Klassifizierung von Korpora für die Rechtsterminologie

Dr. Oliver Streiter, RA Leonhard Voltmer

Klassifizierung von Korpora für die Rechtsterminologie

RA Leonhard Voltmer, Dr. Oliver Streiter,

EURAC research

EUROPÄISCHE
AKADEMIE

ACCADEMIA
EUROPEA

EUROPEAN
ACADEMY

BOZEN - BOLZANO

Oktober 2003

Bestellungen bei:
Europäische Akademie Bozen
Drususallee, 1
39100 Bozen - Italien
Tel. +39 0471 055055
Fax +39 0471 055099
E-mail: press@eurac.edu
Verantwortlicher Direktor: Stephan Ortner

Per ordinazioni:
Accademia Europea Bolzano
Viale Druso, 1
39100 Bolzano - Italia
Tel. +39 0471 055055
Fax +39 0471 055099
E-mail: press@eurac.edu
Direttore responsabile: Stephan Ortner

Klassifizierung von Korpora für die Rechtsterminologie

Voltmer/ Dr. Streiter

Vorstellung des Forschungsbeitrags

Die Fortschritte in der Computerlinguistik und insbesondere der Korpuslinguistik eröffnen der Terminologieforschung ungeahnte neue Möglichkeiten zur effizienteren Erstellung und Darstellung von Terminologie. Korpora versprechen das automatisierte Auffinden von Termkandidaten, ihrer Übersetzungen, Eigenschaften und Kontexte. Auch die Beschreibung und Verknüpfung von Terminologie könnte neue Anstöße bekommen.

Vor allem könnte die Anpassung der Darstellung an die Bedürfnisse unterschiedlicher Nutzer von Korpora profitieren. Voraussetzung für jeden ernst zu nehmenden korpusterminologischen Ansatz ist natürlich die Erstellung eines gewichteten, annotierten und klassifizierten Korpus. Diese Aktivität erfordert Wissen und Aufwand, die viele terminologische Projekte auf den ersten Blick zu überfordern scheint. Daher wird hier eine Methode vorgestellt, mit der rechtsterminologisch relevante Dokumente semiautomatisch erlangt und klassifiziert werden können. Der semiautomatische Ansatz stellt keine großen Anforderungen an den Entwicklungsgrad der Terminologie, so dass er auch für Projekte in Frage kommt, die erst in Planung oder im Aufbau sind. Darüber hinaus kann der Aufwand den Möglichkeiten des einzelnen Projekts angepasst werden, so dass sich kleine Projekte stärker auf die Automatisierung und größere Projekte mehr auf die Präzision der Methode stützen können.

Dazu werden hier grundlegende Dimensionen und Klassen zur Einteilung rechtlicher Dokumente vorgestellt. Anschließend wird die Validität der wissenschaftlichen Methode mit Experimenten untersucht.

Forschungshintergrund

Der Fachbereich Sprache und Recht der EURAC betreibt angewandte Forschung im Schnittbereich von Linguistik und Rechtspraxis. Hintergrund für die terminologische Arbeit ist die rechtliche Gleichstellung der deutschen und italienischen Sprache in Südtirol und zusätzlich des Ladinischen in einigen Gemeinden.

Ein Forschungsschwerpunkt ist die Entwicklung deutschsprachiger Rechts- und Verwaltungssprache im italienischen Rechtssystem unter Beachtung der deutschen Rechtssprache in Österreich, der Schweiz und Deutschland. In geringerem Maß wird die Verwendung der ladinischen Sprachen im Gader-, Grödner- und Fassatal in Recht und Verwaltung beschrieben und unterstützt. Die deskriptiven und präskriptiven Daten werden Beamten, Übersetzern und allgemein der Öffentlichkeit kostenlos im Internet zur Verfügung gestellt.

Korpora in der Rechtsterminologie

Im letzten Jahrzehnt sind große elektronische Korpora allgemein verfügbar geworden und die Computerlinguistik hat sich mit großem Erfolg korpuslinguistischen Ansätzen zugewandt. Anwendungen sind part-of-speech tagging[1], parsing[2], computerunterstützte Übersetzung[3] und Disambiguierung[4]. Auch die Translatologie hat sehr von Translation Memories[5], Termdatenbanken und Termextraktions-

---

1 Brill, E., Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging, Computational Linguistics, 1995.

2 Charniak, E., Tree-bank grammars, in 13th National Conference on Artificial Intelligence, AAAI-96, 1996, S. 1031-1036.

3 Furuse, O. and Iida, H., An example-based method for transfer-driven Machine Translation, in The Third International Conference on Theoretical and Methodological Issues, Empiristic vs. Rationalist Methods in MT, Montréal, 1992.

4 Disambiguierung, Desambiguierung, Entambiguisierung oder Monosemierung bezeichnen die Erlangung sprachlicher oder außersprachlicher Kontextinformation zur Reduzierung lexikalischer oder struktureller Mehrdeutigkeit eines sprachlichen Ausdrucks. Ein Beispiel für den Vergleich mehrerer Methoden zur Disambiguierung beschreiben Dominic Widdows, Stanley Peters, Scott Cederberg, Chiu-Ki Chan, Diana Steffen, Paul Buitelaar in „Unsupervised Monolingual and Bilingual Word-Sense - Disambiguation of Medical Documents using UMLS“ in Natural Language Processing in Biomedicine, ACL workshop proceedings, S. 9-16, 2003, http://citeseer.nj.nec.com/article/widdows03unsupervised.html .

5 Carl, M., Schaible, J., Pease, C., Enhancing translation memory (TM) technologies with linguistic intelligence, MULTI-DOC Deliverable D4.1, Europäische Kommission, Luxemburg, 1998.

werkzeugen[6] sowie von den weit verbreiteten Terminologieverwaltungsprogrammen[7] profitiert.

In der Terminologie sind Korpora mehr als bloßes Ausgangsmaterial für eine Termextraktion oder einen Kontextnachweis:

1. Korpora sind wichtigstes Mittel zum Verständnis eines Fachgebiets, sowohl in sprachlicher Hinsicht für den Fachterminologen als auch in inhaltlicher Hinsicht für die Fachleute
2. Korpora sind authentische Äußerungen und enthalten daher Informationen, die der Zielgruppe von Terminologie nützlich sein können.
3. Korpora enthalten implizit Informationen über die Begriffe und ihre Verwendung wie idiomatische Redewendungen, Jargon, Wortwertigkeiten und Kollokationen, die meist vollständiger und aktueller sind als das in Wörterbüchern oder Termdatenbanken der Fall sein kann.
4. Korpora zeigen Bedeutungsänderungen in verschiedenen Kontexten auf und sind damit ein erster Schritt in Richtung einer Begründung oder Theoriebildung der Bedeutungsverschiedenheit.
5. Korpora decken auch Benennungsvarianten und Synonyme auf.

Als weiterer Punkt kommt hinzu, dass die Benutzer eine Termdatenbank oft unterschiedlich nutzen. Ein deutschsprachiger Jurist sucht andere Informationen als ein italienischsprachiger Übersetzer, so dass auch andere Darstellungsweisen als wünschenswert erscheinen. Schnell stößt die bloß veränderte Anzeige von terminologisch aufgearbeitetem Material aber auf ihre Grenzen, da eben diese Vorauswahl durch den Terminologen auf bestimmten Annahmen über die Zielgruppe beruht.[8] Die Anpassungsfähigkeit an die verschiedenen Benutzer, ein Qualitätsmerkmal im Fernunterricht und Multimediaanwendungen mit unterschiedlichen Benutzergruppen, wird heute im Wesentlichen durch geschickten Einsatz von annotierten und klassifizierten Korpora erreicht.[9]

---

6 Z.B. Bonnet, E., Gaussier, E., Langé, J.-M., A method for automatic extraction of terms from bilingual corpora, in AVIGNON-94, 1994.

7 Schmidt-Wigger, A., Building consistent terminologies, in Proceedings of COMPUTERM '98, 1998.

8 Faber, P., Lopés Rodriguez, C. I. and Tercedor Sánchez, M. I., Utilización de técnicas de corpus en la representatición del conocimiento médico, Terminology, 2002, 7(2), S. 167-197.

9 De Carolis, B., de Rosis, F. and Pizzutilo, S., Generating user-adapted hypermedia from discourse plans, in Lenzerini, M., ed., LNAI 1321, Springer, 1997.

Je größer Korpora sind, umso mehr Informationen enthalten sie.[10] Das größte Korpus ist das Internet. Das gilt mittlerweile wohl auch für rechtliche Informationen, soweit man rechtshistorische Untersuchungen ausklammert, weil die allgemeine Zugänglichkeit von Normen eine Voraussetzung für gelungene Verhaltenssteuerung ist. Normgeber versuchen daher, ihre Normen über den kostengünstigen Weg offizieller Informationsdienste im Internet zu verbreiten. Auch juristische Verlage vertreten ihre Produkte längst in elektronischer Form und sind aus Werbegründen oft aufgeschlossen gegenüber Forschungseinrichtungen, die ohnehin nicht Zielgruppe des oft kostenpflichtigen Angebots sind. Die Voraussetzungen für korpusgestützte Terminografie, nämlich große, frei erhältliche Datenmengen, sind damit gegeben.

Die elektronischen Daten können entweder in ein lokales Korpus eingebracht werden oder an ihrem ursprünglichen Ort belassen und durch spezielle Indizierung und Metasuchwerkzeuge angesteuert werden. Beide Vorgehensweisen haben Vor- und Nachteile. Das lokale Korpus ist sicher besser zu kontrollieren, d.h. es kann genauer annotiert, klassifiziert und allgemein aufbereitet werden, was auch Voraussetzung für die Ausgewogenheit eines Korpus ist. Das externe Korpus kann hingegen sehr viel größer sein und trotzdem, ohne eigenen Aufwand, stets aktuell gehalten werden.

An der EURAC werden daher die Vorteile beider Ansätze vereint. Das Korpus soll vor allem zur Termerstellung und Termdarstellung verwendet werden. Seit 2001 wird dazu eine korpuslinguistische Terminografieplattform BISTRO[11] aufgebaut. Entscheidend für die Qualität der terminografischen Ergebnisse war der Aufbau des Korpus, der eine geschickte Weiterentwicklung erfordert, aber auch erleichtert.

---

10 Michele Banko, Eric Brill, „Scaling to Very Very Large Corpora for Natural Language Disambiguation“, Meeting of the Association for Computational Linguistics", S. 26-33, 2001. http://citeseer.nj.nec.com/banko01scaling.html. In diesem Artikel wird vertreten, dass das Anhäufen von Korpora daher Priorität habe vor der Feinauswahl zwischen verschiedenen funktionierenden Methoden.

