Computational Lexicology

Ausarbeitung eines Referates, gehalten am 30. Juni 1993
Proseminar »Lexikon«, veranstaltet von B. Zimmermann,
Universität des Saarlandes, Computerlinguistik

Sascha Brawer · Stengelstraße 18 · D–66117 Saarbrücken
e-Mail: brawer@coli.uni-sb.de

Zusammenfassung

Thema dieser Arbeit ist das Umwandeln von Satzdateien für ein für Menschen gedachtes Wörterbuch in eine (relationale) lexikalische Datenbank für ein sprachverarbeitendes System. Die skizzierte Methode ist unabhängig von einem bestimmten Syntax- oder Semantikformalismus, ebenso spielt es keine Rolle, von welchem gedruckten Lexikon man ausgeht.
Besonders betont werden die Vorteile einer globalen Sichtweise (im Gegensatz zum früher verwendeten *»Localist Approach«*). Es werden Techniken vorgestellt, um vollautomatisch Über-/Unterbegriffs-Hierarchien zu erstellen sowie dem semantischen Attribut »natürliches Geschlecht« einen Wert zuzuweisen. Abschließend wird ein Verfahren gezeigt, mit dessen Hilfe man halbautomatisch weitere Mitglieder einer semantisch zusammengehörigen Gruppe ermitteln kann.
Die Grundlage für die Arbeit bilden die im Literaturverzeichnis angegebenen Artikel von Byrd et al. und von Boguarev.

Einleitung

Wohl jede Anwendung der Computerlinguistik benötigt ein Lexikon. Oft macht man sich Gedanken über dessen Struktur, aber die Frage, woher die einzelnen Einträge im Lexikon eigentlich kommen sollen, wird manchmal etwas vernachlässigt.
Natürlich kann man die Einträge von Hand eingeben. Die Leistungsfähigkeit eines sprachverarbeitenden Systems hängt aber nicht zuletzt von der Größe seines Lexikons ab; man möchte daher möglichst viele Einträge im Lexikon stehen haben, ohne dafür übermäßig Zeit und teure Arbeitskräfte einsetzen zu müssen. Nun gibt es bekanntlich bereits Lexika, an denen viele Menschen jahrelang gearbeitet haben: Von der Taschenausgabe bis zum mehrbändigen Wälzer sind in jeder Buchhandlung Wörterbücher zu finden. Der Gedanke wäre natürlich verlockend, all das dort versammelte Wissen für ein NLP-System nutzbar zu machen. Man muß dabei jedoch bedenken, daß ein solches Wörterbuch dafür bestimmt ist, von Menschen benutzt zu werden, die bereits — im Gegensatz zum Computer — über viel

Vorwissen, auch lexikalisches, verfügen. Es gilt also einige Schwierigkeiten zu bewältigen; auf der anderen Seite steht der im Vergleich zum Selbermachen gewaltig verringerte Aufwand, der sich vor allem dann lohnt, wenn man auch versucht, semantische Information aus dem Wörterbuch zu gewinnen.
Man versucht schon lange, sich gedruckte Wörterbücher für die Sprachverarbeitung nutzbar zu machen. Es kann daher sinnvoll sein, sich als erstes zu überlegen, welche Gründe dazu geführt haben, daß manche dieser frühen Ansätze keinen großen Erfolg hatten.

Woran sind frühere Versuche gescheitert?

