Computerlinguistische Methoden für die Rechtsterminologie

Stephan Walter

Fachrichtung 4.7 Allgemeine Linguistik
Gebäude C 7.1, C 7.2 + C 7.4
Postfach 15 11 50
Universität des Saarlandes
Saarbrücken
stwa@coli.uni-sb.de

Schlagworte: Computerlinguistik, Terminologie, Rechtssprache, Definitionen, Ontologien, linguistisch gesteuerter Informationszugriff

Abstract: Dieser Beitrag stellt das Projekt CORTE (Computerlinguistische Methoden für die Rechtsterminologie)[1] vor, das sich mit der Entwicklung computerlinguistischer Verfahren zur automatischen Extraktion und Verarbeitung von Definitionen in deutschen Gerichtsentscheidungen beschäftigt. Mittels eines robusten, semantisch orientierten Parsingsystems werden linguistische Strukturen für den Text von Urteilsbegründungen ermittelt, in einem XML-Format abgelegt und dann nach sprachlichen Mustern durchsucht, die charakteristisch für Definitionen sind. Die Fundstellen werden dann automatisch in die strukturellen Bestandteile einer Definition (Definiendum, Definiens usw.) zergliedert. Durch die Ergebnisse kann beispielsweise eine gezielte Informationssuche auf begrifflicher Basis und das textbasierte Update juristischer Ontologien unterstützt werden.

1. Motivation

Bisherige Ontologien der Rechtssprache erfassen häufig ein „Upper model“, das neben generischen Common Sense-Begriffen Grundbegriffe der Rechtsordnung (vgl. Breuker und Hoekstra 2004) und einige in Gesetzestexten durch Legaldefinitionen festgelegte zentrale Rechtsbegriffe umfassen kann. Solche Ontologien enthalten in der Regel hand-kodiertes Expertenwissen. Wenn die repräsentierten Konzepte sich dagegen nicht ausschließlich auf die oberen ontologischen Ebenen beschränken sollen, ist automatische Unterstützung bei der Ontologieerstellung von großem Interesse. Sprachtechnologische Verfahren können hierzu einen wichtigen Beitrag leisten.

---

[1] DFG PI 154/10-1. Siehe http://www.coli.uni-saarland.de/projects/corte/

Dieser Beitrag stellt das Projekt CORTE (Computerlinguistische Methoden für die Rechtsterminologie)[1] vor, das sich mit der Entwicklung computerlinguistischer Verfahren zur automatischen Extraktion und Verarbeitung von Definitionen in Texten von Gerichtsentscheidungen beschäftigt. Ausgehend von einem zunächst weit gefassten, in erster Linie syntaktisch charakterisierten Definitionsbegriff untersucht das Projekt außerdem, welche Definitionstypen nach juristischen und linguistischen Gesichtspunkten unterschieden werden können und wie juristische Kategorien bei der Formulierung von Definitionen sprachlich umgesetzt werden.

2. Definitionen in Entscheidungstexten

Viele Rechtsbegriffe werden durch Definitionen im Gesetzestext nicht ein für alle Mal erschöpfend bestimmt, sondern in ihrer Anwendbarkeit nur partiell festgelegt. Über die Anwendbarkeit auf den konkreten Rechtsfall kann nur nach weiterer Präzisierung entschieden werden. Im Zuge dieser Präzisierung werden Gesetzesbegriffe auf den Fall hin konkretisiert, Hilfsbegriffe eingeführt und oft auch alltagssprachliche Begriffe definitorisch festgelegt. Nicht nur in Gesetzestexten, sondern auch in Gerichtsentscheidungen sind deshalb (insbesondere in den Entscheidungsgründen) verschiedene Arten von Definitionen wesentliche Bestandteile.[2] Obwohl formal nicht bindend, haben die begrifflichen Festlegungen in Urteilsbegründungen eine hohe faktische Bindungswirkung für die anschließende Rechtsprechung. Wenn in Ontologien auch die in solchen Definitionen betroffenen „unteren Ebenen“ rechtssprachlicher Begriffe dargestellt werden sollen, sind Gerichtsentscheidungen daher - neben Gesetzestexten - eine zweite wichtige Wissensquelle. Für die Nutzung dieser Wissensquelle ist die Unterstützung durch automatische Verfahren kaum verzichtbar, allein aufgrund der enormen Menge des zu berücksichtigenden Textmaterials und der Notwendigkeit, ständig neue Entscheidungen in die Textgrundlage einzubeziehen.