11 BISTRO steht für „Bozner Informationssystem für Rechtsterminologie“. Der Internetzugang zur Datenbank ist unter http://www.eurac.edu/bistro ohne Registrierung kostenlos. BISTRO hält sich an alle gängigen IT-Standards, es erfüllt z.B. die W3C Normen für XHTML und CSS und besteht die AAA Prüfung von „Bobby“.

Korpora in der Terminografie

Voraussetzung für den Einsatz von Korpora in der Terminografie

Jedes Korpusprojekt muss sich den Bedingungen der Informatik, der Linguistik und des jeweiligen Fachgebiets unterwerfen.

Die Bedingungen der Informatik sind:

- frühere Arbeit in das Korpusprojekt einfließen zu lassen,
- elektronische Daten zu sammeln (zunächst durch evtl. Verhandeln mit den Verlagen, auch bei allgemein zugänglichen Quellen durch Klärung der Urheberrechtsfragen, schließlich dann auch rein technisch),
- Datenspeicherung (technisch, aber auch verwaltungstechnisch mit verschiedenen Zugriffsrechten für Terminografen und Benutzer, für automatisierte oder einzelne Bearbeitung),
- Sammeln und Speichern von weiteren Informationen über die gesammelten Daten (Untersuchung des Dokumentinhalts durch Menschen oder computerlinguistische Methoden, Untersuchung des Dokuments selbst oder von anderen Informationsquellen über das Dokument),
- Klassifizierung der gespeicherten Daten (explizit oder implizit, physisch oder nicht, nach situationsgerechten oder allgemein anerkannten Klassifikationen),
- Aufbereitung der Daten für verschiedene Zwecke einschließlich dem Datenaustausch und der Wiederverwendbarkeit in anderen Projekten (durch Verwendung von technischen Standards und Offenheit für zukünftige Standards, indem aktuell nicht verwendete Daten, insbes. Metadaten, aufbewahrt werden und alle Veränderungen protokolliert werden),
- Aktualisierung des Korpus (inhaltlich und technisch).

Die linguistische Bedingung ist:

- Ausgewogenheit des Korpus in jeglicher Hinsicht, also bezüglich aller Kategorien, nach denen die Dokumente unterteilt werden (Sprache, Fachgebiet, Rechtssystem, usw.).

Die fachspezifischen Bedingungen für Recht sind:

- Trennung der Rechtssprache vom Sprechen über Recht, von anderen Fachgebieten und von der Allgemeinsprache,

- Trennung von Dokumenten mit Normen, autoritativer Normauslegung (Rechtssprechung) und der Rechtslehre,
- Trennung gültiger Normen von nicht mehr gültigen und Aufzeichnung der Änderungen und ihrer Abfolge (entsprechend für Rechtsprechung und Rechtslehre, die sich auf ungültig gewordene Normen beziehen),
- Trennung von Dokumenten nach der Normenhierarchie (Verfassung, Gesetze, Satzungen usw.),
- Trennung nach dem Rang der Normgeber (EU, Staat, Land, Gemeinde usw.),
- Vereinheitlichung der Klassifikationen für die verschiedenen Rechtssysteme (Land und Regierungsbezirk in Deutschland, Region und Provinz in Italien usw.),
- Verarbeitung von ausdrücklichen und impliziten Verweisen in Dokumenten.

Rechtliche Vorfragen eines Korpus

Man kann zwei Arten elektronischer Quellen unterscheiden. Auf der einen Seite stehen elektronische Dokumente, die uns von offizieller Seite oder von Verlagen zu Forschungszwecken zur Verfügung gestellt werden. Hat man erst einmal eine Erlaubnis zur Benutzung dieser Dokumente, dann haben sie den Vorteil, dass sie bereits bearbeitet sind und daher in gleicher, für gewöhnlich hervorragender Qualität sind. Sofern nicht bereits eine übertragbare Klassifikation vorliegt, kann fast immer die Klassifikation der gedruckten Ausgaben verwenden.[12] Das Datenformat ist ein überwindbares Problem, im Gegensatz zum Kopierschutz. Die Daten können bei einem benutzerorientierten und interaktiven Ansatz nicht intern bleiben, sondern müssen auch den Benutzern zugänglich gemacht werden. Genau dies macht aber die Arbeit, mit der Verlage ihr Geld verdienen, allgemein zugänglich. Selbst Gesetzestexte, die als solche nicht geschützt sind, werden durch ihre Aufbearbeitung (Unterteilung, Verknüpfung usw.) geschützt. Dokumente von Verwaltungen können Datenschutzrechtlich geschützt sein. Selbst wenn man eine Zustimmung erhalten würde, wäre ein so erstelltes Korpus ständig von einem Widerruf der Genehmigung bedroht und würde das gesamte Projekt auf tönerne Füße stellen.

---

[12] Die Klassifikation kann online abgerufen werden, das MAB2 Format unter http://www-opac.bib.bvb.de. Damit liegen die Klassifikationsdaten sofort in elektronischer Form vor. Vergl. auch [23].

Auf der anderen Seite gibt es öffentlich zugängliche elektronische Dokumente von hoher Qualität und Aktualität im Internet. Das Kopieren ganzer Datenbanken ist verboten, für sprachliche Informationen gibt aber mittlerweile genügend verschiedene Quellen.[13] Das Darstellen fremder Inhalte auf eigenen Seiten (sog. „Framing") ist in den meisten Rechtssystemen jedoch verboten. Suchmaschinen (z.B. GOOGLE) kopieren zwar im großen Stil Internetseiten und haben bisher noch keine rechtlichen Konsequenzen zu spüren bekommen, das muss allerdings nicht auf Dauer so bleiben.[14]

Auf den ersten Blick scheint weder der rechtlich sichere aber tatsächlich dornenreiche Weg über Verlage und offizielle Stellen noch der effektive aber rechtlich verfängliche Weg durch Kopieren fremder Inhalte gangbar zu sein. Andererseits ist das bloße Verweisen auf Internetinhalte wegen der Unbeständigkeit der Seiten keine Lösung. Weil Seiten zu rasch ihre Adresse und ihren Inhalt verändern ist ein lokal gespeichertes Korpus fast unumgänglich.

Hier hilft das wissenschaftliche oder publizistische Zitat, das mittlerweile von der Rechtssprechung als stets erlaubt anerkannt zu werden scheint.[15] Das Darstellen von Zitatbruchstücken der elektronischen Dokumente dürfte damit sowohl für die Verwendung von Verlagsmaterial wie auch von nicht mehr veröffentlichten Internetseiten rechtlich gedeckt sein. Als einzige rechtliche Gefahr bleiben mögliche Schadensersatzansprüche durch Verlage, wenn deren Material über die Korpusdarstellung von Dritten geraubt wird.[16] Damit reduziert sich die Wahl auf Internetdokumente, die allerdings nicht ganze Datenbanken umfassen dürfen und deren Inhalte nicht als eigene dargestellt werden dürfen.

---

13 Bei einem Terminologiekorpus geht es ja gerade nicht um die „Wahrheit" oder Güte der Inhalte, sondern um die Versprachlichung von Inhalten eines bestimmten Fachs.

14 Einen Überblick zur Thematik Framing, Inline Linking und Deep Framing bietet Stefan Ott, JurPC Web-Dok. 14/2003, http://www.jurpc.de/aufsatz/20030014.htm : 23.9.2003. Problematisch wird es spätestens dann, wenn Inhalte angezeigt werden, die der Autor aus dem Verkehr gezogen hat, denn damit entfällt jede Basis für die Konstruktion einer stillschweigenden Zustimmung zur Indexierung dieser Inhalte.

15 In diese Richtung geht der BGH für deutsches Recht im Urteil v. 11.07.2002, I ZR 255/00, „Elektronischer Pressespiegel" JurPC Web-Dok. 302/2002, http://www.jurpc.de/rechtspr/20020302.htm, wo allerdings eine Volltextrecherche zum Schutz der Informationshersteller ausgeschlossen sein muss.

16 Der Verlag könnte argumentieren, dass die überlassenen Inhalte nicht ausreichend gegen solche Angriffe gesichert wurden. Nach deutschem Recht könnte das eine pVV des Überlassungsvertrags darstellen.

Korpusaufbau

Internetdokumente können durch einen Web-Robot[17] lokal gespeichert werden. Dazu gibt man eine beliebige URL[18] ein. Der Robot holt sich die Seite von der angegebenen Adresse, prüft sie auf die angegebenen Mindestanforderungen (Kontrolle der robots.txt Datei auf Kopierschutz, Kontrolle der robots meta-tags auf Erlaubnis zum Speichern der Verweise, Mindestzeichenzahl, Sprache usw.) und speichert den Inhalt der Seite dann lokal ab. Die Verweise der gespeicherten Seite werden in eine Einkaufsliste geschrieben und weiter bearbeitet. Bilddateien wie .gif, .tif, jpg. oder cgi-Programme des Servers werden an der Dateinamenerweiterung erkannt und von der Liste gestrichen. Gehen die Verweise auf geeignete Seiten, werden auch sie gespeichert und ihre Verweise zur Einkaufsliste hinzunotiert. Dieses Schneeballsystem kann und muss auf geeignete Weise gesteuert werden. Dazu kann die Suche auf eine Domain und auf Schlagwörter beschränkt werden und eine maximal zu speichernde Seitenzahl angegeben werden.[19] Außerdem wird ein zeitlicher Mindestabstand für die automatischen Zugriffe eingebaut, um die Server nicht zu überlasten. Der nötige Web-Roboter kann im Internet frei heruntergeladen werden.[20]

Dieser Web-Roboter eignet sich auch vorzüglich zum Aktualisieren des Korpus. Es genügt, die Anfragen zur Erstellung zu speichern und nach einiger Zeit erneut in dieser oder ähnlicher Weise selbständig laufen zu lassen. Der Robot ist in der Lage zu erkennen, ob ein Dokument bereits gespeichert ist und ob der Inhalt noch der selbe ist. Dann wird es nicht erneut gespeichert.

Zu jeder Seite werden das zum wissenschaftlichen Zitieren nötige Datum und die für die Suche verantwortliche Person als Metadaten gespeichert. Diese Metadaten werden wie unten beschrieben zur Klassifikation verwendet.

---

17 Auch bot, crawler, agent oder spider genannt. Programm, das einen Auftrag nach dem Initialisieren selbständig weiter ausführt.