Boguarev nennt in seinem Artikel im wesentlichen zwei Gründe, wieso frühe Versuche zur automatisierten Lexikonerstellung mißlungen sind.
Der eine Punkt mag vielleicht etwas banal tönen: Man muß sich zuerst einmal im klaren darüber sein, was für eine Ausgabe die Konversion liefern soll. Für die Syntax bestehen darüber seit langem mehr oder weniger klare Vorstellungen: Man möchte die syntaktische Kategorie kennen, morphologische Eigenschaften, vielleicht den Subkategorisierungsrahmen eines Verbs; bei Systemen, die gesprochene Sprache erkennen oder erzeugen sollen, muß man natürlich auch phonologische Information in geeigneter Weise repräsentieren. Große Schwierigkeiten bereitet jedoch die Semantik: Für welche Bedeutungsinformationen interessiert man sich überhaupt? Was ist für die Sprachverarbeitung relevant? Wie stellt man die Bedeutung eines Wortes am besten dar? Es leuchtet ein, daß es sinnvoll wäre, sich zuerst einmal klar darüber zu werden, was im Lexikon stehen soll, bevor man damit beginnt, sich eines aufzubauen. Anders formuliert: Es fehlte eine fundierte formale Theorie für die lexikalische Semantik, wobei dies (leider) teilweise auch heute noch der Fall ist.
Bei einem weiteren Punkt, der früher oft vernachlässigt wurde, kann man sich fragen, wieso nicht schon früher daran gedacht wurde: Geliefert wird einem ein Band mit allen Befehlen für die Satzmaschine — also nicht nur dem eigentlichen Text im Wörterbuch, sondern auch allen typografischen Variationen. Oft hat man nun als erstes diese ganzen Codes herausgefiltert, so daß in den folgenden Schritten nur noch der reine Text verarbeitet wurde. Übersehen wurde jeweils, daß die Typografie eines Wörterbuches ausgesprochen viel implizit kodiert; durch das Herausstreichen der entsprechenden Codes geht auch sehr viel interessante Information verloren. Zum Teil wird in Lexika sogar ausdrücklich geschrieben, durch welche Schrift, durch welchen Schriftschnitt was ausgedrückt werden soll. Zur Illustration ein Auszug aus den Umschlagseiten des Duden-Wörterbuchs [Drosdowski 1989]:

2. Gibt es zu einem Stichwort eine zweite Form, erscheint diese, durch Komma getrennt, ebenfalls halbfett gedruckt. Angaben zu dieser zweiten Form stehen in runden Klammern davor.

   Fries, der; -es, -e, (Fachspr. auch:) **[1]Frie | se,** die; -, -, …

14. Bedeutungsschattierungen, Kontextbedeutungen und die Bedeutungen der idiomatischen Ausdrücke stehen in runden Klammern […] und sind kursiv gedruckt.

   [1]Futter, das; -s [mhd. vuoter, ahd. fuotar]: *Nahrung für [Haus]tiere:* F. schneiden; Ü dieses F. (salopp; *Essen*) paßt dir wohl nicht? der Lesezirkel hat ihm neues F. (ugs.; *neuen Lesestoff*) gebracht; […]

Deutlich zeigt auch der folgende Wörterbucheintrag [Boguarev 1991] die Mittel, welche die Typografie den Lexikografen zur Verfügung stellt:

nui·sance /»nju˘s´ns||»nu˘-/ *n* **1** a person or animal that annoys or causes trouble: PEST: *Don't make a nuisance of yourself: sit down and be quiet!* **2** an action or state of affairs which causes trouble, offence, or unpleasantness: *What a nuisance! I've forgotten my ticket* **3 Commit no nuisance** (as a notice in a public place) Do not use this place as **a** a LAVATORY **b** a TIP[4]

Würde man sich ausschließlich auf den reinen Text beschränken, was früher offenbar oft genug gemacht wurde, verliert man viel Infomation. »Do not use this place as a a lavatory b a tip« ist schon für den Menschen einiges schwerer verständlich als der obige Lexikoneintrag; ein Computer wird damit vermutlich nicht mehr gerade viel anfangen können.
Im folgenden soll nun ein Prinzip gezeigt werden, nach dem man ein gedrucktes Wörterbuch in ein Lexikon für ein sprachverarbeitendes System umwandeln kann, wobei sich die Frage stellen wird, wie man das gewonnene Wissen am besten speichert. Anschließend werden einzelne Verfahren vorgestellt, um semantische Information aus den Definitionen des Wörterbuchs zu gewinnen.

Parsen der Satzdatei

Wie eben gezeigt wurde, ist die typografische Information von großer Bedeutung für das Verständnis der Wörterbucheinträge. Wie wandelt man aber die Satzdatei — einschließlich aller relevanter typografischer Codes — in eine für unsere Zwecke besser geeignete, das heißt klarere, strukturiertere Darstellung um?
Die Computerlinguistik bietet für solche Aufgaben ein seit langem erprobtes Werkzeug an: den Parser. Man läßt sich die einzelnen Einträge im Wörterbuch von einem geeigneten Parser in ein anderes Format umsetzen, man strukturiert also die Einträge. Das folgende Beispiel ist wiederum dem Artikel von Boguarev entnommen:

book[1] / … / *n* **1** a collection of sheets of paper fastgened together as a thing to be read, […] **3** the words of a light musical play: *Oscar Hammerstein II wrote the book of "Oklahoma", and Richard Rogers wrote the music* — compare LIBRETTO

