Grenzen der Sprachenidentifizierung

Stefan Langer

CIS, Universität München
stef@cis.uni-muenchen.de

Zusammenfassung: Die automatische Sprachenidentifizierung für elektronische Dokumente, deren Mindestlänge eine bestimmte Wortzahl überschreitet und die regulären Text enthalten, kann als weitgehend gelöstes Problem betrachtet werden. Problematisch ist die korrekte Erkennung der Sprache jedoch für alle Typen von Sprachidentifizierungssystemen bei der Verarbeitung von kurzen und/oder nicht regulären Dokumenten. Zudem gibt es bestimmte Spracheigenschaften (etwa die Existenz einer sehr nahe verwandten Sprache), die eine Entscheidung erschweren können. Dieser Artikel versucht, die Grenzen der Performanz von Sprachenidentifizierungssystemen auszuloten., indem Dokumenten- und Spracheigenschaften analysiert werden, die zu Identifizierungsproblemen führen. In einigen Fällen können dann auch Methoden aufgezeigt werden, die eine höhere Erkennungsrate ermöglichen.

1 Einleitung

Sprachenidentifizierung ist die Ermittlung der Sprache(n), in denen ein Dokument verfasst ist. Die automatische Sprachenidentifizierung ist in den letzten Jahren durch die Zunahme von multilingualen elektronischen Textsammlungen, allen voran das Internet, zu einem unentbehrlichen Werkzeug zur Dokumentenklassifikation geworden. Auf der Identifizierung von Sprache und Zeichensatzkodierung eines elektronischen Dokumentes bauen weitere Verarbeitungsschritte auf - von der korrekten Indizierung über die Grundformenreduktion bis hin zur automatischen Übersetzung und der intelligenten Generierung von Zusammenfassungen.

Angesichts der Tatsache, dass die bekannten Algorithmen - allen voran der N-Gramm-Algorithmus (beschrieben etwa in Cavnar&Trenkle 1994) - sehr gute Erkennungsraten haben, könnte man annehmen, dass die Identifizierung der Sprache eines elektronischen Texts ein abgeschlossenes Problem darstellt. Im Folgenden versuche ich zu jedoch zeigen, dass diese Aussage nicht generalisiert werden kann, da sie nur für reguläre Dokument mit bestimmten Mindestanforderungen gilt. Für weitere Fortschritte in der Sprachenidentifizierung ist es wichtig, die Parameter herauszuarbeiten, entlang derer Verbesserung der Algorithmen erfolgen können. Zu diesem Zweck werden im Folgenden Eigenschaften von Dokumenten und Sprachen bzw. Sprachgruppen herausgearbeitet, die einen Einfluss auf die Erkennungs- und Diskriminierungsraten von Sprachenidentifizierungssystemen haben. Der Fokus liegt dabei auf eine qualitativen, nicht auf einer quantitativen/ statistischen Analyse der noch bestehenden Probleme.

1.1 Algorithmen zur Sprachenidentifizierung

Bevor ich zur Analyse der Problemquellen komme, ist es kurz notwendig, nochmals kurz die vorherrschenden Ansätze zur Sprachenidentifizierung zu vergegenwärtigen. Unter den bisher beschriebenen Systemen lassen sich zwei grundsätzlich verschiedene Ansätze unterscheiden (s.a. Grefenstette 1995):

1. **N-Gramm-Techniken**: Das trainierte System kennt die Wahrscheinlichkeit gängiger Bytefolgen in elektronischen Texten für die Sprachen (und Zeichensatzkodierungen), die erkannt werden sollen. Die meisten Sprachenidentifikationssysteme dieser Art arbeiten mit Trigrammfolgen, genauer gesagt mit Byte-Trigrammen. Eine ausreichende Größe des Trai-

ningskorpus vorausgesetzt, macht dieser Ansatz das Training des Systems relativ einfach. Die Ergebnisse dieser Sprachenidentifizierer sind für Standardtexte (dieser Begriff wird weiter unten erläutert werden) sehr gut. Allerdings können bei der Unterscheidung sehr nah verwandter Sprachen (d.h. Sprachen mit ähnlichen Byte-N-Grammen) Probleme auftreten; Ursachen fehlerhafter Klassifizierung einzelner Dokumente lassen sich in Systemen, denen dieser Ansatz zugrunde liegt, nicht ohne weiteres erkennen und eliminieren.