18 URL heißt unified resource locator und ist die Adresse eines Dokuments im Internet. Die Adresse besteht regelmäßig aus dem Namen des Protokolls (z.B. http, ftp, dap, file), der Domain, dem Verzeichnis und dem Dateinamen.

19 Weitere Einschränkungen können sein, ob die am URL ablesbare Hierarchie nach oben weiterverfolgt werden darf, ob eine bestimmte Zeichenfolge wie „Recht“ in der URL vorkommen soll. Die hierarchische Strukturierung von Domains hilft oft, nur auf die gewünschten Dokumente zuzugreifen. Wenn Textdateien nur am Ende der Hierarchie zu finden sind, wird die Hierarchie bis nach unten abgefragt und wenn die Dokumente alle auf paralleler Hierarchieebene gespeichert sind, kann auch dies angegeben werden.

20 Z.B. wget von Open Source Robot: http://www.gnu.org/manual/wget/:23.9.2003.

Kodierung des Korpus

In einem Korpus werden Dokumente und Informationen über die Dokumente gespeichert. Anfragen an das Korpus sollten über alle Daten gleichzeitig laufen können, was am leichtesten durch ein gemeinsames Datenformat zu gewährleisten ist. Wir haben uns für XML entschieden, weil XML der Standard für Datenrepräsentation ist[21] und weil viele Werkzeuge für dieses Format kostenlos im Web erhältlich sind[22], so dass nicht alle korpuslinguistischen Werkzeuge selbst erstellt werden müssen.

Alle Daten müssen also in XML Format umgewandelt werden. Relativ einfach umzuformen sind bibliographische Angaben, die im MAB2-Standard von Online-Verbundkatalogen abgerufen werden.[23] Internetdokumente haben aber verschiedene Formaten, z.B. html, pdf und txt. Daher wurde ein Konvertierungsprogramm geschrieben, das zunächst alle Formate in html und dann in XML umformt. Es gibt aber auch einen Standard für die Kodierung von Korpora, den CES (corpus encoding standard). Dessen XML-Version ist XCES[24]. CES normiert die Angaben über die Annotierung der Dokumentstruktur (Titel, Überschrift, Absatz Satz), über linguistische Annotierung im Dokumenttext (flektierte Formen, Phrasen) und über Textalinierung[25]. Für die Speicherung dieser Daten im Header gibt es wiederum einen Standard, den TEI (text encoding initiative)[26], der normiert, wie und wo die tags sein sollen, in denen die Angaben stehen.[27] Die Standards sind alle untereinander kompatibel, so dass die Daten allen Standards gleichzeitig entsprechen können und sollten.

Nun müssen die standardmäßig vorgesehenen Felder natürlich noch ausgefüllt werden. Internetdokumente enthalten in den seltensten Fällen überhaupt verwendbare Metadaten, geschweige denn in standardisierter Form. Außerdem müssen Metafelder nicht nur beim Aufbau eines Korpus ausgefüllt werden, sondern

---

21 http://www.w3.org/XML :23.9.2003.

22 http://xml.apache.org :23.9.2003.

23 Man muss nur einmal ein Mappingprogramm schreiben, d.h. ein Programm, das über die ISBN die Daten zu diesem Werk aus dem Verbundkatalog abruft und automatisch die Daten bestimmter Felder des MAB2 Standards in bestimmte XML Angaben überführt. Das Programm läuft unter http://dev.eurac.edu:8080/cgi-bin/bib/biblio und kann kostenlos online benutzt werden.

24 http://www.cs.vassar.edu/XCES :27.9.2003.

25 Alinierung ist die Angabe, welche Textstellen einander entsprechen, insbesondere welche die Übersetzung zu einem Ausgangstext ist.

26 http://www.tei-c.org/ :27.9.2003.

27 Ein TEI Header muss Angaben über die elektronische Quelle, das entsprechende Druckwerk und das Verhältnis der beiden zueinander machen. Man kann Bibliographische Angaben als Druckwerke eingeben, deren elektronische Version nicht im Korpus ist. Das hat nicht nur den Vorteil, dass später nur noch die elektronische Version angegeben werden muss, sondern auch, dass Header für Dokumente von Verlagen und aus dem Internet gleich aussehen und behandelt werden können.

bei jedem neu hinzukommenden Dokument, ja, bei jeder Änderung der Inhalte der Dokumente und ihrer Beziehungen. Man sollte also computerlinguistische Hilfen so wie möglich ausnutzen. Im Folgenden werden Methoden zur automatischen Erstellung der nötigen Metadaten für Dokumente aus dem Internet vorgestellt.

Klassifizierung rechtlicher Dokumente

In welche Klassen teilt man rechtliche Dokumente ein? Die Einteilung muss nicht nur den Anforderungen einer Informationssuche sprechen, also die Dokumente in geeigneter Weise repräsentieren[28], sondern für alle Nutzer unmittelbar durchschaubar und benutzbar sein. Da die Nutzer sowohl Rechtsexperten wie Linguisten sind, dürfen die Klassen nicht nur rechtlicher Natur sein. Ideal wären Metadaten, die einem Standard entsprechen und die zugleich als wissenschaftliche Quellenangabe[29] dienen können.

Existierende **Katalogisierungsstandards**[30] kommen mit einem einzigen, tief verzweigten Klassifizierungsbaum aus. Sie passen aber nicht richtig zur Aufgabenstellung, denn sie beschäftigen sich nur mit Druckwerken, klassifizieren aus der Sicht eines Außenstehenden und vor allem gehen sie von der Annahme aus, dass der zu klassifizierende Inhalt „der selben Welt" angehört. Rechtstext bezieht sich aber nicht auf die konkrete Welt, sondern auf eine Ideenwelt und darüber hinaus auf eine bestimmte von vielen konkurrierenden Ideenwelten. Diese Ideenwelt heißt Rechtssystem. Normen aus verschiedenen Rechtssystemen gehen fast immer völlig beziehungslos aneinander vorbei. Wenn ein österreichischer Jurist nach § 12 Einkommensteuergesetz sucht, dann hilft ihm § 12 des italienischen Einkommensteuergesetzes nichts. Es gibt keinen Klassifizierungsbaum, der rechtlichen Anforderungen genügt.[31] Ein rechtsvergleichender Standard zur Klas-

---

28 Voltmer L., Dr. Streiter O.: „Textindexierung durch beispielbasierte Termextraktion", EURAC online working paper no. 1, 2003, http://dev.eurac.edu:8080/autoren/publs/wp1.pdf :27.9.2003.

29 Wie oben angesprochen kann grundsätzlich jedes Dokument mit seiner URL zitiert werden. Oft soll aber unabhängig von der Veröffentlichung jener Seite auf den weiterbestehenden (Norm-)inhalt verwiesen werden. Korrekter wäre dann die Angabe der Norm oder offiziellen Fundstelle (v.a. bei Urteilen), damit der wahre Autor an der Quellenangabe ersichtlich ist.

30 UDC, DDC, Dewey, Regensburger Verbundklassifikation (http://www.bibliothek.uni-regensburg.de/rvko_neu/mytree.php3#P:27.9.2003) usw.

31 Man könnte vorhandene Klassifikationen erweitern, zumal die dezimale Systematik eine Erweiterung zuließe. Da diese Erweiterung aber weder Standard ist noch wird, kann man auch gleich eine eigene Klassifikation entwerfen.

sifikation von Dokumenten aus verschiedenen Rechtsordnungen hat sich noch nicht herausgebildet.

Folgende fünf Kriterien könnten für eine Klassenbildung interessant sein:

- Sprache des Dokuments,
- Rechtsordnung, auf die sich der Inhalt bezieht,
- Rechtsqualität des Dokumentinhalts, also ob es eine Norm, eine offizielle Norminterpretation (Urteile etc.), oder nichtoffizielle Informationen sind,[32]
- Evtl. die Ebene in der Normen- und Staatsorganisationshierarchie, aus der die Norm oder Norminterpretation hervorgegangen ist,
- Rechtsgültigkeit, also ob die Norm, auf die sich der Inhalt bezieht, noch Gültigkeit besitzt und
- das Fachgebiet

Entscheidend sind aber nicht nur die Kriterien der Klassen, sondern auch die Festlegung der Kategorien: Was wird als Sprache des Dokuments eingegeben, wenn der Text mehrsprachig ist? Was, wenn eine Norm für mehrere Rechtssysteme gilt oder nur dann subsidiär anwendbar ist. Was, wenn in einem nichtoffiziellen Dokument eine offizielle Norm veröffentlicht wird? Entsprechen die deutschen und österreichischen Bundesländer den schweizer Kantonen und alle ihrerseits einer italienischen Region? Was ist der Gültigkeitsstatus, wenn die Norm ab oder bis zu einem bestimmten Datum gilt?

Für **Sprachen** bietet es sich unmittelbar an, die geltenden ISO Normen zu verwenden. Diese kennen keine Sprachkombinationen, so dass die Kodierung eines mehrsprachigen Texts durch mehrfache Etikettierung erfolgt. Das entspricht in aller Regel auch den Bedürfnissen der Benutzer, die auf alles Sprachmaterial einer Sprache zugreifen möchten, selbst wenn es in einem mehrsprachigen Text steht.[33]

Am schwierigsten ist die Lösung für die **Gültigkeit** von Normen. Normen können formal noch gültig sein, aber obsolet sein. Bei Normenkontrollverfahren kann gerade die Gültigkeit einer Norm Streitgegenstand sein. Wenn selbst Fachleute

---

[32] Vergleiche die „Dokumenttypen“ bei Unger W., „Methoden juristischer Dokumentrecherche“, http://www.juralink.de/8LITERATUR/Umgang/Recherche.htm:27.9.2003.

[33] Es gibt also keine eigene Klasse „deutsch-italienisch“, sondern ein Dokument kann mehreren Klassen angehören. Diese Klassifizierung könnte verbessert werden, indem man die Klassifizierung nur auf den Teil des Textes anwendet, der tatsächlich in der jeweiligen Sprache geschrieben ist, so dass keine falschen Treffer für die weitere Sprache entstehen.

nicht jederzeit eindeutig die Gültigkeit einer Norm feststellen können, wie soll es dann bei der Korpusklassifikation oder gar automatisch geschehen? Die einzige Information, die sich in aller Regel auf die Gültigkeit bezieht und die automatisch herausgefunden werden kann, ist die Chronologie von aufeinanderfolgenden Normen. Wenn im Korpus zwei Versionen eines Gesetzes sind, dann spricht viel dafür, dass das neuere Gesetz das alte ersetzt hat. Damit ist das alte ungültig und das neue gültig. Mehr Informationen über die Gültigkeit würden bei der korpuslinguistisch erforderlichen Menge an Dokumenten die Datenpflege überfordern.