---

[2] Ein “klassisches” Beispiel der Rechtsliteratur für die sukzessive Präzisierung von Rechtsbegriffen ist die BGB-Definition von “Sache” als beweglicher Gegenstand. Aufgrund der Legaldefinition allein war nicht zu klären, ob es sich bei der unbefugten Entwendung von elektrischem Strom um Diebstahl handelt. Diese Frage wurde erst durch definitorische Festlegung in einem Gerichtsurteil - negativ - entschieden (RG, IV. Strafsenat, 20.10.1896 g. W. Rep. 2609/96, RGSt. Bd. 29, S.111f.).

3. Computerlinguistische Analyse von Urteilstexten

Bei den Arbeiten im Projekt CORTE können wir auf den Datenbestand der Firma juris GmbH, Saarbrücken, zurückgreifen, der derzeit über 8 Millionen juristische Dokumente umfasst. Zurzeit basieren unsere Arbeiten auf einem Bestand von 6000 Urteilen im Umweltrecht (das entspricht 237 935 Sätzen in Leit- und Orientierungssätzen sowie Entscheidungsgründen). Die Dokumente liegen als Textdateien vor, die zunächst vorverarbeitet werden müssen (z.B. XML-Strukturierung in Anlehnung an den Saarbrücker Standard für Gerichtsentscheidungen (Gantner und Ebenhoch 2001), Markierung von Satzgrenzen). Für die linguistische Analyse kommt dann ein robustes, semantisch orientiertes Parsingsystem zum Einsatz, das in dem BMBF-geförderten Projekt COLLATE[3] entwickelt wurde. Das System wird in (Braun 2003) genauer beschrieben. Es wurde zunächst zur Analyse von Zeitungstext eingesetzt und dann für CORTE an die spezifischen Merkmale juristischer Texte angepasst (unter anderem Satzkomplexität, spezifische Named Entities wie z.B. Angaben zu Belegstellen). Komponenten des Systems wurden auch im Rahmen einer empirischen Untersuchung zur Verständlichkeit von Entscheidungen des Bundesverfassungsgerichts eingesetzt (Braun et al. 2005).

Das Parsingsystem kombiniert verschiedene flache Verfahren in einer Kaskade. Dadurch ist es möglich, auch für so komplexen Text wie Gerichtsurteile robust zu einer verhältnismäßig tiefen Analyse zu gelangen. Grundlage der Analyse ist die Ermittlung der topologischen Satzstruktur der Eingabesätze. Diese wird durch weitere linguistische Information (Morphologie, Named Entities und Nominalphrasen in den topologischen Feldern) angereichert. Auf dieser Basis wird dann für jeden Satz eine teil-resolvierte Prädikat-Argument-Struktur (Preds, *partially resolved dependency structure*) erzeugt, in der die Beziehungen der ermittelten Bestandteile untereinander sowie zum Prädikat des Satzes dargestellt werden.

Ambiguitäten, die zum Zeitpunkt der Analyse nicht auflösbar sind (zum Beispiel die Anbindung von Präpositionalphrasen), können in diesen Strukturen in vielen Fällen unterspezifiziert dargestellt werden. Es werden dabei nur die gesicherten gemeinsamen Teile der möglichen Lesarten festgelegt. Ungesicherte Strukturen werden als solche markiert und können von späteren Komponenten weiter behandelt werden. Durch diese Technik kann das System trotz der massiven

[3] Computational Linguistics and Language Technology in Real Life Applications

Ambiguität natürlicher Sprache auch ohne die extrem verarbeitungsaufwändige Verfolgung alternativer Lesarten robust ein Kern an semantischer Information für den zu analysierenden Text ermitteln. Dies hat u.a. den Vorteil dass für eine anschließende semantische Suche, wenn diese sich auf die eindeutige Teilinformation beschränkt, nur eine Struktur pro Satz durchsucht werden muss.

