
Termextraktion durch Beispielterme

Ansätze und Versuchsergebnisse
für eine Minderheitensprache

Leonhard Voltmer, Oliver Streiter, Daniel Zielinski, Isabella Ties

Termextraktion durch Beispielterme

Ansätze und Versuchsergebnisse für eine Minderheitensprache

Leonhard Voltmer, Oliver Streiter, Daniel Zielinski, Isabella Ties

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

Nachdruck und fotomechanische Wiedergabe
sind auch auszugsweise nur unter Angabe der
Quelle (Herausgeber und Titel) gestattet.

Riproduzione parziale o totale del contenuto
è autorizzata soltanto con la citazione della
fonte (titolo e edizione).

Termextraktion durch Beispielterme

Voltmer, Streiter, Zielinski, Ties

Abstract

This paper discusses various approaches to Term Extraction for minority languages. In particular the example-based approach is explained and applied to Ladin, a typical minority language. Results on sparse computerized language resources starting from a relatively small set of example terms are better than simple statistical approaches to term extraction.

Inhaltsangabe

In diesem Artikel werden die verschiedenen Ansätze zur Termextraktion für Minderheitensprachen vorgestellt. Auf den beispielbasierten Ansatz wird genauer eingegangen, weil er mit relativ wenigen Beispielen und ohne großes Hintergrundkorpus bereits gute Ergebnisse bringt. In Experimenten zur Termextraktion in der typischen Minderheitensprache Ladinisch lieferte der beispielbasierte Ansatz bessere Ergebnisse als einfache statistische Termextraktionsverfahren.

Einleitung

Terminologie und Terminografie beschäftigen sich mit Termen¹. Terme werden in Texten oder Kontexten verwendet. Es ist zeitaufwendig und mühsam, Term für Term aus dem Text oder Kontext zu lösen. Daher versucht die Computerlinguistik, Terme automatisch aus elektronischen Texten herauszusuchen.²

¹ Term wird hier gleichbedeutend mit Terminus verwendet, also als zusammengehöriges Paar aus einem Begriff und seiner Benennung als Element einer Terminologie. (DIN 2342 1992:3).

² Die computergestützte Termextraktion (*computer aided terminology extraction*) ist ein neueres Forschungsgebiet in der maschinellen Sprachverarbeitung (*natural language processing* oder kurz NLP) und wird oft als Hilfswissenschaft eingesetzt.

In diesem Artikel werden zunächst die verschiedenen Methoden zur Termextraktion vorgestellt (Kapitel 1). Danach werden die Vor- und Nachteile der verschiedenen Methoden für Minderheitensprachen und Terminologieprojekte ohne große elektronische Ressourcen diskutiert (Kapitel 2). Das dritte Kapitel untersucht die beispielbasierte Termextraktion für die Minderheitensprache Ladinisch und im letzten Kapitel werden die Ergebnisse bewertet.

Termextraktionsmethoden

Die computergestützte Termextraktion (*computer aided terminology extraction*) ist ein Teilgebiet der maschinellen Sprachverarbeitung (*natural language processing* - NLP).

Termextraktion holt aus einem Dokument Termkandidaten. Die computergestützte Termextraktion analysiert einen maschinenlesbaren Text und filtert die darin enthaltenen Termkandidaten mit Hilfe des Computers heraus. Termkandidaten sind Wörter oder Phrasen, wie sie in Glossaren oder Wörterbüchern stehen oder wie sie als Schlagwörter und Indizes verwendet werden.

Die Termextraktion wird im Wesentlichen für drei Zwecke gebraucht. Man kann

1. einen Index, Schlagwörter o.ä. erzeugen (*indexing*),
2. bereits bekannte Terme in Texten wieder finden (Terminologieerkennung - *term recognition*), um die Texte zu annotieren oder klassifizieren oder
3. neues Wissen automatisch erzeugen (*automatic knowledge acquisition*).

Die automatische Wissenserzeugung durch Termextraktion ist meist das Finden noch nicht beschriebener Terminologie (*term discovery*). Sollen die gefundenen Terme einen Terminologiebestand erweitern, spricht man von *term enrichment*, soll eine Terminologie ganz neu aufgebaut werden von *term acquisition*.

Die Begriffe, mit denen sich die Terminologie und Terminografie beschäftigt, werden als bereits gegeben vorausgesetzt und die Termextraktion beschränkt sich auf das Auffinden und den expliziten Erwerb dieser Terme, es werden aber keine Terme neu geschaffen.³ Das Bestimmen von Schlagworten und Indizes zum Suchen, Wiederfinden und Einteilen von Dokumenten kann man hingegen als kreativen Prozess betrachten, weil Elemente verwendet werden können, die außer-

³ Zur Einteilung der Termextraktion siehe Zielinski D., Computergestützte Termextraktion aus technischen Texten, Diplomarbeit Universität des Saarlands, Saarbrücken 2002, <http://www.iai.uni-sb.de/~mt-dept/texte/zielinski.pdf>: 11.10.2003.

halb des Indexes keine Bedeutung haben.⁴ Hier sind die Terme nur Mittel, während sie bei der Terminologieerkennung bereits selbst das Ziel darstellen. Eine Termextraktion zur Indizierung kann schlechtestenfalls unzweckmäßig oder nicht zielführend sein, eine Terminologieerkennung kann jedoch fehlerhaft sein. Daher wird praktisch nur bei der Terminologieerkennung von Hand nachgearbeitet, und zwar sehr häufig. In diesem Artikel wird nur die Terminologieerkennung behandelt. Zur Termextraktion zur Indizierung siehe Voltmer et al.⁵

Termextraktion wird in **zwei computerlinguistische Teilaufgaben** zerlegt. Einerseits muss entschieden werden, welche Teile eines Textes zusammengehören (*unithood problem*), andererseits muss erkannt werden, welche zusammengehörenden Teile tatsächlich ein Term sind (*termhood problem*). Ein Adjektiv gehört stets zu seinem Substantiv (Bsp.: indirekte Steuer), aber nicht jeder Kombination aus Adjektiv und Substantiv wird Termwert zugesprochen (nicht: hohe Steuer). Oft wird ein Text zuerst in Einheiten zerlegt, die dann auf ihre Termqualität untersucht werden.

Es gibt folgende Ansätze zur Termextraktion⁶:

- linguistische Ansätze
- statistische Ansätze und
- beispielbasierte Ansätze.

Während die linguistischen Ansätze sprachliches Wissen (z.B. morphologische oder syntaktische Informationen) einsetzen, versuchen statistische Methoden über den Vergleich vieler Textbruchstücke (z.B. nach Häufigkeitsverteilung, Assoziationskoeffizienten oder mit statistischen Testverfahren) zu Erkenntnissen über die Struktur der Texte und die Lage von Termen zu gelangen. Ersteres erfordert tiefgehendes Wissen über die Sprache, letzteres eine breite Untersuchungsgrund-

⁴ Ein einfaches Beispiel sind Wortwurzeln, die in Texten nur mit Endungen vorkommen, aber bei der Suche hilfreich sein können.