Die Ausgabe des Parsers könnte zum Beispiel folgendermaßen aussehen (natürlich ist auch eine andere Ausgabe, zum Beispiel als Attribut-Wert-Matrix, denkbar):

```
entry
-hdw: book
-homograph
   -print_form: book
   -hom_number: 1
   -syncat: n
   -pronunciation
        -primary
             -pron_string: [...]
```

```
-sense_def
     -sense_no: 1
     -defn
          -def_string:  a collection of sheets of paper fastgened together as a thing to be read, […]
[…]
-sense_def
     -sense_no: 3
     -defn
          -def_string:  the words of a light musical play
     -example
          -ex_string:   Oscar Hammerstein II wrote the book of "Oklahoma",
                        and Richard Rogers wrote the music
     -explicit_xref
          -how: compare
          -implicit_xref
              -to: libretto
          -exrf_string: libretto
```

Beim Erstellen des Parsers sollte man einige Punkte beachten. Zum einen ist es klar, daß ein Parser für die Satzdatei eines Wörterbuches eine andere Token-Definition als ein »normaler« Parser benötigen wird: Die typografischen Codes stehen unmittelbar neben den Wörtern, vielleicht sind sie sogar die einzige Abgrenzung. Außerdem wird man einen solchen Parser möglichst deklarativ schreiben, da man ja nicht ganz von vorne mit seiner Arbeit beginnen möchte, wenn man ein anderes Wörterbuch mit anderen typografischen Konventionen verarbeiten will.
Ganz trivial ist dieses Parsing übrigens nicht; im nächsten Abschnitt seien daher kurz die wichtigsten Probleme skizziert.

Schwierigkeiten beim Parsen

Eine Besonderheit von Lexika sind die vielen Abkürzungen: Der Platz ist knapp, und daher ist es klar, daß man möglichst überall sparen möchte. Vor dem Weiterverarbeiten muß man daher versuchen, diese Abkürzungen aufzulösen. Aus *Abend, -essen, -land* etc. wird so *Abend, Abend-essen, Abend-land*. Man wird übrigens die Information, aus welchen Lexemen das Kompositum zusammengesetzt ist, vermutlich ebenfalls abspeichern wollen, schon wegen des Prinzips, auf keine Information allzu frühzeitig zu verzichten.
Ein weiteres Problem sind Mehrdeutigkeiten. Kapitälchen können in ein und demselben Wörterbuch, ja sogar im selben Eintrag recht unterschiedliche Funktionen haben, ebenso Klammern und anderes. Es ist nicht immer einfach, diese Ambiguitäten aufzulösen; man hofft darauf, aus dem Kontext, in dem beispielsweise ein durch Kapitälchen gekennzeichneter Querverweis auftritt, auf dessen Funktion schließen zu können.
Das bei weitem größte Problem, das für die meisten fehlerhaften Analysen verantwortlich ist, sind laut [Byrd et al. 1987] Ellipsen. In einem englisch–italienischen Lexikon kann zum Beispiel für »si sta facendo buio« 'es wird dunkel' »it is growing *o* getting dark« stehen. Zwischen welchen Ausdrücken kann man nun wählen — »it is growing« bzw. »it is getting dark«, oder ist »it is growing dark« resp. »it is getting dark« richtig? Für den Menschen mit seinem Vorwissen stellt sich eine solche Frage gar nicht erst, wohl aber für den Rechner.
Trotz dieser Hindernisse nennen Byrd et al. eine Erfolgsquote von 95%, nur fünf Prozent der Einträge wurden falsch analysiert.

Speichern in einer lexikalischen Datenbank

Im nächsten Schritt wird die durch das Parsen gewonnene Struktur in einer Datenbank abgelegt. Dies hat den Vorteil, einigermaßen unabhängig von den restlichen Teilen des Systems zu sein: Beschließt man, den für Syntax oder für Semantik zuständigen Teil auszuwechseln, muß man höchstens die Schnittstelle zur Datenbank neu schreiben; man verliert dadurch nicht das gesamte Lexikon. Das Benutzen einer Datenbank bringt zudem weitere Vorteile: bessere Wartbarkeit, Übersichtlichkeit, es können allenfalls mehrere Benutzer gleichzeitig auf das Lexikon zugreifen, usw.
Seltsamerweise schreiben Byrd et al. [Byrd et al. 1987, S. 221]: »It quickly became apparent that the usual data base architecture, characterized by a fixed number of fields for each entry and a fixed amount of space for each, would never work for dictionary entries, beause there might be one value for some fields […], but multiple values for others […], or no value at all for still others«. Diese Argumentation trifft meiner Meinung nach zumindest auf die relationale Datenbankarchitektur nicht zu: Eine gewöhnliche 1:*mc*-Relation löst das Problem mit der variablen Anzahl von Werten. Die Autoren verwenden denn auch offensichtlich eine dem relationalen Modell zumindest äußerst nahestehende Datenbank, und ihre Abfragesprache LQL sieht dem relationalen SQL zum Verwechseln ähnlich: Das »Output«-Feld von LQL entspricht den `select`- und `from`-Teilen, die »Conditions« dem `where`-Teil einer SQL-Anfrage.
Das Institut für Wissensbasierte Systeme der IBM Deutschland hat übrigens das Lexikon seines Projekts für die maschinelle Übersetzung erst kürzlich (im Mai 1992) auf eine relationale Datenbank umgestellt. Im entsprechenden Bericht [Jantzen 1992] findet sich eine ausführliche Begründung.

