kann zu einer Frage lediglich ein Fokus mit mangelnder Aussagekraft ('it' o.ä.) ermittelt werden, ist PRISMATIC in der Lage, anhand der syntaktischen Verknüpfungen des Fokus den LAT für naheliegendere POS auszugeben (im Beispiel für das Verb, dessen Object/Subjekt der Fokus ist)

PRISMATIC

Anwendungen

Fehlende Verbindungsglieder zwischen Begriffen aus Clues und Antwortkandidaten aufspüren

- Liefert eine Frage keine Hinweise auf die Antwort, die etwa in text-passage-search brauchbar wären, kann PRISMATIC ausgehend von vorhandenen Stichworten mehrere Kanten semantischer Verknüpfungen abschreiten und damit wahrscheinliche Antworten auch über zwei Ecken ausfindig machen.

PRISMATIC

Anwendungen

Weitere Lösungen, die mit terminologischen und assertionalen Axiomen und dazugehörigen statistischen Auswertungen realisierbar sind:

- Ableiten des wahrscheinlichsten infrage kommenden Typs einer named entity (NER)
- Auflösen von Koreferenzen mithilfe vorhandenen Wissens über wahrscheinliche Entitätstypen
- Auflösen von Mehrdeutigkeiten bei Begriffen

PRISMATIC

Probleme und Verbesserungsmöglichkeiten

Trotz des ansehnlichen Beitrags, den PRISMATIC an präzisen Ergebnissen in Antwortvorschlägen und -Bewertung leistet, gibt es noch Unzulänglichkeiten und Nachholbedarf:

- Immer noch Wissenslücken aufgrund von Leerstellen im Textkorpus
- Fehlendes Einbeziehen des übergeordneten Kontext erschwert die Auswertung von Aussagen bei Begriffen mit mehreren Bedeutungen
- Durch Verwertung der *confidence*-Parameter aller zugrundeliegenden NLP-Komponenten könnte eine höhere Präzision erreicht werden

Literatur

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P. (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge university press.

Chu-Carroll, J., Fan, J., Schlaefer, N., and Zadrozny, W. (2012). Textual resource acquisition and engineering. *IBM Journal of Research and Development*, 56(3):4.

Fan, J., Ferrucci, D., Gondek, D., and Kalyanpur, A. (2010). Prismatic: Inducing knowledge from a large scale lexicalized relation resource. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 122–127. Association for Computational Linguistics.

Fan, J., Kalyanpur, A., Gondek, D. C., and Ferrucci, D. A. (2012). Automatic knowledge extraction from documents. *IBM Journal of Research and Development*, 56(3):5.

Hovy, D., Zhang, C., Hovy, E., and Peas, A. (2011). Unsupervised discovery of domain-specific knowledge from text. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1466–1475.

Krötzsch, M., Simancik, F., and Horrocks, I. (2012). A description logic primer. *arXiv:1201.4089*.

McCord, M. C. (1990). Slot grammar: A system for simpler construction of practical natural language grammars. In *Proceedings of the International Symposium on Natural Language and Logic*, pages 118–145, London, UK, UK. Springer-Verlag.

Welty, C., Fan, J., Gondek, D., and Schlaikjer, A. (2010). Large scale relation detection. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 24–33. Association for Computational Linguistics.