⁵ Voltmer L., Streiter O., Textindizierung durch beispielbasierte Termextraktion, EURAC Online Working Paper, Bolzano 2003, <http://dev.eurac.edu:8080/autoren/pubs/wp1.pdf>: 11.10.2003.

⁶ Eine Einteilung in linguistische, statistische und hybrid linguistisch-statistische Ansätze findet sich bei Cabré Castellvi, M. T., Estopà Bagot, R., Palatresi, J. V.: Automatic term detection: A review of current systems. In: Bourigault, Didier; Jacquemin, Christian; L'Homme, Marie-Claude (Hrsg.): Recent Advances in Computational Terminology. In: Natural Language Processing Band 2. Amsterdam, Philadelphia: John Benjamins, 2001, S. 53-89.

lage. Heute werden diese beiden Ansätze meistens zu einem hybriden Ansatz verbunden.⁷

Eine andere Möglichkeit der Einteilung ergibt sich nach der Herkunft der Information über die Termkandidaten. Wenn die Information im Termkandidaten selbst steckt (morphologische, syntaktische oder semantische Informationen), dann wird der Ansatz intrinsisch (*intrinsic approach*) genannt, wenn die Information von außerhalb des Termkandidaten gewonnen wird, extrinsisch (*extrinsic approach*). Extrinsische Information kann syntagmatisch (syntaktische oder kontextuelle Information) oder paradigmatisch sein (Information über die Beziehungen zwischen Termkandidaten und Termen).

Übersicht über Termextraktionsmethoden

| Methode | | | Methode | verwendet von |
|-----------------------|-------------|----------------|--------------------------|-------------------------------------|
| Linguistische Methode | intrinsisch | | POS-tagging und chunking | Bourrigault & Jacquemin 1999 |
| | | | Stoppwörter | Merkel & Mikael 2000 |
| | extrinsisch | syntagmatisch | volles Parsing | Arppe 1995; Soininen et al. 1999 |
| | | paradigmatisch | Termvariation | Jaquemin 1999 |
| Statistische Methode | intrinsisch | | mutual information | Church & Hanks 1989 |
| | | | likelihood ratio | Hong et al. 2001 |
| | extrinsisch | syntagmatisch | nc-Wert | Maynard Ananiadou 1999 |
| | | | Entropie | Merkel & Mikael 2000 |
| | | paradigmatisch | c-Wert | Nakagawa 2001 |
| | | | weirdness | Brekke et al. 1996 |

Linguistische Ansätze verwenden morphologische, syntaktische oder semantische Informationen aus sprachspezifischen Anwendungen. Ihr Hauptziel ist das Erkennen von Spracheinheiten. Das Sprachmodul soll die Zusammensetzung von Termen besonders effizient und genau analysieren. Das sprachliche Wissen bezieht sich beispielsweise auf die Anzahl der Wörter eines Terms, auf besondere Vor- oder Nachsilben und auf grammatikalische Anpassungen des Terms (Mehrzahl, Konjugation). Diese Analyse erledigen morphologische Analyseprogramme,

⁷ Maynard, D., Ananiadou, S: Term extraction using a similarity-based approach. In: Bourigault, Jacquemin, L'Homme: Recent Advances in Computational Terminology. Amsterdam/Philadelphia: John Benjamins, 2001, S. 261-279.

Part-of-Speech-Tagger und Parser.⁸ Dazu werden z.T. auch Stopwortlisten verwendet, in denen Wörter stehen, die nie in einer bestimmten Position (Terminfang, letztes Wort des Terms oder Mittelstellung) eines Terms auftauchen.

Statistische Ansätze der Termextraktion beruhen auf der Annahme, dass Terme aus lexikalischen Einheiten bestehen, die statistisch signifikant öfter gemeinsam auftauchen, als durch die Kombination unabhängiger Einheiten zu erwarten wäre. So können Mehrworttermkandidaten gefunden werden. Im Folgenden werden die wichtigsten Methoden kurz vorgestellt.

Die Häufigkeit des gemeinsamen Auftretens von lexikalischen Einheiten ist aber noch kein alleiniger Garant für deren Termeigenschaft, weil häufig nebeneinander stehende Funktionswörter keine Termkandidaten sind. Man verschärft daher die Annahme und sucht nicht nur nach häufigen Begriffen, sondern nach Fachbegriffen. Man nimmt an, dass Fachbegriffe eines Dokuments in diesem häufiger verwendet werden als in anderen Dokumenten. Man sucht also jene Kombination bestimmter lexikalischer Einheiten (Fachwörter) oder morphosyntaktischer Konstruktionen (Fachjargon), deren Frequenz in wenigen Dokumenten hoch, in allen Texten hingegen niedrig ist. Eine im Information Retrieval beliebte Formel dafür ist TF.IDF. In dieser Formel wird die Häufigkeit eines Terms (TF = *term frequency*) in einem Dokument durch die Häufigkeit dieses Terms in allen Dokumenten geteilt, bzw. mit der *inverted document frequency* multipliziert:

$$TF.IDF_x = \frac{TCF_x}{DF_x}$$

mit TC = Termkandidat, D = Dokument, F_x = Häufigkeit (*frequency*) von x.

Dieselbe Idee steht auch hinter der *weirdness ratio*⁹, in der relative Häufigkeiten für TF und IDF verwendet werden:

$$weirdness\ ratio_x = \frac{\frac{TCF_x}{\#\{TC\}}}{\frac{DF_x}{\sum_{d=1}^{d=m} doc_j}}$$

⁸ Die Wirkungsweise dieser Instrumente ist am leichtesten durch ein Beispiel zu demonstrieren, und zwar über die Links unter <http://www.ifi.unizh.ch/CL/InteractiveTools.html>: 13.10.2003.

⁹ Brekke M., Myking J. & Ahmad K. (1996). Terminology management and lesser-used living languages: A critique of the corpus-based approach. In Sandrini P., Ed. (1996). Proceedings of Terminology and Knowledge Engineering (TKE'96), Innsbruck. Term-Net, S. 179-189.