Als **Rechtsordnung** wurden alle souveränen Staaten gewählt und, juristisch fehlerhaft aber aus informationstechnischen und praktischen Gründen, die beiden Auffangkategorien EU-Recht und international. Ein Dokument kann sich zugleich auf mehrere Rechtsordnungen beziehen.[34]

Die Ebene in der **Normen- und Staatsorganisationshierarchie** wurde nur sehr schwach an die territorial units for statistics (NUTS)[35] der EU angelehnt und werden eher untechnisch und pragmatisch zur weiteren Unterteilung angewandt. Da nur offizielle Normen und Norminterpretationen hierarchisch zueinander stehen, gibt es nur einen oder keinen Eintrag in dieser Klasse.

Aufstellung der Klassifizierung

Sprache	Rechtsordnung	Normen- und Staatsorganisationshierarchie	Rechtsqualität
Italienisch	International	Staat	Gesetz
Deutsch	EU	selbständiger Staatsteil/Teilstaat	andere Norm
Grödnerisch	Italien	Region/Provinz	Urteil
Gadertalerisch	Österreich	Gemeinde	Anderes Rechtsdokument
Fassanisch	Deutschland	nicht anwendbar	Wörterbuch
Standard Dolomitenladinisch	Schweiz		Anderes Dokument

Die bei Weitem komplexeste und inhaltsorientierteste Unterteilung ist die **Unterteilung in Fachgebiete**. Die Schwierigkeit besteht darin, dass verschiedene Fachgebietseinteilungen konkurrieren. Zum einen konkurriert die korpuslinguistische Einteilung (gleich große Datenmengen je Klasse) mit der sprachlichen Ein-

---

[34] Z.B. ein Rechtsvergleich oder eine internationale Norm, die zugleich unmittelbares anwendbares Recht in mehreren Staaten ist.

[35] http://europa.eu.int/comm/eurostat/ramon/nuts/codelist_en.cfm: 27.9.2003.

teilung (eigenes Fachvokabular und eigene Ausdrucksweise) und der rechtlichen Einteilung (allgemeines Schuldrecht und besonderes Schuldrecht unterscheiden sich stark in Menge, nicht in Vokabular oder Ausdrucksweise, aber in ihrer rechtssystematischen Stellung). Es konkurrieren die Einteilungen der verschiedenen Rechtssysteme (Während in Deutschland und Österreich das Handelsrecht ein eigenständiges Rechtsgebiet darstellt und unter dem Zivilrecht, parallel zum bürgerlichen Recht eingeordnet wird, ist es in Italien Teil des Zivilrechts. In Deutschland und Österreich gibt es traditionell ein eigenes Handelsgesetzbuch, in Italien finden sich die Regelungen im Codice Civile/Zivilgesetzbuch). Schließlich konkurrieren verschiedene Sichten des Rechts (Nach einer Ansicht gibt es die drei Rechtsgebiete Zivilrecht, Öffentliches Recht und Strafrecht, nach einer anderen Ansicht ist das Strafrecht ein Teil des Öffentlichen Rechts). Schließlich konkurrieren auch Rechtstraditionen (Deutschland und Österreich mit Abstraktionsprinzip und konstitutivem Grundbuch, Italien und Frankreich ohne).

Deswegen konnte sich weder eine Norm für die Klassifizierung von Recht, noch ein verbreiteter Gebrauch für einen bestimmten Bereich durchsetzen. Jede Klassifizierung wird daher höchst angreifbar bleiben, was für eine pragmatische und nicht allzu tief verzweigte Einteilung spricht. Von Vorteil erscheint jedenfalls eine Dezimalklassifikation, um flexibel hinsichtlich der Einteilung und Bezeichnung der Klassen zu bleiben. Da an der EURAC terminologische Einträge bereits in Fachgebiete eingeteilt waren, wurden zunächst einmal diese auch für die Einteilung von Korpusdokumenten ausgewählt.

Eine wichtige Entscheidung ist, ob die Fachgebiete alle nebeneinander stehen oder hierarchisch aufgebaut sind. Wenn es eine Hierarchie gibt ist wichtig, ob nur die jeweils niedrigste, konkreteste Ebene Dokumente enthalten darf oder ob weiter unterteilbare Klassen kein eigenes Fachgebiet, sondern nur Unterteilungskriterien sind. Man sollte sich jedenfalls entscheiden, wenn also eine Hierarchie besteht, sollte sie alle Klassen zueinander in Beziehung setzen. Ebenso sollte man vermeiden, dass einigen Unterteilungen Dokumente zugeordnet werden können und anderen nicht.

Wir haben uns an der Systematik des Bundesrechts des Juristischen Internetprojekts Saarbrücken orientiert.[36], an der Systematik der Schweizer Gesetzestexte Online[37] und der Documentazione Giuridica[38] für Italien orientiert und eine aus eigenen Klassen bestehende Hierarchie entschieden:

---

[36] http://www.jura.uni-sb.de/BGBl/BGBLSYST.HTML:27.9.2003. Herberger, M., Systematik des Bundesrechts, Projektbericht, Juristisches Internetprojekt Saarbrücken, 1997. Diese Klassifikation dürfte die Materialmenge in den jeweiligen Fachgebieten berücksichtigen und schien uns daher ein Kompromiss zwischen den korpuslinguistischen und fachlichen Bedürfnissen.

[37] http://www.gesetze.ch/ :27.9.2003.

0 Recht allgemein
1 Öffentliches Recht
- 10 Verfassungsrecht
- 11 Staatsorganisationsrecht
- 13 Bundesgrenzschutz
- 17 Europarecht
- 18 Völkerrecht

2 Verwaltungsrecht
- 21 besonderes Verwaltungsrecht
  - 211 Straßenverkehrsrecht
  - 213 Baurecht
  - 219 Polizeirecht
  - 221 Universitätsrecht
- 27 Auswärtiger Dienst

3 Rechtspflege
- 30 Gerichtsverfassungsrecht
- 31 Prozessrecht
  - 310 Zivilprozessrecht
  - 312 Strafprozessrecht
  - 315 freiwillige Gerichtsbarkeit
  - 318 Verwaltungsprozessrecht

4 Zivil- und Strafrecht
- 40 bürgerliches Recht
  - 401 bürgerliches Recht allgemeiner Teil
  - 402 Schuldrecht
  - 403 Sachenrecht
  - 404 Familienrecht
  - 405 Erbrecht
- 41 Handelsrecht
  - 411 Börsenrecht
  - 412 Gesellschaftsrecht
- 45 Strafrecht

5 Militärrecht
6 Finanzwesen
7 Wirtschaftsrecht
- 76 Geld-, Kredit- und Versicherungswesen
- 79 Umweltrecht

8 Arbeits- und Sozialrecht
- 80 Arbeitsrecht
- 86 Sozialrecht

x nicht Fachgebiet Recht

---

[38] http://www.idg.fi.cnr.it/banche/dogi/dogi.htm .

Die Entscheidung für eine Hierarchie aus Klassen bedeutet nicht, dass eine automatische Klassifikation nicht alle Klassen gleich behandeln kann. Die Information über die Beziehung der Klassen zueinander mag einerseits wertvoll sein, kann andererseits aber auch die Automatisierung verkomplizieren.

Automatische Klassifizierung rechtlicher Dokumente

Automatische Dokumentklassifikationen nutzen in aller Regel die Ähnlichkeit des zu klassifizierenden Dokumentes mit bereits klassifizierten Dokumenten.[39] Die zugrunde liegende Annahme ist, dass ähnliche Dokumente der gleichen Klasse angehören. Die Ähnlichkeit bezieht sich auf lexikalische Gleichförmigkeit (dieselben Wörter in den Dokumenten) oder n-Gramm[40]-Gemeinsamkeit.[41] Solche Werkzeuge sind im Internet frei erhältlich.[42] Sie werden zumeist als Spracherkenner verwendet und funktionieren bei hinreichend viel Untersuchungsmaterial und hinreichender Unterschiedlichkeit der zu erkennenden Sprachen praktisch fehlerfrei.[43] Die Unterschiede zwischen Fachgebieten werden aber nicht so groß sein

---

39 Chien, L.-F. und Chen, C.-L., Incremental extraction of domain-specific terms from online text resources, in Bourigault, D., Jacquemin, C. and L'Homme, M.-C., Hg., Recent Advances in Computational Terminology, John Benjamins, Amsterdam, Natural Language Processing, 2001.
Nohr, H., Automatische Indexierung, Einführung in betriebliche Verfahren, System und Anwendungen, Materialien zur Information und Dokumentation, Verlag für Berlin-Brandenburg, Potsdam, 2001.

40 N-Gramme sind die n (=natürliche Zahl) aufeinanderfolgenden Zeichen oder Worte eines Textes. Bsp.: Recht besteht aus den Buchstaben-4-Grammen –Rec, Rech, echt und cht-.

41 Damashek, M., Gaugin similarity via n-grams: Language-independent sorting, categorization, and retrieval of text, Science, 1995, 267, S. 843-848, der die beiden Verfahren vergleicht und zu dem Schluss kommt, dass n-Gramme dem Wortansatz ohne linguistische Zusatzinformation meist überlegen sind. Cowie, J., Ludovik, E. and Zacharski, R., An autonomous, web-based multilingual corpus collection tool, in Proceedings of the International Conference on Natural Language Processing and Industrial Applications, 1998 sehen den Grund dafür darin, dass n-Gramme Wortstamm, -beugung und -zusammensetzung als ähnlich erkennen. Möglicherweise kann man sich mit N-Grammen auch an die optimale Menge vergleichbarer Textteilchen annähern. Es wäre interessant, den Zusammenhang zwischen der sprachlichen Dichte einer Schrift und dem besten n für n-Gramme zu untersuchen.

42 http://odur.let.rug.nl/~vannoord/TextCat/. Meist wurden die Programme zum Erkennen von ähnlichen Sprachen geschrieben.

43 „Die automatische Sprachenidentifizierung für elektronische Dokumente, deren Mindestlänge eine bestimmte Wortzahl überschreitet und die regulären Text enthalten, kann als weitgehend gelöstes Problem betrachtet werden." Langer, S. (2002): Grenzen der Sprachenidentifizierung. Tagungsband KONVENS 2002, Saarbrücken S. 99-106, S. 99, http://konvens2002.dfki.de/cd/pdf/19V-langer.pdf: 27.9.2003. Nicht immer bekommt man regulären Text im Internet, weil Inhalte oft in Fenster mit anderen Inhalten (Werbung, Bilder, Navigationshilfen, Menüleisten) eingebunden werden und weil der übersandte Quelltext (anders als die Anzeige in einem Browser) oft mit Formatierungshinweisen (statt Inhaltshinweisen, die das Formatieren dem Browser überließen) übersäht ist.

wie zwischen Sprachen, weshalb hier weitere Anhaltspunkte für eine Klassifizierung vorgestellt werden.