2. **Wortbasierte Ansätze**: Das trainierte System kennt Wortformen in den Sprachen (und Zeichensatzkodierungen), die erkannt werden sollen und deren Frequenz im Trainingskorpus. Im einfachsten Fall enthält das Lexikon nur hochfrequente Wortformen. Es kann allerdings beliebig mit weniger häufigen Wortformen ergänzt werden, um die Erkennungsrate für kurze Dokumente zu erhöhen. Das Training solcher Systeme ist aufwändiger als bei dem zuerst genannten Ansatz, da für eine optimale Performanz die Wörterbücher manuell (oder halbautomatisch) gereinigt werden sollten, um Internationalismen und ins Trainingskorpus eingestreute Wörter aus anderen Sprachen weitgehend aus den Wortlisten zu eliminieren (ein 100% reines Trainingskorpus ist in vielen Fällen nicht ohne weiteres verfügbar). Während solche Wortformen bei langen Dokumenten keine größeren Probleme bereiten, können Sie bei sehr kurzen Dokumenten leicht zu einer Fehlklassifikation führen. Die Erkennungsrate - aber auch der Trainingsaufwand - eines solchen Systems wächst mit der Größe der Wörterbücher für jede Sprache. Soll das System nur reguläre Dokumente mit einer garantierten Minimallänge und vollständigen Sätzen erkennen, reicht es aus, wenn das Wörterbuch häufige Funktionswörter enthält. Ist dies nicht garantiert - wie etwa im Falle der Dokumentklassifizierung für eine Internetsuchmaschine - sollte das Wörterbuch hoch- und mittelfrequente Wortformen aller Wortarten enthalten.

So unterschiedlich die beiden genannten Ansätze sind, haben sie doch eine wichtige Gemeinsamkeit: die Klassifizierung beruht letztlich auf dem Abgleich von Byte-Sequenzen im Dokument mit Byte-Sequenzen in einem vor der Laufzeit erstellten Lexikon von spezifischen Sequenzen des entsprechenden Typs. Bei N-Gramm-Algorithmen sind dies Sequenzen mit fixer Länge, bei wortbasierten Ansätzen von Leer- und Satzzeichen begrenzte Einheiten variable Länge.

Die vorgestellten Untersuchungen beruhen auf dem Einsatz des hybriden Elexir Sprachenidentifizierers (Langer 2001), der in der Internetsuchmaschine AllTheWeb eingesetzt wird. Da die Ausführungen nicht auf die Identifizierung asiatischer Sprachen eingehen (für die der wortbasierte Algorithmus nicht anwendbar ist), kann der Elexir Sprachenidentifizierer im Rahmen dieses Dokuments als wortbasiert angesehen werden, denn nur für Chinesisch, Japanisch, Koreanisch und Thai wird ein N-Gramm-basierter Ansatz verwendet.

Insofern Aussagen gemacht werden über Differenzen zwischen wörterbuchbasierten und den statistischen, N-Gramm basierten Algorithmen, wurden diese Wertungen, wenn nicht eine andere Quelle explizit genannt wird, auf Grundlage der Identifizierungsresultate von TextCat von van Noord (http://odur.let.rug.nl/~vannoord/TextCat/) und der Demonstration des Basis Tech Sprachenidentifizierers Euclid abgegeben (http://demos.basistech.com/euclid/).

1.2 Anforderungen an Sprachenidentifizierungssysteme

Als optimale Performanz eines Sprachenidentifizierungssystems kann gelten, dass die Resultate möglichst nahe an der Zuordnungsfähigkeit eines idealen menschlichen Klassifizierers liegt. Die Grenze der automatischen Spracherkennung liegt dort, wo auch eine angenommene ideale Vergleichsperson, die alle Sprachen beherrscht, die vom System erkannt werden, nicht mehr in der Lage ist, ein Dokument zu klassifizieren, weil dieses keine Textsequenzen enthält, die einer Sprache zuzuordnen wären.