Häufigkeitszählungen eignen sich gut für die Berechnung ununterbrochener Zeichenketten und feststehender Phrasen. Wenn die lexikalischen Einheiten jedoch nicht unmittelbar nebeneinander stehen (z.B. wegen Partikelverben oder Einschüben in die Phrase), dann muss die Beziehung zwischen Wörtern in einem Satz bestimmt werden. Man berechnet dazu die Abstände zwischen den Wörtern. Als Abstandsmaße werden der statistische Mittelwert (*mean*), die Varianz (*variance*) oder die Standardabweichung (*deviation*) verwendet.¹⁰

Verfahren, die sich nach der Häufigkeit richten, funktionieren gut für Einwortterme, sie können aber nicht maßstäblich auf Zwei- und Dreiwortterme vergrößert werden. Wenn eine derartige Termextraktionsmethode für 10.000 Wörter gut funktioniert, dann würde die Erweiterung auf Zweiwortausdrücke 100.000.000 Wörter benötigen, um gleich gute Ergebnisse aufzuweisen. Für Termkandidaten aus drei Wörtern wären $10.000^3 = 1.000.000.000.000$ Wörter nötig. Dieser Zusammenhang ist als *sparse-data problem*¹¹ bekannt und findet sich in allen statistischen Maßen, die die Häufigkeit eines Termkandidaten verwenden.

Die Häufigkeit eines Termkandidaten ist aber nur ein Referenzmaß.¹² Man kann auch die Korrelation der lexikalischen Einheiten A und B eines Termkandidaten untereinander messen. Dazu trägt man in eine Tabelle ein, wie häufig man die vier logisch möglichen Fälle (A, B), (A, nicht B), (nicht A, B) und (nicht A, nicht B) auftreten:

1 Kontingenztabelle der beobachteten Ereignisse

| | Wort ₂ = B | Wort ₂ ≠ B | Σ |
|-----------------------|-----------------------|-----------------------|----------------|
| Wort ₁ = A | O ₁₁ | O ₁₂ | R ₁ |
| Wort ₁ ≠ A | O ₂₁ | O ₂₂ | R ₂ |
| Σ | C ₁ | C ₂ | N |

O ist *occurrence*/Auftreten. Die Zahl 1 im Index bedeutet „tritt auf“ die Zahl 2 bedeutet „tritt nicht auf“. C₁ ist die Häufigkeit von B, C₂ die von „nicht B“, R₁ die von A und R₂ die von „nicht A“. N ist die Summe aller Dokumente.

¹⁰ Das Abstandsmaß geht auf Smadja, F. zurück, zuletzt in: Retrieving collocations from text: Xtract. Computational Linguistics 19 (S. 143-177) m.w.Nachw., <http://acl.ldc.upenn.edu/J/J93/J93-1007.pdf>: 30.12.2002. Zur Berechnung ausführlicher Zielinski D., 2002, a.a.O.

¹¹ Problem der geringen Häufigkeiten oder der erforderlichen Datenmenge.

¹² Die meisten Referenzmaße sind im Perl Modul *N-gram Statistics Package* implementiert, das von CPAN kostenlos heruntergeladen werden kann.

Aus der Häufigkeit von A und B lässt sich eine Erwartung (*expectancy*) für das Zusammentreffen von A und B in einem Dokument berechnen. Ebenso lassen sich die Wahrscheinlichkeiten für (A, nicht B), (nicht A, B) und (nicht A, nicht B) aus den Summen der beobachteten Häufigkeiten schätzen. Die entstehende Wahrscheinlichkeitstabelle heißt Kontingenztabelle des Wortpaars (A, B):

2 Kontingenztabelle der erwarteten Ereignisse

| | Wort ₂ = B | Wort ₂ ≠ B |
|-----------------------|------------------------------------|------------------------------------|
| Wort ₁ = A | $E_{11} = \frac{R_1 \cdot C_1}{N}$ | $E_{12} = \frac{R_1 \cdot C_2}{N}$ |
| Wort ₁ ≠ A | $E_{21} = \frac{R_2 \cdot C_1}{N}$ | $E_{22} = \frac{R_2 \cdot C_2}{N}$ |

E ist die Erwartung (*expectance*). R, C und N sind die beobachteten Häufigkeiten der Kontingenztabelle 1. Die Zahl 1 im Index bedeutet „tritt auf“ die Zahl 2 bedeutet „tritt nicht auf“.

Je weiter der tatsächlich gemessene Wert vom errechneten Wert abweicht, umso außergewöhnlicher ist das Ereignis und umso kleiner ist die Wahrscheinlichkeit, dass es sich um Zufall handelt. Es kann sich um eine positive oder negative Abweichung handeln.

Es gibt viele verschiedene Maße, die Abweichung zwischen Tabelle 1 und 2 zu berechnen. Ein in der Korpuslinguistik häufig gebrauchtes Maß ist die *mutual information* (MI):

$$MI = \frac{O_{11}}{E_{11}}$$

Die MI ist also die Wahrscheinlichkeit des gemeinsamen Auftretens von A und B (also das Feld O_{11}) durch das Produkt der Auftretenswahrscheinlichkeiten jedes Ereignisses geteilt durch die Anzahl der Dokumente (E_{11}). Das bedeutet erstens, dass nur zwei der acht informationstragenden Felder miteinander verglichen werden. Zweitens wird keine Wahrscheinlichkeit mit einem Wert zwischen Null und Eins berechnet, sondern ein Wert auf einer nicht linearen Skala von Null bis Unendlich. Dadurch kann man mit MI-Werten nur Aussagen über höhere oder niedrigere Wahrscheinlichkeit machen, man kann diese Wahrscheinlichkeit aber

nicht quantifizieren oder mit ihr rechnen. Tatsächlich ist die MI vor allem bei kleinen Häufigkeiten sehr schlecht.¹³

Außerdem ordnen Häufigkeitsmaße nur Termkandidaten mit gleicher Wortzahl korrekt. Sobald aber Einwort- und Mehrwortterme gegeneinander konkurrieren, werden Birnen mit Äpfeln verglichen und eine Gruppe wird benachteiligt. Viele Ähnlichkeitsmaße können mehrere Worte überhaupt nicht zueinander in Beziehung setzen, weil die Berechnungsformel nicht definiert ist. Selbst das relativ einfache MI-Maß erforderte zwei dreidimensionale Kontingenztabelle, die weder statistisch noch informationstechnisch definiert sind.

Andere Assoziationsmaße (*association measures*) sind das χ^2 -Maß (= *chi*²-Maß = *chi-square* = *chi*-Quadrat), der *t-score* und die *likelihood ratio*.

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O und E beziehen sich auf obige Kontingenztabelle. i und j können die Werte 1 oder 2 annehmen. Es gibt daher vier χ^2 -Formeln.

Das χ^2 -Maß nimmt keine normalverteilten Wahrscheinlichkeiten an. Es funktioniert erst ab 5 Ereignissen oder mehr. Außerdem gibt es nur an, ob ein Zusammenhang besteht, aber nicht wie stark der Zusammenhang ist und ob er negativ oder positiv ist. Daher wird es häufig in Formeln integriert, die solche Aussagen treffen können.