Beim oben beschriebenen Korpusaufbau fallen weitere elektronische Informationen an:

	Information	Art und Herkunft
1.	URL des Dokuments	Oft im Adressbalken des Browsers angezeigt
2.	Schlagwörter	in den Metadaten eines Dokuments
3.	Inhaltsbeschreibung	in den Metadaten eines Dokuments
4.	Titel	im Dokumentinhalt und den Metadaten
5.	Verweisadresse im Dokumentinhalt[44]	URL, auf die ein Link verweist.
6.	Mitarbeiterprofil	Loginname des Mitarbeiters und Startzeit des Web-Robots

Diese Daten müssen vom Web-Robot jeweils getrennt gespeichert werden. Zum Ähnlichkeitsvergleich bietet es sich ganz offensichtlich an, die URLs zweier Dokumente heranzuziehen. Auch ein Vergleich ungleicher Klassen wie etwa Schlagwörter und Inhaltsbeschreibung könnte aber Sinn machen. Chakrabarti et al. zeigen beispielsweise, dass der Vergleich der URL eines Dokuments mit der Verweisadresse eines anderen ein wertvolle Informationen über deren Ähnlichkeit enthält.[45] Es ist auch nachvollziehbar, dass eine Seite auf eine sehr ähnliche Seite verweist, die ihrerseits nicht mehr auf ähnliche Seiten weiter verweist, so dass ein solcher Klassifikator zusätzliche Informationen einbringen kann.

---

[44] Mit dem Open Source Browser LYNX kann man die Verweisadressen leicht vom Dokumentinhalt trennen: http://lynx.browser.org. Die Verweisadresse in EURAC ist http://www.eurac.edu. Man könnte auch den Verweistext (Im Beispiel wäre das „EURAC“)mit einbeziehen, wie das die Suchmaschine GOOGLE tut, um Suchergebnisse zu sortieren: „link structure and link text provide a lot of information for making relevance judgments and quality filtering“ Brin, S., Page, L., The anatomy of a large-scale hypertextual Web search engine, in Computer Networks and ISDN Systems Vol. 30, Nr.1-7, S. 107-117, 1998, http://citeseer.nj.nec.com/brin98anatomy.html. Mit Hilfe von GOOGLE könnte man sogar die in-links, also die Seiten, die auf eine URL hinverweisen, bekommen. Hier wurde auf die weitere Abfrage von Daten verzichtet, es handelt sich also immer um out-links, die wegverweisen.

[45] Chakrabarti, S., Dom, B., Indyk, P., Enhanced hypertext categorization using hyperlinks, in ACM SIGMOD 1998, Seattle, Washington, 1998, untersuchten allerdings die Identität von Verweisadresse und URL. Es scheint offensichtlich, dass die konkret gemeinte Seite inhaltlich mit der Ausgangsseite zu tun hat. Sehr viel gewagter ist die Hypothese, dass Seiten, die ähnliche n-Gramme in der Adresse haben, dem verweisenden Dokument statistisch gesehen ähnlicher sind als eine Zufallsauswahl.

Hier werden also auch die URL des zu klassifizierenden Dokuments mit den Verweisadressen in den klassifizierten Dokumenten und der Verweisadressen des zu klassifizierenden Dokuments mit den URLs der klassifizierten Dokumente verglichen.

	Vergleich	erhoffte Information
1.	Dokumentinhalt – Dokumentinhalt	Sprache, Rechtsordnung, Fachgebiet, interne Hierarchie
2.	URL - URL	Sprache, Rechtsordnung, Fachgebiet, interne Hierarchie
3.	Schlagwörter - Schlagwörter	Sprache, Fachgebiet, interne Hierarchie
4.	Inhaltsbeschreibung - Inhaltsbeschreibung	Sprache, Fachgebiet, interne Hierarchie
5.	Titel - Titel	Sprache, Rechtsordnung, Fachgebiet, interne Hierarchie
6.	Verweisadresse - Verweisadresse	Sprache, Rechtsordnung, Fachgebiet
7.	Mitarbeiterprofil - Mitarbeiterprofil	Fachgebiet[46]
8.	URL – Verweisadresse[47]	Sprache, Fachgebiet
9.	Verweisadresse - URL[48]	Sprache, Fachgebiet

Kombinationsproblem

Das Kombinieren von mehreren Ähnlichkeitskriterien in einem einzigen Klassifikator[49] kann zum peaking Phänomen führen: Mit den Ähnlichkeitskriterien müssen die Trainingsdaten exponentiell zunehmen, damit die Ergebnisse nicht schlechter werden. [50] Bei der Kombination von eigenständigen Klassifikatoren ist hingegen seit langem bekannt, dass selbst gute Klassifikatoren durch eine Kombination schwacher Klassifikatoren geschlagen werden können.[51] Daher schlägt in der vorgestellten Methode jedes Ähnlichkeitskriterium eine Klasse vor. Später können die Klassifikatoren dann kombiniert werden.

---

46 Die einzelnen Mitarbeiter sind mit einem je eigenen Fachgebiet betraut und werden Dokumente für ihre eigene Arbeit speichern.

47 Hinverweisender Link oder „in-neighbor".

48 Wegverweisender Link oder „out-neighbor"

49 Ein Klassifikator (classifier) ist eine Rechenvorschrift (Algorithmus) für die Zuordnung einer Klasse.

50 Jain, A. K., Duin, R. P. and Mao, J., Statistical pattern recognition: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1), S. 4-37.

51 Larkey, L. S., Croft, W. B., Combining classifiers in text categorization, in Proceedings of SIGIR-96, 19. Internationale Konferenz über Wissenschaft und Entwicklung im Information Retrieval, ACM Press, New York-Zürich, 1996, S. 289-297. In diesen Versuchen ist die Kombination beliebiger Klassifikatoren besser als der beste Klassifikator alleine. Die Klassifikatoren müssen dafür wohl aber hinreichend unterschiedlich sein, damit ihre jeweilige Stärke die Schwäche des Partnerklassifikators wettmachen kann.

Nearest Neighbour Verfahren

Ein Nearest Neighbor (NN) classifier oder k-NN-Klassifikator berechnet zu jedem Dokument einen Merkmalsvektor. Die Annahme ist, dass das Dokument der Testdatenmenge, dessen Merkmalsvektor den geringsten Abstand zu dem zu klassifizierenden Dokument hat, das Ähnlichste ist. Dessen Klasse wird dem zu klassifizierenden Dokument zugewiesen. Das k steht für die maximale Anzahl herangezogener Nachbarn. Unter mehreren Nachbarn wird durch Abstimmen (voting) ausgewählt. Je mehr nahe Nachbarn ihre Information in die Abstimmung einbringen können, umso besser wird das Ergebnis, aber je mehr ferne Nachbarn einbezogen werden, umso schlechter wird es. Meistens wird k zwischen 3 und 20 gewählt.

Das Verfahren bringt bereits bei einem klassifizierten Dokument ein Ergebnis. Es hat eine steile Lernkurve, verbessert sich also schnell mit zunehmenden Lerndaten. Das bedeutet für den Korpusneuaufbau, dass bereits ab dem zweiten Dokument der Übergang zur automatischen Erkennung beginnt,[52] weil bereits eine Klassifikation vorgeschlagen wird. Die intellektuelle Korrektur der Klasse (semiautomatische Klassifizierung) wird immer seltener nötig und schließlich sollte die Fehlerschwelle so niedrig werden, dass die Klassifizierung unbeaufsichtigt vonstatten gehen kann. Während statistische Verfahren mit steigenden Lerndaten immer schlechter korrigiert werden können[53] und im Extremfall gar nicht mehr lernen,[54] ist das bei NN Verfahren nicht der Fall, weil jedes Dokument einen eigenen Ort hat[55] und als eines von wenigen Dokumenten (nicht als eine statistische Nichtigkeit unter vielen) entscheidend sein kann.

Entscheidend ist aber, welches Ähnlichkeitsmaß für die Berechnung des „nächsten“ Nachbardokuments zugrunde gelegt wird.[56]

---

52 Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P., Vilain, M., Mixed-initiative development of language processing systems, in 5. Conference on Applied Natural Language Processing, Association for Computational Linguistics, Washington D.C., 1997.
Streiter, O., Corpus-based parsing and treebank development, in ICCPOL 2001, 19. International Conference on Computer Processing of Oriental Languages, Seoul, Korea, 2001, S. 115-120.

53 Day et al. a.a.O.

54 Kurohashi, S., Nagao, M., Building a Japanese parsed corpus while improving the parsing system, in First International Conference on Language Resources & Evaluation, Granada, Spain, 1998, S. 719-724.

55 Aha, D. W., Editorial- lazy learning, Artificial Intelligence Review, 1997, (11), S. 1-3.

56 Eine nützliche Einführung in die verschiedenen Methoden der Dokumentklassifizierung bieten Han, E.-H. S., Karypis, G., Centroid-based document classification: Analysis & experimental results, 2000, http://www.cs.umn.edu/karypis.

Ähnlichkeitsmaß und k-NN Algorithmus

Kaum ein Klassifizierungsverfahren wendet das NN-Verfahren in Reinform an, weil dazu ja den Abstand eines Merkmalsvektors zu allen Tausend oder Millionen anderen berechnen müsste. Daher werden die zu vergleichenden Dokumente oft vorher schon reduziert, möglichst ohne die nächsten Nachbarn dabei zu verlieren, und dann wird das NN-Verfahren angewandt. Wenn die Reduktion zu stark ist, dann fällt die Auswahl des NN-Verfahrens nicht mehr ins Gewicht.

Wegen der z.T. sehr kurzen und nicht als Einzelwort vorliegenden Informationen (Verweisadresse) wird das Ähnlichkeitsmaß durch 3 und 4-Gramm Ähnlichkeit berechnet. Aus Termfrequenz[57] (TF) und Dokumentfrequenz[58] (DF) ergibt sich nach der anerkannten Formel TF.IDF=TF/DF die Relevanz eines n-Gramms für die Klassifizierungsaufgabe. Das bedeutet, dass ein n-Gramm umso mehr über die Ähnlichkeit zweier Dokumente aussagt, je öfter es in einem Dokument und je seltener es in allen Dokumenten vorkommt.