»Localist Approach« vs. globale Sicht

Bisher wurde noch nicht erklärt, wozu man überhaupt die ganzen Querverweise speichert. Interessiert man sich lediglich für morphologische und phonologische Informaton, kann man sich den Aufwand mit dem Parsing, wie zu Beginn bereits gesagt wurde, auch sparen. Betreibt man in seinem System jedoch eine ausgefeilte Semantik, wird das Wissen, daß *libretto* etwas mit *Buch* zu tun hat, durchaus interessant.
Setzt man die einzelnen Lexikoneinträge miteinander in Beziehung, erhält man nämlich — bei entsprechend geschicktem Vorgehen — manche Information, auf die man verzichten müßte, würde man nur die einzelnen Einträge voneinander getrennt betrachten. Sogar die Syntax läßt sich auf diese Weise weiter verbessern: Man wird durch die globale Betrachtung des Lexikons zum Beispiel verbesserte Subkategorisierungsrahmen erhalten.
Es gibt noch ein weiteres Argument gegen den *localist approach:* Man möchte wohl meistens ein System bauen, das wenigstens ein bißchen der menschlichen Sprachverarbeitung ähnelt — zumindest sollte das künstliche System nicht völlig anders als die menschliche Kognition aufgebaut sein. Nun weiß man zwar nicht gerade besonders viel über die Sprachverarbeitung im Gehirn, aber es ist doch äußerst unwahrscheinlich, daß die einzelnen Wörter im menschlihchen »Lexikon« vollständig voneinander getrennt abgespeichert sind.
Kommen war aber zu unseren Computern zurück: Wir stehen immer noch vor dem Problem,

welche Informationen wir überhaupt aus dem gedruckten Lexikon gewinnen können. In den folgenden Abschnitten werde ich auf einige Beispiele vertieft eingehen.

Semantische Hierarchien

Betrachtet man Definitionen, wie sie in einem gedruckten Lexikon stehen, fällt eine strukturelle Eigenschaft auf, die immer wieder auftritt: Häufig ist der syntaktische Kopf der Definition ein Überbegriff: *Futter* ist beispielsweise definiert als ***Nahrung*** *für [Haus-]Tiere*.
Trifft man einige Annahmen darüber, wie Definitionen in einem Lexikon im allgemeinen aufgebaut sind, kann man also weitere Information gewinnen. [Byrd et al. 1987] verwenden die Annahme, der syntaktische Kopf der Definition sei der Oberbegriff, um ein Netz von Ober- bzw. Unterbegriffen aufzubauen — aus unserem Beispiel-Eintrag würde der Computer schließen, *Nahrung* sei ein Oberbegriff von *Futter*. Daß dieses Verfahren ein Netz und keinen Baum ergibt, hat seine Berechtigung: Die einzelnen Teilbedeutungen eines Wortes können ja mehrere, ganz unterschiedliche Überbegriffe besitzen.

Bestimmen des natürlichen Geschlechts

In Sprachen wie dem Englischen, die nur ein grammatisches Geschlecht haben, kann es von Vorteil sein, das natürliche Geschlecht eines Wortes zu kennen. Mit diesem Wissen läßt sich manche Ambiguität klären, was vor allem das Auflösen von Anaphern erleichtert. Bei Sprachen wie Deutsch oder Italienisch ist das grammatische Geschlecht wichtiger, das in den Lexika ausdrücklich angegeben ist; dort stellt sich diese Frage also gar nicht erst.
Man geht ähnlich an das Problem heran wie beim oben gezeigten Erstellen von semantischen Hierarchien: Als erstes sucht man nach Gemeinsamkeiten in den Lexikoneinträgen:

aviatrix *n*	a woman aviator, called also aviatress
churchwoman *n*	a woman adherent of a church
king *n*	a male monarch of a major territorial unit