Der *t-score*

$$t\text{-score} = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$$

und die *log-likelihood ratio*

$$\text{Log-likelihood} = 2 \sum_{ij} O_{ij} \cdot \log_2 \left(\frac{O_{ij}}{E_{ij}} \right)$$

sind zwar besser für kleine Häufigkeiten geeignet, aber letztere ist nicht definiert für den Fall, dass eines der Wörter eines Termkandidaten nicht zumindest auch einmal alleine auftritt.¹⁴

¹³ Die MI ist allerdings einfach zu berechnen, kann auf jeden Text angewandt werden und ist voraussetzungsloser (keine linguistischen Informationen oder Termdatenbanken nötig). So Zielinski D., 2002, a.a.O.

Insgesamt kann man festhalten, dass alle Häufigkeitsmaße Termen einen numerischen Wert zuweisen, nach dem die Termkandidaten geordnet werden. Unterhalb eines Schwellenwerts wird die Liste abgeschnitten. Die Zusammengehörigkeit von Wörtern wird nicht richtig untersucht. Zum einen werden nur Wortsequenzen von vorher bestimmter Länge untersucht (z.B. Zweiwortterme), zum anderen werden Phrasengrenzen nicht beachtet, so dass zu einem Term gehörende Worte unterschlagen oder nicht zu einem Term gehörende Worte hinzugefügt werden.

Daher versucht ein anderer statistischer Ansatz, die Grenzen der Termkandidaten zu erkennen. Wenn man die Grenze einer Nominalphrase als zwischen einem Wort und dem ersten Wort eines Termkandidaten sowie zwischen dem letzten Wort eines Termkandidaten und dem nachfolgenden Wort festlegt, dann können Termkandidaten beliebige Länge haben. Damit umgeht man das *sparse-data* Problem, denn man kann auch in kurzen Texten lange Termkandidaten finden. Wenn man Termanfang und Termende aufeinander bezieht, also den gesamten Termkandidaten als Grenze definiert, dann hat man wieder das *sparse-data* Problem. Eine Grenze wird normalerweise da gesetzt, wo die Entropie hoch ist, aber man kann im Grunde jedes Ähnlichkeitsmaß verwenden, um zwischen schwach korrelierenden Worten eine Phrasengrenze zu setzen.

Die Entropie (entropy) ist der Grad der Unordnung bzw. des Zufalls und dient als Maß für die Schwierigkeit der Vorhersage eines Ereignisses. Eine Definition dieses Maßes ist die durchschnittliche Bitlänge zur Beschreibung dieses Ereignisses. Wenn alle möglichen Ereignisse n gleich wahrscheinlich sind, dann muss zur Beschreibung des tatsächlichen Ereignisses die entsprechende Zahl zwischen 1 und n berichtet werden. Dazu benötigt man $\log n$ Bits. Wenn nicht alle Ereignisse gleich wahrscheinlich sind, dann wählt man für die häufiger auftretenden Ereignisse eine kleinere Bitfolge und für die selteneren eine längere. Dadurch kann man Ereignisse mit der Wahrscheinlichkeit p (*probability*) theoretisch durch $-\log_2 p$ Bits beschreiben.

Je größer die Entropie zwischen lexikalischen Einheiten ist, umso schwieriger ist die Voraussage der nächsten lexikalischen Einheit und umso unwahrscheinlicher ist die Zusammengehörigkeit dieser beiden Einheiten.

Evaluierung

Der konkrete Nutzen einer Termextraktion hängt stark davon ab, wie die Termextraktion in den terminografischen Prozess eingebunden wird. Objektiv

¹⁴ Daille B., Combined approach for terminology extraction: lexical statistics and linguistic filtering, Université Paris VII, 1994 = Unit for Computer Research on the English Language Technical Papers 5, Lancaster University, 1995.

vergleichbar werden verschiedene Methoden durch Bewertung der Treffergenauigkeit bzw. Präzision (*precision*), die Vollständigkeit (*recall*), eine Kombination dieser beiden zu einem Mittelwert (mean oder *F-measure*) sowie die Bewertung der Rangliste der Termkandidaten (*ranked recall*).

$$recall = \frac{\#\{\{TC\} \cap \{T_{doc}\}\}}{\#\{T_{doc}\}}$$

Diese Bewertungsmethoden kommen aus dem Informationssuche (*information retrieval*) und funktionieren formal gesehen auch für die Termextraktion. Bei der Übertragung der Bewertungsmethoden muss jedoch stets im Auge behalten werden, dass das Ziel einer Informationssuche ein anderes ist als beim Finden noch nicht beschriebener Terminologie (*term discovery*). Die Schwerpunkte sind daher anders gesetzt und die Ergebnisse müssen anders interpretiert werden.

Die **Vollständigkeit** (*recall*) einer Termextraktion ist das Verhältnis von extrahierten Termkandidaten ($TC = \textit{term candidates}$) zu vorhandenen Termen (T_{doc}). Bei einem *recall* von 80 % bleiben 20 % der Termkandidaten unentdeckt. Für eine Informationssuche sind nicht entdeckte Informationen sehr schlimm, weil diese Informationen dem Nutzer verloren gehen. Bei einer Termerkennung kann es hingegen akzeptabel sein, auf einen Teil der möglichen Terme zu verzichten, weil die nicht entdeckten Terme später noch gefunden werden können. Wenn die gefundenen Terme der Datenbank hinzugefügt wurden, kann erneut eine Termextraktion über denselben Text laufen und je nach Einbindung dieser neuen Informationen können weitere Terme gefunden werden (*bootstrapping*). Noch einfacher wäre es aber, diese Terme, die als Fachwörter per Definition auch in anderen Texten vorkommen, durch Termextraktion in weiteren Texten aufzuspüren.

Oft wird der *recall* gar nicht berechnet, weil dazu alle Terme eines Textes vollständig bestimmt werden müssten, was arbeitsintensiv und im Einzelfall schwierig ist. Man arbeitet daher oft mit relativer Vollständigkeit.

Die **Treffergenauigkeit** (*precision*) ist das Verhältnis der erkannten Termini zu den extrahierten Termkandidaten.

$$precision = \frac{\#\{\{TC\} \cap \{T_{doc}\}\}}{\#\{TC\}}$$

Bei einer *precision* von 80 % sind vier von fünf Termkandidaten Terme. Je geringer die Treffergenauigkeit ist, umso wichtiger ist die manuelle Nachbearbeitung der Ergebnisse. Bei der Informationssuche wird die Bedeutung von Dokumenten (Information) gesucht, während bei der Terminologieerkennung eine bestimmte Formalisierung (Term) von Information (Bedeutung eines Terms) gesucht

wird. Je stärker die Formalisierung ist, umso höher wird die Treffergenauigkeit sein. Wenn nur Nominalphrasen als Terme zugelassen werden, dann hat die Terminologieerkennung einen wesentlichen Vorteil gegenüber der Informationssuche. Selbstverständlich muss auch beim Vergleich verschiedener Termextraktionssysteme die jeweilige Formalisierung von Begriffen berücksichtigt werden.