Neben der Frage nach der theoretisch möglichen Erkennungsrate stellt sich die Frage nach der maximal notwendigen Erkennungsrate, die zunächst aus dem Blickwinkels des Endbenutzers eine Retrieval-Systems betrachtet

werden soll. Die Perspektive, die wir zunächst einnehmen, ist diejenige des Benutzers (oder der Benutzerin) einer Internetsuchmaschine, der sich der Sprachvoreinstellung bedient. Ähnliches gilt natürlich für Benutzer eines beliebigen Text-Retrieval Systems. Die Sprachenidentifikation dient in solchen Retrievalsystemen auf der Benutzeroberfläche dazu, der Benutzerin die Auswahl von Dokumenten zu ermöglichen, die in einer Sprache verfasst sind, die sie lesen kann. Im Zusammenhang mit der Vorwahl der Sprache durch den Benutzer ist eine fehlerhafte oder nicht erfolgte Erkennung der richtigen Sprache problematisch für solche Dokumente, die sich manuell einer Sprache (oder mehreren) zuordnen lassen, und bei denen das Verständnis der Inhalte vom Verständnis der Sprache abhängt.

Unproblematisch ist die nicht erkannte Sprache aus der Benutzerperspektive dagegen bei Dokumenten, die sich auch ohne Sprachkenntnisse verarbeiten lassen - dies gilt vor allem, wenn primär nicht-sprachliche Informationen vermittelt werden (graphische und/oder Audioinformation), die ohne Begleittext auskommen.

Die falsche oder nicht erfolgte Identifizierung der primären Sprache eines Dokumentes ist v.a. problematisch, wenn dem Benutzer Dokumente vorenthalten werden, die für die Suchanfrage relevant sind, d.h. bei zu geringer Ausbeute (Recall). Aber auch die Präzision ist im Fall von Internet-Suchmaschinen nicht zu unterschätzen. Deren Wichtigkeit liegt vor allem an der extrem ungleichen Verteilung der verschiedenen Sprachen auf dem Internet - mit einer Prozentzahl von >60% für englischsprachige Dokumente, und am sehr geringen Anteil von Dokumenten in kleineren Sprachen. Würden etwa 0,5% der englischen Dokumente fälschlicherweise etwa als westfriesisch identifiziert, wäre damit (bei einem Anteil westfriesischer Dokumente am Gesamtweb von < 0.1%) nur jedes sechste der als Westfriesisch klassifizierten Dokumente tatsächlich in dieser Sprache verfasst - die Suche für diese Sprache gäbe wenig tatsächlich relevante Ergebnisse. Die Präzision ist damit für kleinere Sprachen ebenso zentral wie die Ausbeute.

Die Möglichkeit der Auswahl einer Zielsprache für die Suche - d.h. die Präsentation auf der Benutzeroberfläche - ist nicht die einzige Aufgabe von Sprachenidentifizierungssystemen für elektronische Textsammlungen. Gleichzeitig ist die Sprache auch ein Eingabeparameter für weitere Verarbeitungsschritte, wie Grundformenreduzierung, maschinelle Übersetzung u.a. Außerdem dienen die meisten Sprachenidentifizierungssysteme gleichzeitig dazu, den Zeichensatz und die Zeichensatzkodierung eines Dokumentes zu erkennen - d.h. sie sind Sprach- und Kodierungsidentifizierer. Diese häufige Verbindung hat Ihren Grund darin, dass beide Aufgaben eng miteinander verzahnt sind - um die Sprache eines Dokuments zu erkennen - d.h. um Wörter oder Buchstabensequenzen mit einer Referenzliste abzugleichen - muss a) entweder bekannt sein, wie die Byte-Sequenzen eines Dokumentes interpretiert werden, d.h. die Zeichensatzkodierung muss bekannt sein - oder b) die zu identifizierenden Wörter bzw. Sequenzen müssen mit Referenzsequenzen in allen Zeichensatzkodierungen abgeglichen werden. Da a) meist nicht oder nicht mit ausreichender Sicherheit gegeben ist, wird die Strategie b) gewählt, die dann natürlich ermöglicht, gleichzeitig die Zeichensatzkodierung zu identifizieren.