Abb. 1 zeigt eine Visualisierung der für Satz (1), eine Definition in einem Urteil aus dem Verwaltungsrecht, erzeugte Preds:

> *(1) Bei einem Einfamilienhaus liegt ein mangelhafter Schallschutz dann vor, wenn die Haustrennwand einschalig errichtet wurde (…).*[4]

```
vorliegen[verb, sg, ind]
  -PPMod->bei[praep, dat]
            -Arg->*ein#familien#reihen#haus[noun]
  -DSub->*schall#schutz[noun]
           -Mod->mangelhaft[adj]
  -Mod->wenn[subj]
          -Arg->errichten[verb, sg, ind, pres, pass]
                  -DObj->*haus#trenn#wand[noun, defArt]
                  -Mod->einschalig[adj]
  -Mod->dann[adv]
```

Abb. 1: Preds für Satz (1)

Die Repräsentation abstrahiert u.a. über verschiedene Aspekte der Oberflächenstruktur des Satzes (z.B. das Passiv im Nebensatz), zeichnet Prädikate (z.B. vorliegen im Hauptsatz) aus und klassifiziert Bestandteile des Satzes nach ihren Rollen bezüglich des Prädikats (im Beispiel tiefes Subjekt DSub und Objekt DObj, Modifikatoren Mod und PPMod).

4. Definitionsextraktion

Der Einsatz des beschriebenen Parsers dient zunächst zur Verbesserung der Suche nach Definitionen. Die Möglichkeit, neben der textuellen Oberflächenform der Datenbestände auch deren linguisti-

[4] OLG Stuttgart, 22.11.1995, 1 U 199/93, juris

sche Struktur zu berücksichtigen, erlaubt die Formulierung sehr viel einfacherer und präziserer Suchmuster. Gegenüber einer auf regulären Ausdrücken für unanalysierten Text basierenden Vorgehensweise kann ein Suchmuster z.B. über verschiedene Stellungsvarianten abstrahieren und (wie dargestellt) Aktiv- und Passiv-Varianten einer Formulierung abdecken. Modusinformation erlaubt beispielsweise das Ausfiltern in indirekter Rede wiedergegebener Meinungen, denen sich das Gericht nicht angeschlossen hat.

Außerdem ist durch die linguistische Analyse eine weitere Verarbeitung von Fundstellen möglich. So können auf der Basis der Prädikat-Argument-Struktur Fundstellen in Definiens (im Beispiel das tiefe Subjekt DSub) und Definiendum (im Beispiel Mod mit Subjunktion *wenn*) segmentiert werden. Oftmals handelt es sich zudem bei definitorischen Festlegungen in Urteilen nicht um vollgültige Definitionen. Es werden z.B. nur partielle oder explizit auf bestimmte Rechts- oder Sachbereiche eingeschränkte Präzisierungen vorgenommen. Dies ist meist an Modifikatoren zu erkennen, die ebenfalls in der Ergebnisstruktur identifiziert werden können (im Beispiel wird die Definition durch eine modifizierende Präpositionalphrase auf den Sachbereich „Einfamilienhäuser“ eingeschränkt).

Das in Abb. 2 gezeigte Suchmuster dient zur Extraktion und Verarbeitung von Definitionen des in Beispiel (1) dargestellten Typs (*… liegt vor, wenn…*):

```
<pattern>
  number=7
  description=liegt vor + wenn-Nebensatz
  query=sent/parse/preds/word[@stem="vorliegen" and INDPRES
          and WENN]]
  filters=definite
  definiendum=DSub
  definiens=WENN/arg/word
  geltungsbereich=PPMOD{PREP%bei}
</pattern>
```

Abb. 2: Suchmuster für Definitionen des Typs *… liegt vor, wenn…*

Die unter den einzelnen Schlüsseln angegebenen Ausdrücke werden zunächst zu Pfadangaben in XPath-Syntax expandiert. Diese werden dann auf XML-Ausgabestrukturen des oben beschriebenen Parsers evaluiert. Durch Auswertung des unter `query` angegebenen Ausdrucks werden so im vorliegenden Fall alle Sätze extrahiert, deren

Prädikat *vorliegen* (im Indikativ Präsens) durch einen *wenn*-Nebensatz ergänzt wird. Die Angabe unter `filter` bestimmt, dass Sätze ausgeschlossen werden, deren Subjekt definit ist. Durch Verwendung des bestimmten Artikels wird nämlich häufig angezeigt, dass der Satz auf den konkreten Fall referiert, also keine Definition enthält, sondern eine Subsumtion. In den gefundenen Sätzen wird dann das Subjekt als Definiendum, der Inhalt des Nebensatzes als Definiens und der Inhalt einer potentiell vorhandenen *bei*-Präpositionalphrase als Angabe eines Geltungsbereichs ausgezeichnet.