Vollständigkeit und Treffergenauigkeit sind indirekt voneinander abhängig. Werden alle möglichen Wortkombinationen extrahiert, dann ist die Vollständigkeit 1 und die Treffergenauigkeit nimmt den niedrigsten Wert an (nicht Null, sondern abhängig von der Anzahl der Terme). Die Treffergenauigkeit maximiert man dadurch, dass nur der Termkandidat mit dem besten Wert extrahiert wird. Damit wird gleichzeitig die Vollständigkeit minimiert, denn ist dieser Termkandidat kein Term, dann ist die Treffergenauigkeit Null. Daher werden Vollständigkeit und Treffergenauigkeit zur Evaluierung oft zu einem gemeinsamen Wert (F-Wert, *F-score* oder *mean*) kombiniert:

$$F - score = mean = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{1}{\alpha \frac{1}{precision} + (1 - \alpha) \frac{1}{recall}} = \frac{2 \cdot \#\{\{TC\} \cap \{T_{doc}\}\}}{\#\{T_{doc}\} + \#\{TC\}}$$

Mit α = Gewichtung zwischen Vollständigkeit und Treffergenauigkeit.

Da der Zusammenhang nicht linear ist, gibt es einen optimalen F-Wert. Auch wenn die Veränderung des F-Werts oft nur gering erscheint, sollte beim Vergleich verschiedener Termextraktionsmethoden darauf geachtet werden, dass der jeweils beste F-Wert verglichen wird, auch wenn der Schwerpunkt einer Termerkennung auf der Treffergenauigkeit liegt.

Die Treffergenauigkeit wird für alle Termkandidaten berechnet. Bei statistischen Ansätzen werden aber Ergebnisse für alle möglichen Kombinationen berechnet. Ergebnisse unterhalb eines Schwellenwertes werden dann außer Acht gelassen bzw. die Trefferliste wird unten abgeschnitten. Damit wird die Treffergenauigkeit zur Kunst des Abschneidens. Deshalb ist teilweise dazu übergegangen worden, bei der Termextraktion die Rangfolge zu bewerten. Eine Termextraktion, die acht Terme findet, aber an den Stellen 3 bis 10, ist schlechter als eine Termextraktion, die die Terme an die Stellen 1 bis 8 auflistet. Dies ist die Fähigkeit, die guten von den schlechten Termkandidaten zu trennen, der *ranked recall*. Wenn r_i der Listenplatz des i -ten Termkandidaten mit $TC|TC \in \{\{TC\} \cap \{T_{doc}\}\}$, dann ist der *ranked recall* definiert als:

$$\text{ranked recall} = \frac{\sum_i^n i}{\sum_i^n r_i}$$

In einer Liste von zehn Termkandidaten, in der die acht Terme auf den Plätzen drei bis zehn sind, ist der *ranked recall*:

$$\frac{1+2+3+4+5+6+7+8}{3+4+5+6+7+8+9+10} \approx 0.69$$

Sind die acht Terme auf den ersten acht Plätzen, dann ist der *ranked recall*:

$$\frac{1+2+3+4+5+6+7+8}{1+2+3+4+5+6+7+8} = 1$$

Hier ist also jede einzelne Platzierung ausschlaggebend, und zwar umso stärker, je weiter oben der Fehler gemacht wird, weil dadurch alle späteren Termkandidaten einen Platz absacken und einen höheren Platz einnehmen.

Termextraktion durch Beispielterme

Der **beispielbasierte Ansatz** in der Verarbeitung natürlicher Sprache (*natural language processing* = NLP) zeichnet sich dadurch aus, dass das Trainingsmaterial von der gleichen Art ist wie das Ergebnis. Es gibt beispielsweise Programme, die mit Syntaxbäumen gefüttert werden, um das Erstellen von Syntaxbäumen zu lernen. Seit 1981 gibt es die Idee zur beispielbasierten maschinellen Übersetzung, bei der Übersetzungen eingegeben werden, um Übersetzungen zu finden.¹⁵ Der Vorteil beispielbasierter Ansätze gegenüber regelbasierten Ansätzen ist, dass keine abstrakten Regeln erstellt werden müssen. Die als Trainingsmaterial notwendigen Beispiele kann man also durch Auswahl aus existierenden Beispielen händisch oder automatisch erzeugen oder auch erfinden. Es ist keine komplexe Formalisierung des sprachlichen Wissens nötig. Beispiele zu Regeln und Ausnahmen können nebeneinander, also nicht-hierarchisch aufgelistet werden. Außerdem funktioniert der beispielbasierte Ansatz bereits mit wenigen Beispielen, im Extremfall mit einem einzigen. Jedes weitere gefundene Beispiel kann dem Trainingsmaterial hinzugefügt werden, wodurch das System weiter lernt und sich die Leistung weiter verbessert.

¹⁵ Somers H., Machine Translation: Latest Developments, in Mitkov R. (Hrsg.): The Oxford Handbook of computational Linguistics, Oxford University Press Oxford 2003, S. 514. Beispielbasierte Übersetzung ist nicht gleichbedeutend mit Translation Memory (TM), weil bei ersterer eine fertige Übersetzung ausgegeben wird, beim TM nur evtl. Vorschläge zu möglichen Übersetzungen, über deren Verwendung der Übersetzer selbst entscheidet.

Die Termextraktion durch Beispiele verläuft in vier Schritten:

1. Beispiele werden ausgewählt
2. Maschinenlesbare Beschreibung, wie aus Beispielen Muster zu erzeugen sind (Formalisierung der Musterart)
3. Aus den Beispielen werden termtypische Muster erzeugt.
4. Mit Hilfe der Muster werden die Texte durchsucht und Termkandidaten gefunden.

Bei der **Termextraktion durch Beispielterme** werden vorhandene Terme eingegeben, um als Extraktionsergebnis Termkandidaten zu erhalten. Wenn die Trainingsterme aus einer Termdatenbank stammen, dann werden die extrahierten Terme der expliziten oder impliziten Termdefinition dieser Datenbank entsprechen. Enthält die Termdatenbank nur Nominalphrasen, dann werden nur Nominalphrasen extrahiert. Sind auch Verbalphrasen unter den Beispieltermen, dann werden auch die extrahierten Terme dieselbe Zusammensetzung haben. Wenn man noch keine traditionell erstellten Terme hat und stattdessen Wörterbucheinträge verwendet, dann werden Terme extrahiert, die zu diesem Wörterbuch passen. Wenn das nicht gewünscht ist, dann sollte man den Teil des Wörterbuchs als Beispielterme verwenden, der der gewählten Termdefinition entspricht.