Nun wird der Einfluss unterschiedlicher Textmengen auf die Frequenzen ausgeglichen, indem die TF eines n-Gramms mit der TF des häufigsten n-Gramms in Beziehung gesetzt wird: $TF_{neu} = 0.5 + 0.5 \times TF/TF_{max}$. Der neue TF.IDF berechnet sich dann: $TF.IDF_{neu} = TF_{neu} / DF$.

Durch eine Kosinusnormalisierung wird erreicht, dass die unterschiedliche Textmenge der Dokumente nicht ins Gewicht fällt: $TF.IDF_{normal} = TF.IDF_{neu} / \sqrt{\Sigma\, TF.IDF_{neu}^2}$. Dieses Ähnlichkeitsmaß heißt atc-weight (ATC-Gewicht).

Der NN-Algorithmus holt sich also in einem ersten Schritt alle Dokumente, die die relevantesten n-Gramme des zu klassifizierenden Dokuments enthalten. Nur diese Auswahl wird dann auf die Kosinusähnlichkeit (ATC-Gewicht) aller n-Gramme untersucht.[59] Nur die besten NN dürfen dann mit ihrer Klasse über die Klasse des zu klassifizierenden Dokuments abstimmen.

Experimente

Zum Testen wurden 140 Dokumente intellektuell klassifiziert und gelten als ideale Klassifizierung. Dann versucht jeder der neun Klassifikatoren, dieses Ergebnis zu reproduzieren. Das erste Dokument kann mangels klassifizierter Nachbarn nicht klassifiziert werden. Danach kommt dieses erste Dokument mit der Information über seine Klassen zu den Lerndaten, aus denen der Nachbar gewählt

57 Vorkommen des N-Gramms in einem Dokument.
58 Vorkommen des N-gramms in allen Dokumenten.
59 Manning, C. D., Schütze, H., Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, London, 1999, http://www.sultry.arts.usyd.edu.au/cmanning.

wird. Der Klassifikator rät für das zweite Dokument zwangsläufig die Klassen des bisher einzig klassifizierten Dokuments. Danach sollte der Lerneffekt einsetzen. Um die Größe des Lerneffekts abschätzen zu können, wird ein zehnter Klassifikator eingesetzt, der einen zufälligen Nachbarn auswählt (random NN).

Auf der X-Achse ist die Anzahl der klassifizierten Dokumente aufgetragen, also der möglichen Nachbarn. Das Dokument 101 wird aufgrund der Informationen der 100 bis dahin klassifizierten Dokumente klassifiziert. Auf der Y-Achse sieht man die Übereinstimmung des Klassifikators mit der intellektuellen Klassifikation. Ein Wert von 0.5 bedeutet, dass jedes zweite Dokument richtig klassifiziert werden konnte. Je nach Aufgabe und Anzahl der Klassen kann das ein gutes oder schlechtes Ergebnis sein. Nicht korrekte Klassifizierungen können an der Wahl eines falschen NN oder am Fehlen eines NN der richtigen Klasse liegen. Erst wenn jede Klasse zumindest einen NN aufweist, kann ein Klassifikator wenigstens theoretisch stets die richtige Klasse vorschlagen; vorher „kennt“ er die Klasse überhaupt nicht.[60] Dieser Effekt, der natürlich je nach Anzahl der Klassen länger oder kürzer dauert, schlägt sich aber auch auf die Ergebnisse des Zufallsklassifikators nieder, so dass dieser als Vergleichsmaßstab dient. Von Klassifikatoren, die sich nach 140 Trainingsdokumenten noch nicht vom Zufallsklassifikator absetzen konnten, kann man annehmen, dass sie für die Klassifizierung dieser Klasse keine Information tragen.

Wenn für ein richtiges Ergebnis der Wert 1 und für ein falsches der Wert 0 vergeben wird, dann muss das Ergebnis extrapoliert werden, um eine Kurve zu erhalten. Hier werden die 20 letzten Ergebnisse aufaddiert, die Graphen zeigen aber noch deutlich die Schwankungen der binären Ereignisse. Der Endpunkt der Graphen nach 140 Dokumenten bezeichnet also nicht den durchschnittlichen Erfolg bei allen Versuchen, sondern eher die Erfolgswahrscheinlichkeit für den unmittelbar folgenden Versuch. Aus dem Schwanken der Graphen kann man ablesen, dass der Endpunkt auch nicht die Güte der einzelnen Klassifikatoren endgültig klärt, weil sich bei Fortführung des Experiments über weitere 100 oder 10000 Dokumente ein anderer Klassifikator durchsetzen kann.

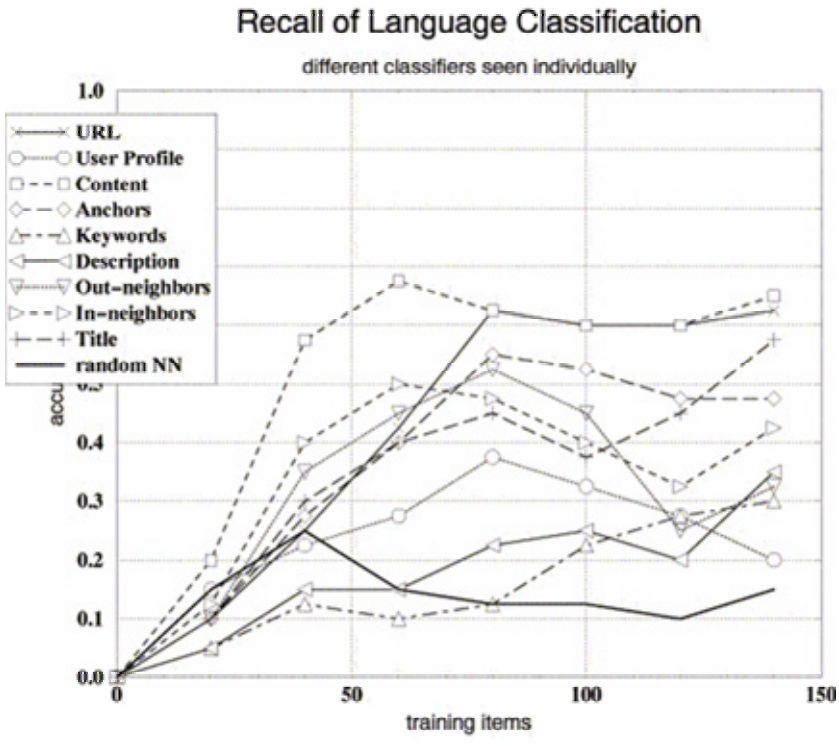
---

[60] Für den Klassifikator kommt mit jedem Beispieldokumente, das eine bisher unbekannte Klasse trägt, eine neue Klassifizierungsmöglichkeit hinzu. Das Hinzufügen einer konzeptuell neuen Klasse zu einem späteren Zeitpunkt ist für diesen Klassifikator also nichts Außergewöhnliches und seine Ergebnisse werden deshalb nur vorübergehend etwas schlechter.

Vorteile von NN-Klassifikatoren

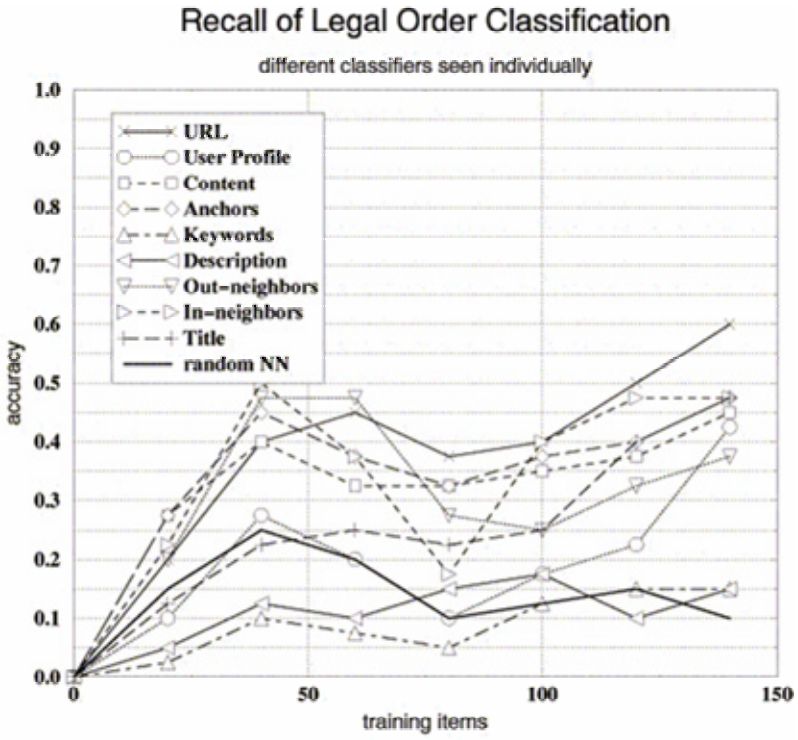
Es wurde erwartet, dass die Präzision der Klassifikatoren von der Anzahl der Klassen abhängt, dass also die Norm- und Staatsorganisationshierarchie mit nur fünf Klassen acht mal leichter zu klassifizieren sein würde als das Fachgebiet mit 41 Klassen. Der Unterschied hat aber nur den Faktor drei. Außerdem wurde erwartet, dass 140 Dokumente leicht ausreichen würden, um fünf Klassen mit guten NN zu belegen, während die Lernkurve bei vierzig Klassen noch weiter nach oben zeigen würde. Die Lernkurven zeigen aber, dass Klassifikatoren selbst mit einer geringen Anzahl Lerndokumenten schon gute Arbeit leisten. Ein Grund kann sein, dass die Klassifikatoren nicht aus dem gesamten Klassifikationsraum auswählen, weil sie ihn noch gar nicht ganz kennen. Das entspricht der Information, dass häufigere Klassen früher im Klassifikationsraum auftauchen und dass die Dokumente ungleichmäßig auf diesen Raum verteilt sind. Dies wiederum deutet darauf hin, dass (noch) kein gewichtetes, also in etwa gleich verteiltes Korpus entstanden ist. Würden alle Klassen in etwa gleichmäßig belegt, dann wäre genau dieser Vorteil wieder zunichte gemacht.