Was fällt auf? Auf [+ female] deuten Suffixe wie *-ess*, *-ette*, *-ix* und Schlüsselwörter wie *woman*, *girl*, *wife*, *daughter* oder natürlich *female*; auf [+ male] das Suffix *-man* und die Schlüsselwörter *man*, *boy*, *husband*, *son* und *male*.
Um das natürliche (semantische) Geschlecht zu bestimmen, sucht man nach diesen Suffixen und Schlüsselwörtern. Es reicht nicht, einfach mittels der oben beschriebenen Methode die Unterbegriffe von z. B. *woman* zu bestimmen, denn der syntaktische Kopf von *aviatrix* im Beispiel ist *aviator*, nicht *woman*.

Erweitern des Wissens mit mehrsprachigen Wörterbüchern

Das Analysieren mittels der in den beiden letzten Abschnitten vorgestellen Methoden ergibt wohl kaum je sämtliche Wörter, die ein bestimmtes Attribut wie [+ female] tragen. Es ist jedoch durchaus nicht unmöglich, die erhaltene Menge noch um weitere Einträge zu erweit-

ern, was am nachfolgenden Beispiel anhand des Attributs [± human] gezeigt werden soll.
Angenommen, man hat in einem einsprachig englischen Wörterbuch durch das Bestimmen der Unterbegriffe von z. B. »Mensch«, vielleicht auch durch Suche nach bestimmten Schlüsselwörtern, eine Anzahl von Wörtern erhalten, die die Eigenschaft [+human] tragen. Es ist nun nicht unwahrscheinlich, daß alle Wörter, die auf Englisch etwas Menschliches bezeichnen, dies ebenso in einer anderen Sprache, vielleicht Italienisch, tun — man nimmt daher ein zweisprachiges englisch – italienisches Wörterbuch und sucht sich alle italienischen Entsprechungen der ursprünglich erhaltenen englischen Wörter heraus. Es empfiehlt sich allerdings, die Übersetzung eines Wortes nur dann zu berücksichtigen, wenn das Wort auf italienisch monosem ist; man weiß schließlich nicht, welcher der Teilbedeutungen der italienischen Übersetzung das englische Wort entspricht.
Nun hat man also eine Menge (monosemer) italienischer Wörter, die vermutlich alle etwas Menschliches bezeichnen. Man braucht nun nur noch mit Hilfe eines italienisch – englischen Wörterbuches die Wörter ins Englische zurückzuübersetzen und hat auf diese Weise zusätzliche englische Wörter erhalten, die im ursprünglichen Lexikon nicht als [+ human] markiert waren. Vermutlich empfiehlt sich allerdings, die neu hinzugewonnen Wörter abschließend noch von Hand durchzusehen.

Schlußbetrachtung

In den obigen Abschnitten war zu sehen, daß mit den vorgestellten Verfahren (und weiteren, die in den dieser Ausarbeitung zugrundeliegenden Artikeln ausfürlich beschrieben werden) mit relativ wenig Aufwand neues lexikalisches Wissen gewonnen werden kann; Wissen, das auf »manuelle« Weise wohl nur recht mühsam zu erhalten wäre. Man sieht aber: Der grundsätzliche Nachteil, daß das Lexikon nicht für eine Maschine, sondern für einen Menschen geschrieben wurde, bleibt. Vielleicht werden mit der Zeit jedoch auch Lexika erhältlich sein, die mehr den Bedürfnissen der Computerlinguistik entsprechen.

Literaturverzeichnis

Boguarev, Branimir: Building a Lexicon: The Contribution of Computers. International Journal of Lexicography, Band 4, Nr. 3, Herbst 1991.

Byrd, Roy J. [et al.]: Tools and Methods for Computational Lexicology. In: Computational Linguistics, Band 13, Nr. 3–4, 1987.

Drosdowski, Günther [Hrsg.]: Deutsches Universalwörterbuch. Mannheim, Wien, Zürich: Dudenverlag, ^{2}1989. ISBN 3-411-02176-4.

Jantzen, Volker: Konzeptuelles Design einer mehrsprachigen lexikalischen Datenbank für die maschinelle Übersetzung. Heidelberg: Institut für Wissensbasierte Systeme [IBM], 1992. IWBS-Report 218, Mai 1992. ISSN 0938-1864.