Eine Besonderheit der beispielbasierten Termextraktion ist, dass die oben beschriebene Einteilung in *termhood* und *unithood* Problem entfällt. Die zugrunde liegende Annahme ist, dass Kombinationen lexikalischer Einheiten, die den Beispieltermen hinreichend ähnlich sind, zusammengehören und selbst einen Term bilden. Auch wenn keine Formalisierung nötig ist, müssen doch Formalisierungsmöglichkeiten angegeben werden, nach denen über die Ähnlichkeit entschieden wird. Man muss also nicht die sprachliche Kodierung selbst, aber die Variablen der Kodierung angeben. Das ist bei natürlichen Sprachen z.B. die Unterteilung in Worte, die Länge der Worte, die Verwendung von Groß- und Kleinbuchstaben, die Bedeutung von Vor- und Nachsilben sowie die Verwendung unterschiedlicher Buchstaben.¹⁶ Die Regeln selbst werden aus den Beispielen automatisch erzeugt.

Ein Beispiel für die Formalisierung eines graphischen Musters ist die Regel klein-klein-klein. Dieses **Groß- und Kleinschreibungsmuster** besagt, dass lexika-

¹⁶ Im Gegensatz zur Bildschrift, der Zeichenschrift, binärer Information oder Kombinationen.

lische Einheiten dann ähnlich sind, wenn drei aufeinander folgende Worte am Wortanfang kleingeschrieben sind. Dieses Muster wurde von dem ladinischen Beispielterm *tofla de comune* erzeugt und wurde für die Extraktion ladinischer Terme verwendet.

Die Erfüllung eines graphischen Musters allein wäre natürlich nicht strikt genug. Muster anderer Art müssen hinzugefügt werden. Der Beispielterm *tofla de comun* erzeugt z.B. das **Affixmuster** **a-de-*n*. Diese Muster passt auf alle Kombinationen von drei Wörtern („-“ steht für ein Leerzeichen), in denen das erste Wort auf a endet („*“ steht für beliebige Buchstaben), dann das Wort „de“ kommt und dann ein auf n endendes Wort. Es passt daher auch auf die ladinischen¹⁷ Terme *ciasa de comun* und *contlamada de comun*. Jeder Beispielterm kann ein Muster erzeugen, aber identische Muster werden nur einmal gespeichert.

Bei den beiden Musterbeispielen fällt auf, dass es im graphischen Muster der Groß- und Kleinschreibung sehr viel weniger Varianz gibt als im Affixmuster. Dieselben Beispielterme erzeugen also sehr viel mehr Affixmuster als graphische Muster, passen aber gleichzeitig sehr viel seltener auf Terme im zu bearbeitenden Text.

Außerdem ist anzumerken, dass die Anzahl der Beispielterme, die hinter einem Muster stehen, keine Rolle spielt. Mit anderen Worten werden Regel und Ausnahme gleich behandelt. Wenn die Beispielterme also in einer Weise von der Verteilung der Terme im Text abweichen (z.B. fast ausschließlich Nominalphrasen und nur einige wenige Verbalphrasen), dann werden die extrahierten Terme doch der Verteilung im Text (ein Teil Nominalphrasen und ein Teil Verbalphrasen) und nicht der Verteilung der Beispielterme gleichen.

Zusätzliche Filter können Effizienz und Qualität der Termextraktion steigern. Man kann etwa die häufigsten Wörter eines Hintergrundkorpus als Funktionswörter definieren und die Regel aufstellen, dass Funktionswörter nie die Randstellung eines Terms einnehmen.¹⁸ Durch einen weiteren Filter kann man von vornherein ausschließen, dass Termkandidaten Satzzeichen (z.B. Punkte oder Klammern) enthalten. In unseren Versuchen (s.u.) zeigte auch ein Filter für zu kurze und zu lange Termkandidaten positive Wirkung. Die gewünschte Länge ei-

¹⁷ Die ladinischen Beispiele sind aus dem Gadertalerischen.

¹⁸ Merkel M., Nilsson B., Ahrenberg L., A phrase-retrieval system based on recurrence in Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2), S. 43-56, Kyoto 1994, <http://www.ida.liu.se/~magne/publications/kyoto--94.pdf>: 15.10.2003.

nes Termkandidaten wurde mit ± 3 Standardabweichungen von der mittleren Länge der Beispielterme festgelegt.

Termextraktion für Minderheitensprachen

Termextraktion dient dem Aufbau von Terminologie. Was für gängige Sprachen ein Geschäft ist, stellt für Minderheitensprachen ein wichtiges Element zur Bildung von Sprachbewusstsein, zur Sprachplanung und -erhaltung dar. Daher wird zwar einerseits viel Aufwand in die Verbesserung von Termextraktionsmethoden gesteckt, aber selten mit Bezug auf Minderheitensprachen. Die meisten Ansätze sind sprachspezifisch und lassen sich nur sehr bedingt oder gar nicht auf Minderheitensprachen übertragen. Das liegt zum einen an den verschiedenen Sprachstrukturen. Während für germanische und slawische Sprachen Komposita analysiert werden müssen, ist für romanische Sprachen eine analytische Zerlegung nötig.

Die Übertragung von Termextraktionsprogrammen scheitert oft auch an den fehlenden Ressourcen von Minderheitensprachen. Für statistische Ansätze fehlt es an Hintergrundkorpora, für linguistische Ansätze an expliziten morphologischen, syntaktischen oder semantischen Informationen. Darüber hinaus mangelt es fast immer auch an den Linguisten und Computerfachleuten, die solche Ressourcen herstellen könnten. Sprachverarbeitungsressourcen für Minderheitensprachen sind wirtschaftlich nicht interessant und die politischen und wirtschaftlichen Mittel sind daher spärlich gesät.

Zwei Experimente zur Termextraktion in Sprachen mit wenig Sprachverarbeitungsressourcen¹⁹ benutzten die *weirdness-ratio*. Brekke et al.²⁰ verwenden einen fachsprachlichen Text mit 10.000 norwegischen Wörtern und ein 100.000 Wort großes allgemeinsprachliches Hintergrundkorpus. Die *weirdness-ratio* funktioniert zwar nur für Einwortterme, das mag für Norwegisch mit vielen Komposita aber passend sein. Auch auf Walisisch wurde die *weirdness-ratio* angewandt, wo-

¹⁹ Norwegisch im Beispiel Fn. 20 gehört nicht zu den Minderheitensprachen, sondern zu den Sprachen mit wenig Sprachverarbeitungsressourcen. Eine weitere Kategorie vor allem in EU-Förderprogrammen sind die weniger verbreiteten und unterrichteten Sprachen (*less widely used less taught languages*) LWULT, zu denen alle Sprachen außer Englisch, Französisch, Spanisch und Deutsch gehören.