Sprachklassifikation



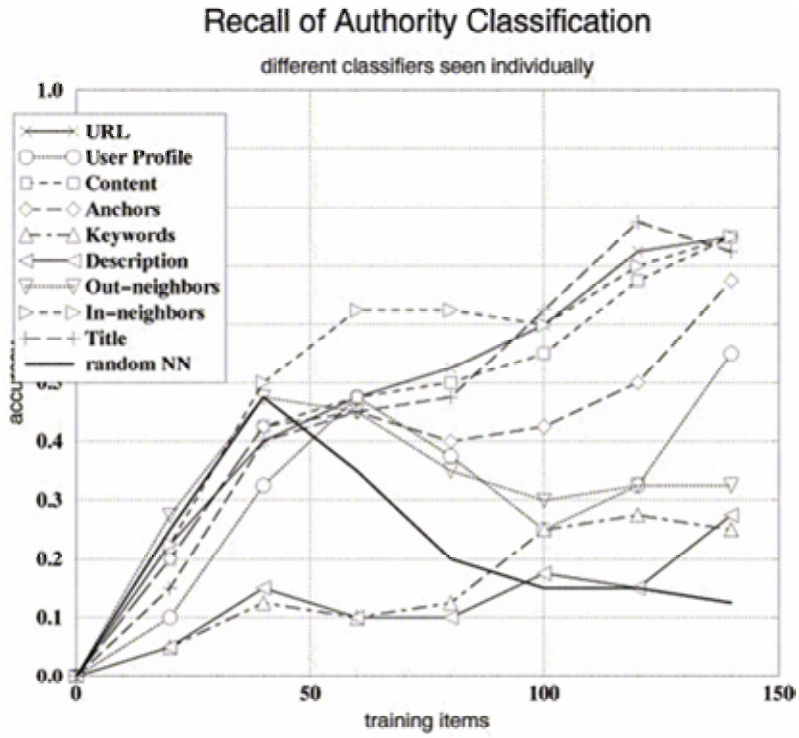
Da die Sprachklassifizierung von einsprachigen Texten praktisch gelöst ist, wurden versucht, auch mehrsprachige Texte zu erkennen. Der Inhalt zeigt sich von Anfang an als bester Klassifikator und erreicht eine 60-70 %ige Wahrscheinlichkeit, die richtige Sprache oder Sprachkombination zu erreichen. Auch die URL erzielt mit dem 4-Gramm „.it/“ oder „.de/“ hervorragende Ergebnisse. Das Mitarbeiterprofil sagt sehr schlecht vorher, was sich darauf zurückführen lässt, dass alle Mitarbeiter mehrsprachig sind und für alle Sprachen gleich viel Material zur Arbeit benötigen. Die doch enttäuschenden Ergebnisse in dieser Sparte ergeben sich durch die englischsprachigen Einschübe vor allem im Wirtschafts- und Bankenrecht. Auch wenn ein Dokument nicht einmal ganze Sätze, sondern nur einige Fachausdrücke in englischer Übersetzung enthält, möchten wir dieses Dokument zur Suche eines englischsprachigen Begriffs einsetzen. Es soll daher nicht nur als einsprachig, sondern auch als Englisch klassifiziert werden, was für den Klassifikator eine eigene, neue Klasse ist. Insofern sind die Ergebnisse durchaus im Rahmen.

Klassifizierung der Rechtsordnung



Bei der intellektuellen Klassifikation einzelner Dokumente, die nicht über Hyperlinks eingeordnet sind, orientiert sich der Fachmann an der URL, die mit den Anhängseln „.ch“, „.eu“ oder „.at“ eine starke Vermutung für eine bestimmte Rechtsordnung enthält. Der URL Klassifikator ist daher mit 60 % der beste zur Vorhersage der Rechtsordnung. Die Metadaten sind zu selten, um konkurrenzfähig zu sein, was umso interessanter ist, weil sie gerade für solche schwer aus dem Inhalt erkennbaren Informationen (deutsches oder österreichisches Einkommensteuergesetz?) ersonnen wurden. Selbst wenn die Identifizierung der Klassen fehlerfrei wäre, müsste das Gros der Dokumente stets durch einen anderen Klassifikator erkannt werden. Dies gilt für alle zu erkennenden Klassen. Allerdings könnten Metadaten in Kombination mit einem anderen Klassifikator die Ergebnisse verbessern.

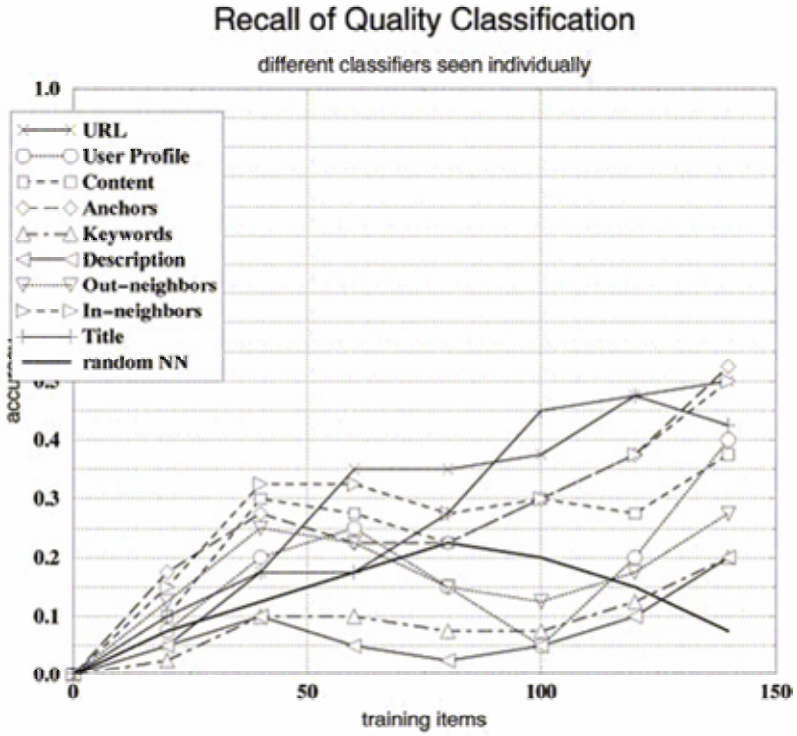
Klassifizierung der Norm- und Staatsorganisationshierarchie



Titel, URL, Inhalt und URL-Verweis (hinverweisende Links) funktionieren etwa gleich gut, und knapp dahinter liegen Verweise und Mitarbeiterprofil. Unerwartet ist, dass die hinverweisenden Links so viel besser abschneiden als die wegverweisenden. Unsere Erklärung dafür ist, dass die Dokumente, die mit einer besonde-

ren Klasse versehen werden müssen, stets von offiziellen Servern geladen werden. Diese seriösen Server bieten meist Text und einen Verweis auf die Wurzel oder die zentrale Seite. Andere Verweise werden hingegen nicht auf den Dokumentseiten, sondern in einer separaten Linkseite angeboten. Wegverweisende Links sind daher sehr selten und bieten zu wenig Information. Der Servername und evtl. der Ordner ist hingegen oft offiziell und enthält charakteristische bedeutungstragende Bruchstücke wie „Bundes“ oder „gesetz“ und erlaubt eine gute Klassifizierung.

Klassifizierung der Rechtsqualität

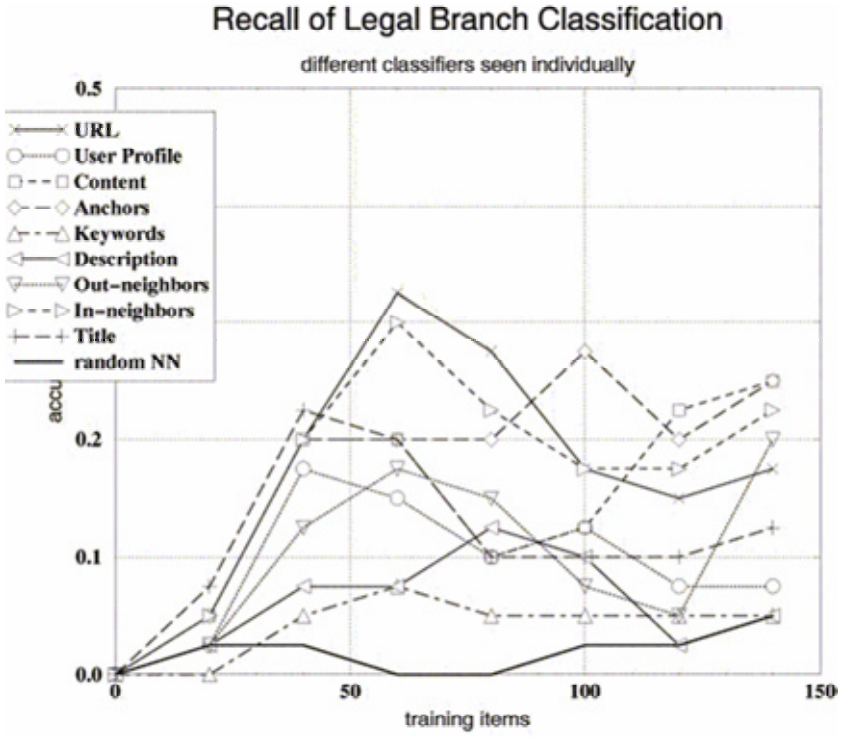


Metadaten und der Unterschied zischen hin- und wegverweisenden Links wurde bereits besprochen. Insgesamt sind die Kurven flach; nach 70 Dokumenten ist praktisch nur die URL deutlich besser als die Zufallsklassifikation. Bei Beendigung der Versuche stiegen acht von neun Graphen noch weiter an. Daraus kann man vielleicht generell den Schluss ziehen, dass unabhängig von der Anzahl der Klassen Versuche mit sehr viel mehr Dokumenten nötig sind, um zu einer verallgemeinerbaren Aussage gelangen zu können. Die hohe Güte, die Spracherkenner er-

reicht haben, macht Hoffnung für die Erkennung von anderen Klassen, sofern die hierfür nötige Menge Lernmaterial oder hinreichend große Textmengen je Dokument erreicht werden.

Fachgebietsklassifikation

Wir erwarteten, dass das Mitarbeiterprofil hier am besten abschneiden würde. Das war wohl deshalb nicht der Fall, weil die Fachgebiete kleiner geschnitten sind als das Arbeitsgebiet eines Mitarbeiters. Ins Arbeitsgebiet Verwaltungsrecht fällt etwa das besondere Verwaltungsrecht, Baurecht, Polizeirecht und Verwaltungsprozessrecht. Außerdem müssen Mitarbeiter zur Bearbeitung eines Gebiets auch die Bedeutung eines Begriffs in anderen Gebieten zumindest eruieren.