Die gleichzeitige Anforderung der Zeichensatzerkennung bedeutet nun allerdings, dass die Performanz dieser Systeme auch über das direkt für den Benutzer der Sprachenauswahl relevante Ergebnisse hinaus gesteigert werden sollte, um die Sprache und Kodierung jedes Dokumentes identifizieren, für die das möglich ist, denn nur die korrekte Erkennung des Zeichensatzes und der Zeichensatzkodierung macht eine korrekte Indizierung möglich. Dies ist zunächst nicht direkt einsichtig, wenn man die Perspektive eines Sprechers einer westeuropäischen Sprachen einnimmt - bei Dokumenten in diesen Sprachen reicht es zur korrekten Indizierung aus, die Standardzeichensatzkodierung anzunehmen (das auf dem Internet als Standard festgelegte ISO-8859-1). Sobald aber der Bereich der westeuropäischen Sprachen verlassen wird, ist ein korrekte Indizierung nur noch bei Erkennung des korrekten Zeichensatzes möglich. Dies gilt bereits für mitteleuropäische Sprachen mit einem erweiterten lateinischen

Zeichensatz, und für Dokumente in nicht-lateinischen Zeichensätzen ist eine Indizierung ohne die korrekte Erkennung der Kodierung überhaupt nicht mehr möglich[1]: Ein Dokument etwa in kyrillischen Buchstaben ist nur indizierbar, wenn vorher festgestellt wurde, dass es sich a) um ein Dokument im kyrillischen Zeichensatz handelt und b) welche der zahlreichen gängigen Zeichensatz-kodierungen für kyrillischen Text vorliegt. Dies gilt nicht nur für Dokumente mit regulärem Text, sondern auch für Eigennamenlisten u.ä. - d.h. die Anforderungen an den Erkennungs-algorithmus gehen über die Anforderungen aus der Benutzerperspektive hinaus.
Es lässt sich festhalten: die maximale Performanz eines Sprachenerkennungssystems ist erreicht, wenn es in der Lage ist, die Erkennungsrate einer angenommenen idealen menschlichen Testperson zu erreichen, die Sprachenerkennung manuell vornimmt. Aus der Sicht des Endbenutzers ist die Sprachen-identifizierung nur relevant, wenn das Dokument tatsächlich nur bei entsprechenden Sprachkenntnissen zu verarbeitende Information enthält. Kommt allerdings die Zeichensatzkodierungserkennung als Anforderung hinzu, ist es wünschenswert, dass auch sehr kurze Dokumente, die u.U. nur Eigennamen und nicht-reguläre Texteinheiten enthalten, korrekt identifiziert werden.

1.3 Standarddokumente / Standardtext

Unter Standarddokumenten werden im folgenden Dokumente verstanden, die folgende Anforderungen erfüllen:

- Sie enthalten mindestens 20 Wörter Text;
- Sie enthalten regulären Text, d.h. Sätze oder Teilsätze, die in etwa eine normale Verteilung (d.h. eine dem Trainings-korpus entsprechende Verteilung) der verschiedenen Wortarten aufweisen - d.h. Produktlisten, Fußballergebnisse u.ä. gelten nicht als regulärer Text;
- Sie sind einsprachig.

Allein für solche Standarddokumente - und mit der Einschränkung auf relativ einfach zu unterscheidende Sprachen - gilt die in der Einleitung erwähnte Aussage, dass die Sprachenidentifizierung als abgeschlossenes Problem angesehen werden kann. Im Folgenden soll gezeigt werden, welche Faktoren im Detail eine Rolle spielen, wenn die Performanz aktueller Systeme hinter die maximale Erkennungsrate zurückfällt.

2 Dimension Dokumentenlänge

Keine Schwierigkeiten für die Sprachen-identifizierung bereiten Standarddokumente einer Länge von mehr als 20 Wörtern, die reguläre Text enthalten - d.h. sie enthalten zumindest einige gängige Funktionswörter oder sonstige hochfrequente Wortformen. Hier liegen die Erkennungsraten aller bekannten Algorithmen über 99%, wenn von der Unterscheidung extrem eng verwandter Sprachen abgesehen wird.

Dokumente unter einer Länger von 20 Wörtern bereiten zunehmend Probleme. Tabelle 1 zeigt die Identifikationsrate des Elexir Sprachen-identifizierers (Langer 2001). Die Statistik beruht auf 2000 (pro Länge genau 100) zufällig ausgewählten Dokumenten aus dem Index von AllTheWeb - darunter auch Dokumente, die keinen regulären Text im Sinne unserer Definition enthalten. Dokumente, die manuell keiner Sprache zuzuordnen waren, wurden allerdings nicht berücksichtigt.
Es wird deutlich, dass die Identifikationsrate für Dokumente mit weniger als 10 Wörter drastisch sinkt - die Wahrscheinlichkeit, dass sich mehrere Wörter aus dem Lexikon im Dokumententext finden, ist hier relativ klein.