²⁰ Brekke M., Myking J., Ahmad K., Terminology management and lesser-used living languages: A critique of the corpus-based approach. In Sandrini P., (Hrsg.) 1996, Proceedings of Terminology and Knowledge Engineering (TKE'96), Innsbruck, TermNet, S. 179-189.

bei der Fachtext und das Hintergrundkorpus beide je 100.000 Wörter umfassen.²¹

Daille et al.²² berichten von zwei Experimenten mit Madagassisch.²³ Im ersten Experiment wurde eine statistische, sprachunabhängige Termextraktionsmethode verwendet.²⁴ Die Treffergenauigkeit ist mit ca. 75 % sehr hoch, die Vollständigkeit mit nur 240 Termkandidaten aus 25.000 Wörtern aber äußerst gering. In einem zweiten Versuch wurde eine hybrid linguistisch-statistische Methode angewandt. Dafür musste zunächst ein Wörterbuch erstellt und ein POS-tagger trainiert werden. Mit 819 Termkandidaten war die Ausbeute höher, es fehlen aber Angaben über die Treffergenauigkeit.

An diesen Versuchen zeigt sich, auf welche Schwierigkeiten die Termextraktion mit nichteuropäischen Sprachen stößt, zeigt aber zugleich Möglichkeiten zur Integration von linguistischen Ansätzen. Es darf vermutet werden, dass die meisten Terminologieprojekte für Minderheitensprachen von vornherein auf eine Termextraktion verzichten und traditionell arbeiten.

Experimente zur Termextraktion mit Beispieltermen im Ladinischen

Die Termextraktion lief über eine ladinische²⁵ Gemeindeordnung aus Gröden mit 994 Wörtern. Eine Terminografin konnte mit der traditionellen Methode 113 Terme finden. Dann wurden maschinell alle 19019 möglichen Wortkombinationen erzeugt und Schritt für Schritt mit einfachen Regeln reduziert. Die Regeln waren: Keine Satzzeichen im Termkandidaten, keine Funktionswörter im Termkandidaten und keine extreme Längenabweichung.

²¹ Ahmad K., Davies A.E., Weirdness in Special-language Text: Welsh Radioactive Chemicals Texts as an Exemplar. *Journal des Internationalen Instituts für Terminologieforschung* 1994, Band 5, Nr. 2, S. 22-52.

²² Daille B., Enguehard C., Jacquin C., Raharinirina R. L., Ralalaoherivony B. S., Lehmann C. Traitement automatique de la terminologie en langue malgache in K. C. et al., (Hrsg.), *Ressources et évaluation en ingénierie des langues, Actualités scientifiques - Universités Francophones*, S. 225-242, De Boek and Larcier S.A. 2000.

²³ Madagassisch gehört zum westindonesischen Zweig der austronesischen Sprachfamilie. Madagassisch ist Haupt- und Nationalsprache in Madagascar und wird von 12 Millionen Menschen gesprochen.

²⁴ Enguehard C., Pantera L., Automatic natural acquisition of a terminology, *Journal of Quantitative Linguistics* 1994, 2(1), S. 27-32.

²⁵ Grödnerisch ist eine rätoromanische Sprache wird im Grödnertal in den italienischen Dolomiten von ca. 5000 Personen gesprochen.

Termextraktion mit einfachen Methoden

| Methode | #{TC} | <i>recall</i> | <i>precision</i> | <i>mean</i> | <i>ranked recall</i> |
|------------------|-------|---------------|------------------|-------------|----------------------|
| ohne | 19019 | 1 | 0,0056 | 0,011 | 0,011 |
| kein Satzzeichen | 8023 | 1 | 0,0134 | 0,026 | 0,0179 |
| Funktionswörter | 6289 | 0,946 | 0,016 | 0,033 | 0,030 |
| Längenabweichung | 2419 | 0,9375 | 0,044 | 0,084 | 0,055 |
| Muster | 489 | 0,848 | 0,202 | 0,326 | 0,388 |

Als Beispielterme wurden drei Mengen ausprobiert. Die ersten 1225 Beispiele waren traditionell ausgewählte Terme aus der Termdatenbank, die zweite Menge waren bearbeitete Wörterbucheinträge und die dritte Menge ein Mischung aus beiden. Mit oben genannten Formalisierungsmöglichkeiten wurden Muster erzeugt und die Ähnlichkeit wurde als gegeben erachtet, wenn ein Termkandidat mindestens mit einem graphischen Muster (Groß- und Kleinschreibung) und einem Affixmuster übereinstimmte.²⁶

Vorversuche zu den Beispieltermen

| Methode | #{TC} | <i>recall</i> | <i>precision</i> | <i>mean</i> |
|--------------|-------|---------------|------------------|-------------|
| Fachbegriffe | 299 | 0,7321 | 0,284 | 0,410 |
| Wörterbuch | 322 | 0,75 | 0,269 | 0,396 |
| Kombination | 390 | 0,839 | 0,248 | 0,386 |

Im Folgenden wurde mit den Fachbegriffen der Datenbank als Beispiele weitergearbeitet. Um die *precision* zu erhöhen wurden nun die einfachen Methoden mit den Mustern kombiniert.

Termextraktion mit kombinierten einfachen Methoden

| Methode | #{TC} | <i>recall</i> | <i>precision</i> | <i>mean</i> |
|---------------------------|-------|---------------|------------------|--------------|
| Muster | 489 | 0,848 | 0,202 | 0,326 |
| Muster + kein Satzzeichen | 489 | 0,848 | 0,202 | 0,326 |
| Muster + Funktionswörter | 390 | 0,839 | 0,204 | 0,386 |
| Muster + Längenabweichung | 477 | 0,839 | 0,203 | 0,328 |

Aus der Tabelle geht hervor, dass die Kombination mit dem Filter für Funktionswörter der *recall* um 1 % sinkt, aber die *precision* um 4 % steigt.

²⁶ Die Termextraktion wurde auf zwölf Sprachen und weitere Sprachkombinationen erweitert und läuft online sowohl über Webseiten als auch selbst geschriebenen Text kostenlos unter der Adresse <http://dev.eurac.edu:8080/cgi-bin/index/TermExtract>: 13.10.2003.

Nun wurde der Versuch mit der weirdness-ratio für Einwortterme durchgeführt. Dabei gehen von vornherein 46 % aller Terme verloren, weil Rätoromanisch eine analytische Sprache ist. Der selbe Ansatz kann aber für synthetische Sprachen gut funktionieren.