Mögliche Verbesserungen

1. Eine große Verbesserungsmöglichkeit bietet wahrscheinlich die Kombination von Klassifikatoren. Für statistisch aussagekräftige Versuche wären aber mehr als 140 Dokumente nötig. Die erprobten Kombinationen lassen so keine Systematik erkennen.

2. Eine weitere Möglichkeit zur Verbesserung könnte die Anwendung der Mengenlehre zur Reduzierung der Klassen sein. Ein zweisprachiges Dokument könnte als Dokument aufgefasst werden, das Eigenschaften zweier Mengen in sich vereint. Wenn man eine Sprache gut erkennt, dann steigen die Chancen, auch die zweite richtig zu erkennen – sowohl für den Algorithmus wie für den menschlichen Klassifizierer.

3. Wie vorher angesprochen, könnte der Linktext als Klassifikator verwendet werden.

4. Weitere Untersuchungen des Einflusses der Textmenge je Dokument und der Trainingsmenge wären hilfreich. Wenn die relativ kleine Textmenge von Internetseiten die Qualität vermindert, dann wären die größeren Dokumente von Verlagen umso interessanter. Außerdem könnte man mehrere Dokumente, etwa vom gleichen Server, für die Klassifizierung zu einem gemeinsamen Dokument zusammenfassen.

5. Auch das Ausnutzen der bekannten Hierarchie- oder Nähebeziehung von einigen Fachgebieten könnte die Fachgebietsklassifizierung verbessern.

6. Man könnte nur einfach zu erkennende Dokumente automatisch klassifizieren lassen und die schwierigen semiautomatisch oder ganz intellektuell klassifizieren. Das würde sich insbesondere bei der Spracherkennung anbieten. Dafür müsste aber ein Kriterium für einfach zu klassifizierende Dokumente gefunden werden, also für Dokumente, bei denen die Wahrscheinlichkeit eines Klassifizierungsfehlers gegen Null tendiert. Würde dies gelingen, dann könnte dieses Kriterium in den Web-Robot integriert werden. Der Robot würde nur noch einfache Dokumente aus dem Web holen. Möglicherweise müsste dann der Korpusaufbau semiautomatisch verlaufen. Ein Mitarbeiter würde die ersten fünf Dokumente intellektuell klassifizieren und der Web-Robot hätte dann Trainingsmaterial, anhand dessen er entscheiden kann, welche weiteren Dokumente leicht zu erkennen und daher zu speichern sind. Wenn der Web-Robot dazu fähig ist, dann kann er das nächste Mal die gleiche Domäne selbständig besuchen und nach geeignetem Material suchen. Damit würde sich das Korpus selbst mit klassifiziertem Material erneuern.

Zusammenfassung

Hier wurden einige Voraussetzungen für erfolgreiche Rechtsterminologie erörtert. Wir plädieren für die Verwendung von Korpora. Ein Korpus ist über das kurzfristige Ziel und das einzelne Projekt hinaus verwendbar und erhaltenswert, stellt er doch Denken und Formulieren einer bestimmten Zeit und eines bestimmten Bereichs dar. Ein Korpus sollte daher auch öffentlich zugänglich sein.[61] Traditionelle Terminologie ist im Grunde genommen erst die Kondensierung und Veredelung dieses Wissens. Mit neuen Technologien kann ein Teil der Kondensierung maschinell ablaufen, beispielsweise die Generierung und Sortierung von Kontextstellen. Damit kann die intellektuelle Arbeit auf die wirklichen Probleme in der Rechtsvergleichung, Termdarstellung oder Nutzeranpassung konzentriert werden.

Hier wurden die Quellen für ein Korpus diskutiert und die Bedeutung des Internets als Quelle hervorgehoben. Da der Nutzen eines Korpus von seiner sinnvollen Zusammenstellung aus gewichteten Klassen abhängt, haben wir einige Klassen für mehrsprachige Rechtsdokumente vorgestellt. Betonen möchten wir noch einmal, dass nicht nur der Inhalt eines Dokuments, sondern eine ganze Reihe anderer Informationen als Klassifikatoren für eine semiautomatische oder später auch automatische Klassifikation in Frage kommen. Es empfehlen sich insbesondere der Dokumentinhalt, die Metadaten, die Mitarbeiterprofile und formale Kriterien als Klassifikatoren. Alle diese Informationen sind wichtig, weil nicht alle Dokumente diese Informationen tragen oder die vorhandenen nicht für eine Klassifizierung hinreichen. Für jedes Dokument sollten also alle Informationen gespeichert werden, auch wenn einzelne Dokumente überspezifiziert sein sollten. Es ist nur wahrscheinlich, dass beim raschen Fortschreiten der Methodenentwicklung früher oder später alle Daten eine sinnvolle Verwendung finden.

Die übergreifende Erkenntnis dieses Beitrags scheint uns zu sein, dass der Aufbau eines Korpus kein unüberwindliches Hindernis ist, wenn Open Source Programme wie unter LINUX verwendet werden, auf Standards geachtet wird, allen Informationen Beachtung geschenkt wird und von einer semiautomatischen Klassifikation ausgehend der Automatisierungsgrad zunehmend erhöht wird.

[61] Erst Recht dann, wenn er Minderheitensprachen wie hier die Ladinischen Sprachen enthält.

Bibliografie

Aha D.W., Editorial- lazy learning, Artificial Intelligence Review, 1997/11, S. 1-3.

BGH für deutsches Recht, Urteil vom 11.07.2002, I ZR 255/00, Elektronischer Pressespiegel, JurPC Web-Dok. 302/2002, http://www.jurpc.de/rechtspr/20020302.htm: 10.10.2003

Brill E., Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging, Computational Linguistics, 1995.

Bonnet E., Gaussier E. und Langé J.-M., A method for automatic extraction of terms from bilingual corpora, in AVIGNON-94, 1994.

Carl M., Schaible J. und Pease C., Enhancing translation memory (TM) technologies with linguistic intelligence, MULTI-DOC Deliverable D4.1, Commission of the European Communities, Luxembourg, 1998.

Chakrabarti S., Dom B. und Indyk P., Enhanced hypertext categorization using hyperlinks, in ACM SIGMOD 1998, Seattle, Washington, US, 1998.

Charniak, E., Tree-bank grammars, in 13th National Conference on Artificial Intelligence, AAAI-96, 1996, S. 1031-1036.

Chien L.-F. und Chen C.-L., Incremental extraction of domain-specific terms from online text resources, in Bourigault D., Jacquemin C. und L'Homme M.-C., eds., Recent Advances in Computational Terminology, John Benjamins, Amsterdam, Natural Language Processing, 2001.

Cowie J., Ludovik E. und Zacharski R., An autonomous, web-based multilingual corpus collection tool, in Proceedings of the International Conference on Natural Language Processing and Industrial Applications, 1998.

Damashek M., Gaugin similarity via n-grams: Language-independent sorting, categorization, and retrieval of text, Science, 1995, 267, S. 843-848.

Day D., Aberdeen J., Hirschman L., Kozierok R., Robinson P. und Vilain M., Mixed-initiative development of language processing systems, in Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Washington D.C., 1997.

De Carolis B., De Rosis F. und Pizzutilo S., Generating user-adapted hypermedia from discourse plans, in Lenzerini M., ed., LNAI 1321, Springer, 1997.

Ekmekçioglu F., Lynch M., Robertson A., Sembok T. und Willett P., Comparison of n-gram matching and stemming for term conflation in English, Malay, and Turkish texts., Text Technology, 1996, 6, S. 1-14.

Faber P., Lopés Rodriguez C. I. und Tercedor Sánchez M. I., Utilización de técnicas de corpus en la representatición del conocimiento médico, Terminology, 2002, 7(2), S. 167-197.

Find J., Kobsa A. und Nill A., User-oriented adaptivity and adaptability in the AVANTI project, in Conference "Design for the Web: Empirical Studies, Microsoft, Redmond, WA, 1996.

Furuse, O. und Lida H., An example-based method for transfer-driven Machine Translation, in The Third International Conference on Theoretical and Methodological Issues, Empiristic vs. Rationalist Methods in MT, Montréal, 1992.

Han E.-H. S. und Karypis G., Centroid-based document classification: Analysis & experimental results, 2000, URL, http://www.cs.umn.edu/karypis: 10.10.2003.

Herberger M., Systematik des Bundesrechts, Project report, Juristisches Internetprojekt Saarbrücken, 1997,http://www.jura.uni-sb.de/BGBl/BGBLSYST.HTML: 27.9.2003.

Jain A. K., Duin R. P. und Mao J., Statistical pattern recognition: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1), S. 4-37.

Kurohashi S. und Nagao M., Building a Japanese parsed corpus while improving the parsing system, in First International Conference on Language Resources & Evaluation, Granada, Spain, 1998, S. 719-724.

Langer S., Grenzen der Sprachenidentifizierung. Tagungsband KONVENS 2002, Saarbrücken S. 99-106, S. 99, http://konvens2002.dfki.de/cd/pdf/19V-langer.pdf: 27.9.2003.

Larkey L. S. und Croft W. B., Combining classifiers in text categorization, in Proceedings of SIGIR-96, 19th (ACM) International Conference on Research and Development in Information Retrieval, ACM Press, New York, US, Zürich, CH, 1996, S. 289-297.

Manning C. D. und Schütze H., Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, London, 1999, URL, http://www.sultry.arts.usyd.edu.au/cmanning/ 10.10.2003.

Nohr H., Automatische Indexierung, Einführung in betriebliche Verfahren, System und Anwendungen, Materialien zur Information und Dokumentation, Verlag für Berlin-Brandenburg, Potsdam, 2001.

Ott S., Linking und Framing: Ein Überblick über die Entwicklung im Jahre 2002, JurPC Web-Dok. 14/2003, http://www.jurpc.de/aufsatz/20030014.htm: 23.9.2003.

Schmidt-Wigger A., Building consistent terminologies, in Proceedings of COMPUTERM'98, 1998.

Streiter O., Corpus-based parsing and treebank development, in ICCPOL 2001, 19th International Conference on Computer Processing of Oriental Languages, Seoul, Korea, 2001, S. 115-120.

Unger W., Methoden juristischer Dokumentrecherche,

http://www.juralink.de/8LITERATUR/Umgang/Recherche.htm 27.9.2003.

Voltmer L., Dr. Streiter O., Textindexierung durch beispielbasierte Termextraktion, EURAC online working paper no. 1, 2003,
http://dev.eurac.edu:8080/autoren/publs/wp1.pdf 27.9.2003.

Wendt H., Fischer Lexikon: Sprachen, Fischer Taschenbuchverlag, Frankfurt am Main, 1987.