[1] Die meisten Internet-Dokumente enthalten Metainformationen über den Zeichensatz im Dokument selbst, oder die Information wird über den Http-Header geliefert. Allerdings finden sich immer wieder Dokumente, wo diese Information nicht vorliegt oder augenscheinlich falsch ist.

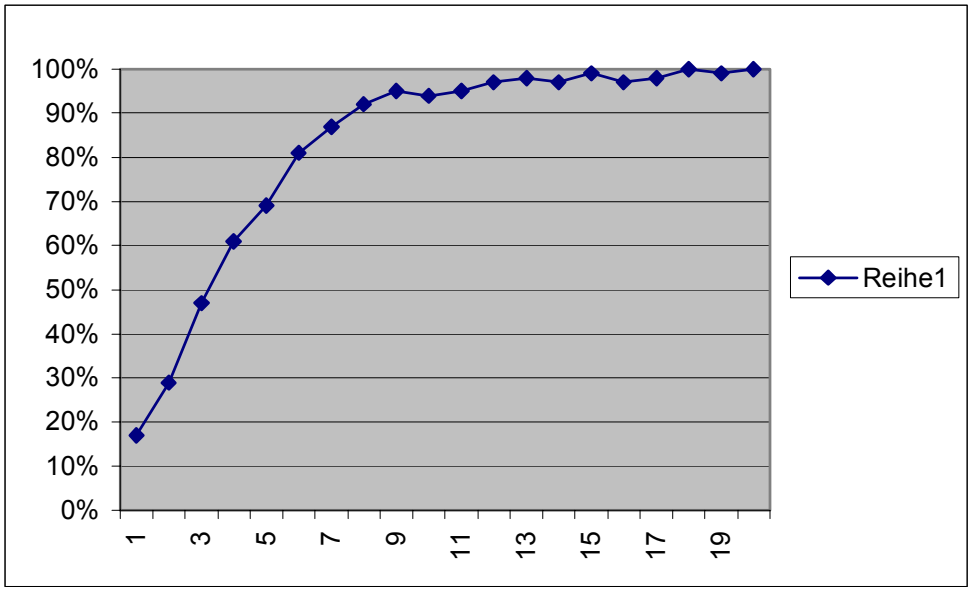

Diagramm 1: Performanz des Elexir Sprachenerkenner für Dokumente mit weniger als 20 Wortformen (gemischtes Korpus; auch nichtreguläre Dokumente)

Probleme mit Sprachenerkennung für kurze Dokumente beschreiben auch Cowie, Ludovik&Zacharski (1999), wobei sie feststellen, dass N-Gramm-Algorithmen weniger Probleme mit der Identifikation der Sprache sehr kurzer Texte hatten als wortbasierte - sie gingen allerdings von einer Liste von nur 1000 Wörtern pro Sprache aus, während die Listen des Elexir Sprachenidentifizierers im Durchschnitt 3000 Wörter umfassen und hier für die wichtigeren Sprachen eine minimale Abdeckung der Trainingsdokumente von 70% erreicht wurde. Aus den Tabellen im genannten Artikel wird deutlich, dass auch der beste der genannten Algorithmen eine Fehlerrate von über 2% für Dokumente mit weniger als 100 Byte aufweist, und für Dokumente mit weniger als 20 Byte Text die Fehlerrate auf über 10% steigt. Allerdings handelt es sich bei dem Testset um Teile aus längeren Dokumenten, d.h. die meisten der kurzen Dokument dürften Bruchstücke regulären Texts sein, während bei dem von uns benutzten Testset sehr viel nicht-regulärer Text vorhanden war.

Die Erkennungsrate eines N-Gramm-Systems kann für kurze Texte nicht ohne weiteres in die Höhe geschraubt werden, wohingegen der wörterbuchbasierte Ansatz durch die Vergrößerung des Wörterbuchs für jede Sprache - und damit die Erhöhung der Abdeckung von Texten - eine Verbesserung der Performanz ermöglicht.