Aktueller Schwerpunkt der Arbeiten in CORTE ist die Erstellung solcher Suchmuster und Verarbeitungsregeln und die Erfassung der Qualität der Suchergebnisse, zudem die strukturierte Ablage und Abfragemöglichkeiten zu zielgerichteten Zugriff auf die Fundstellen. Ein weiteres Thema ist die Nutzung statistischer Information zur Erzeugung einer Rangfolge unter den Fundstellen. Neben sprachlich motivierten statistischen Maßen (*tf-idf*-Maß[5] zur Ermittlung von besonders interessanten Fundstellen, Kollokationsmaße[6] zur Ermittlung terminologisch relevanter Wortkombinationen) sollen hierbei auch spezifisch juristische Informationen berücksichtigt werden (z.B. Hierarchie der Gerichte, Zitationshäufigkeit. Siehe auch unter 5.). Geplant ist zudem, den rein regelbasierten Ansatz bei der Extraktion um eine Komponente zu erweitern, die verschiedene Indikatoren für das Vorliegen einer Definition unter Berücksichtigung maschinell gelernter Gewichte kombiniert.

5. Ausblick

Im weiteren Verlauf des Projekts sollen Verfahren entwickelt und korpusbasiert erprobt werden, um Zusammenhänge zwischen Begriffen im Text von Fundstellen zu ermitteln. Ziel ist, zum einen Bedeutungsrelationen (wie Hyponymie, Antonymie), zum anderen Information über juristisch-übergreifende Zusammenhänge (Fallgruppen, Definitionstypen, Präzisierungen oder widerstreitende Meinungen, Revision von Bedeutungsfestlegungen) in möglichst vielen Fällen automa-

---

[5] Term frequency-inverse document frequency. Dieses Maß berücksichtigt sowohl die Häufigkeit eines Terminus innerhalb eines Dokuments als auch seine Verteilung innerhalb der gesamten Dokumentkollektion, um eine Aussage über seine Relevanz für das betreffende Dokument zu machen.

[6] Solche Maße geben für Wortpaare eine Anhaltspunkt, ob es sich um potentiell terminologisch relevante Fügungen oder um zufällige Kombinationen handelt. Dabei werden die Häufigkeiten der einzelnen Bestandteile sowie der Kombination in der Dokumentkollektion berücksichtigt.

tisch zu erfassen. Neben linguistischer Information im engeren Sinne (z.B. explizite begriffliche Unterordnung, Ermittlung der Kopfnomen bei Nominalkomposita, Auswertung von Partikeln wie noch, erst, schon usw.) soll dabei auch auf nicht-linguistische strukturierende Elemente zurückgegriffen werden (z.B. Angaben zu Zitaten und Normen).

Durch die Nutzung solcher Information können extrahierte Fundstellen geordnet und gewichtet präsentiert werden, um dem juristischen Benutzer die zielgerichtete Navigation durch große Datenbestände zu ermöglichen. Darüber hinaus können aber auch Vorschläge zur Aufnahme und Einordnung von neuen Begriffen in terminologische Ressourcen erzeugt und damit das Update von Ontologien hinsichtlich aktueller Rechtsprechung unterstützt werden. Bestehenden formalen Ontologien liegen zumeist Varianten beschreibungslogischer Formalismen zu Grunde (vgl. Baader et al. 2003). Ein linguistisch fundiertes Modell natürlichsprachlicher Definitionen eröffnet die Perspektive, extrahierte Definitionen in einen solchen Formalismus zu übersetzen und damit die Integration des extrahierten Wissens in formale Ontologien ebenfalls (zumindest teilweise) zu automatisieren.

6. Literatur

Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., und Patel-Schneider, P.F. (Hg.) The Description Logics Handbook: Theory, Implementations, and Applications. Cambridge University Press, 2003.

Braun, C. Parsing German text for syntacto-semantic structures. In Prospects and Advances in the Syntax/Semantics Interface, Lorraine-Saarland Workshop Series, Nancy, France, 2003, 99-102.

Braun, C., Hansen-Schirra, S., Kunz, K. und Neumann, S., The syntactic complexity of German legalese - An empirical approach. International Association of Forensic Linguistics, 7th Biennal Conference on Forensic Linguistics/Language and Law, Cardiff, UK, 1.-4. Juli 2005

Breuker, J. und Hoekstra, R. Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law. In Proceedings of EKAW Workshop on Core ontologies. CEUR, 2004.

Gantner, F. und Ebenhoch, P. Der Saarbrücker Standard für Gerichtsentscheidungen (kommentierte Fassung), JurPC Web-Dok. 116/2001