Termextraktion mit der *weirdness-ratio* (w-r)

| Methode | #{TC} | <i>recall</i> | <i>precision</i> | <i>mean</i> | <i>ranked recall</i> |
|---|-------|---------------|------------------|-------------|----------------------|
| w-r | 345 | 0,544 | 0,188 | 0,280 | 0,363 |
| w-r + Muster | 312 | 0,544 | 0,210 | 0,303 | 0,415 |
| w-r + Längenabweichung | 316 | 0,544 | 0,205 | 0,298 | 0,400 |
| w-r + Muster + Längenabweichung | 302 | 0,544 | 0,215 | 0,308 | 0,416 |
| w-r + Funktionswörter | 281 | 0,544 | 0,225 | 0,318 | 0,367 |
| w-r + Funktionswörter + Muster | 250 | 0,544 | 0,254 | 0,346 | 0,404 |
| w-r + Funktionswörter + Muster + Längenabweichung | 249 | 0,544 | 0,255 | 0,347 | 0,404 |

Alle Methoden finden gleich viele richtige Terme, daher ist die restriktivste Methode hier die beste. Nun wurde der Versuch auch noch für die Mutual Information und Zweiwortterme durchgeführt.

Termextraktion mit der Mutual Information (MI)

| Methode | #{TC} | <i>recall</i> | <i>precision</i> | <i>mean</i> | <i>ranked recall</i> |
|--------------------------------|-------|---------------|------------------|-------------|----------------------|
| MI | 807 | 0,098 | 0,013 | 0,024 | 0,007 |
| MI + Muster | 160 | 0,098 | 0,063 | 0,074 | 0,064 |
| MI + Muster + Längenabweichung | 69 | 0,098 | 0,144 | 0,110 | 0,144 |

Auch hier war die größte Einschränkung die erfolgreichste Methode. Man kann die Einwortmethode und die Zweiwortmethode als sich ergänzend ansehen. Sie laufen nicht beide über alle 19019 Möglichkeiten, sondern jede nur über einen eigenen, sich nicht überlappenden Teil. Die Ergebnisse einer Kombination von *weirdness-ratio* und die *Mutual Information* ergeben sich daher rechnerisch.

Es wären $249 + 69 = 318$ Termkandidaten extrahiert worden, von denen $63 + 10 = 73$ auch Terme waren, so dass der *recall* 0,646 gewesen wäre, die *precision* 0,230 und der *mean* 0,339. Nur der *ranked recall* ist nicht errechenbar, solange

man sich nicht für ein Maß entscheidet, wie die Termwahrscheinlichkeit von Einwort- und Zweiworttermen zueinander berechnet wird.

Damit ist diese Kombination etwas schlechter im recall und daher auch im *mean* im Vergleich zur besten beispielbasierten Termextraktion.

Bewertung

Die Ergebnisse zeigen, dass die Termextraktion durch Beispielterme eine Alternative zu den statistischen und linguistischen Methoden darstellt, die besonders für Minderheitensprachen interessant sein dürfte. Die notwendigen Voraussetzungen sind geringer als bei den anderen Ansätzen. Einige aus einem Printmedium abgetippte oder ausgedachte Beispiele genügen, um Muster zu finden, die fast alle Terme aus dem Text herausholen können. Die Eingabetexte können im Unterschied zum statistischen Ansatz beliebig kurz sein. Sie müssen nicht wie beim linguistischen Ansatz durchanalysiert werden. Der Nachteil dieser Anspruchslosigkeit des beispielbasierten Ansatzes ist, dass Wortkombinationen, die oberflächlich Ähnlichkeiten zu Termen aufweisen, als Termkandidaten ausgewählt werden. Das empfiehlt diesen Ansatz für eine Kombination mit anderen, weil sie unterschiedliche Stärken und Schwächen haben. Auch praktisch können die durch Beispielterme gefundenen Termkandidaten jederzeit mit einem Hintergrundkorpus oder mit linguistischen Hilfsmitteln weiter gefiltert werden.

Die Termextraktion kann unter <http://dev.eurac.edu:8080/perl/all.tar.gz> heruntergeladen werden. Sie läuft unter der graphischen Oberfläche von BISTRO auf der Webseite <http://dev.eurac.edu:8080/cgi-bin/index/TermExtract>.

Bibliographie

Brekke M., Myking J., Ahmad K., (1996): Terminology management and lesser-used living languages: A critique of the corpus-based approach, in: Sandrini P. (Hrsg), Proceedings of Terminology and Knowledge Engineering (TKE'96), Innsbruck, TermNet, S. 179-189.

Cabré Castellví M.T., Estopà Bagot R., Palatresi J. V., (2001): Automatic term detection: A review of current systems, in: Bourigault D., Jacquemin C., L'Homme M.C. (Hrsg.), Recent Advances in Computational Terminology, in: Benjamins J., Natural Language Processing Band 2. Amsterdam/ Philadelphia, S. 53-89.

Daille B., (1994): Combined approach for terminology extraction: lexical statistics and linguistic filtering, Université Paris VII= Unit for Computer Research on the English Language Technical Papers 5, Lancaster University, 1995.

Daille B., Enguehard C., Jacquemin C., Raharinirina R. L., Ralalaoherivony B. S., Lehmann C.: Traitement automatique de la terminologie en langue malgache in K. C. et al. (Hrsg.), (2000): Ressources et évaluation en ingénierie des langues, Actualités scientifiques - Universités Francophones, S. 225-242, De Boek and Larcier S.A.

Enguehard C., Pantera L., (1994): Automatic natural acquisition of a terminology, Journal of Quantitative Linguistics , 2(1), S. 27-32.

Maynard D., Ananiadou S.: Term extraction using a similarity-based approach, in: Bourigault D., Jacquemin C., L'Homme M.C. (Hrsg.), Recent Advances in Computational Terminology, in: Benjamins J., Natural Language Processing Band 2. Amsterdam/ Philadelphia, 2001, S. 53-89.

Merkel M., Nilsson B., Ahrenberg L., (1994): A phrase-retrieval system based on recurrence, in: Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2), S. 43-56, Kyoto

<http://www.ida.liu.se/~magne/publications/kyoto--94.pdf>: 15.10.2003.

Smadja F.: Retrieving collocations from text: Xtract, Computational Linguistics 19 (S. 143-177) m.w.Nachw.,

<http://acl.ldc.upenn.edu/J/J93/J93-1007.pdf>: 30.12.2002

Somers H.: Machine Translation: Latest Developments, in: Mitkov R. (Hrsg.), (2003): The Oxford Handbook of computational Linguistics, Oxford University Press Oxford, S. 514

Voltmer L., Streiter O., (2003): Textindizierung durch beispielbasierte Termextraktion, EURAC Online Working Paper, Bolzano,

<http://dev.eurac.edu:8080/autoren/publs/wp1.pdf>: 11.10.2003.

Zielinski D., (2002): Computergestützte Termextraktion aus technischen Texten, Diplomarbeit Universität des Saarlands, Saarbrücken,

<http://www.iai.uni-sb.de/~mt-dept/texte/zielinski.pdf>: 11.10.2003.