3 Lexik und Morphologie

Problematisch für wortbasierte Systeme sind Dokumente, die sich von Standarddokumenten dadurch unterscheiden, dass sie eine stark abweichende Lexik aufweisen. Zum problematischen Vokabular zählen in erster Linie Eigennamen, aber auch seltene Wortformen (Fachwortschatz, Archaismen, Neologismen). N-Gramm-basierte Algorithmen habe hier nur Probleme wenn die graphemische Struktur des vom Trainingskorpus abweichenden Wortschatzes nicht dem regulären Wortschatz entspricht.

3.1 Eigennamen

Zu den Dokumenten, deren Sprache häufig nicht oder falsch identifiziert wird, gehören solche, bei denen ein großer Prozentsatz der Lexik aus Namen (etwa in Form von Namenslisten) besteht, insbesondere wenn es sich um internationale Namenslisten handelt. Besonders häufig im Internet sind Dokument mit Listen von Personennamen, Produktnamen, Firmennamen und geographischen Namen.

Wörterbuchbasierte Algorithmen können diese Namen nur teilweise (und dann evt. falsch) zuordnen, N-Gramm-basierte Algorithmen werden durch Namen, die nicht der Sprache des

Standardtexts im Dokument zuzuordnen sind, auf die falsche Fährte gelockt. Ein Beispiel ist die Verarbeitung einer Bordeaux-Weinliste eines deutschen Händlers mit einem kleinen Prozentsatz deutschem erklärendem Text - als Sprache wird hier aufgrund der Eigennamen und Zusätze von N-Gramm-Sprachenidentifizierern Französisch erkannt, obwohl zum Verständnis des Dokuments und zum Abwickeln einer Bestellung Deutschkenntnisse (aber keine Französischkenntnisse) erforderlich sind.
Als Lösung für diese Probleme kommen in Betracht:

- Abgleich mit Wörterbuch von Eigennamen und Eliminierung der Eigennamen aus der Evaluierungsbasis für die Identifizierung;
- gezielte Aufnahme sprachspezifischer Eigennamen bei wortbasierten Systemen (dies wäre allerdings im Falle des Beispiels der Weinliste problematisch, da hier ja gerade die erkannten Eigennamen zu einer Fehlklassifikation führen);
- Strukturanalyse zur Erkennung von Listen und Tabellen und Bevorzugung von laufendem Text und Tabellenüberschriften in der Sprachenerkennung.

Problematisch sind Dokumente mit Eigennamen vor allem dann, wenn sie Text enthalten, der nicht in ISO-Latin-1 verfasst sind, da hier die Zeichensatzkodierung zur korrekten Indizierung notwendig ist. Eigennamen in Sprachen mit solchen Zeichensätzen müssen daher in das Wörterbuch aufgenommen werden. Bei nichtlateinischen Zeichensätzen sind auch sehr häufig internationale Namen (etwa Bush, New York) in den nativen Zeichensatz transliteriert - auch solche Eigennamen können dann ins Wörterbuch aufgenommen werden, da sie nicht zu einer Verschlechterung des Ergebnisses der Sprachenidentifizierung führen, sofern die Transliteration sprachspezifisch ist.
Ähnliche Probleme wie mit Eigennamen ergeben sich für wortbasierte Sprachenerkenner mit Listen seltener Wörter (etwa Terminologielisten) - N-Gramm Algorithmen haben hiermit weniger Schwierigkeiten.

3.2 Wortwiederholungen

Wortwiederholungen irreführender Sequenzen stellen für beide Typen von Systemen ein Problem dar.

Bei der ersten Version des Elexir Sprachenidentifizierers kam es zu Fehlklassifikationen, wenn irreführende Buchstabenkombinationen häufig wiederholt wurden. So wurde ein an sich englisches Dokument mit einer Tabelle die sehr oft die Buchstabenkombination "na" als Abkürzung für "not applicable" (nicht zutreffend) enthielt, als Tschechisch einstuft, da "na" ein hochfrequentes tschechisches Funktionswort ist. Als Lösung für dieses Problem wurde der prozentuelle Anteil jeder Wortform an der Gesamtzahl der Wörter im Text auf 3% festgesetzt; weitere Vorkommen einer Wortform werden nicht mehr zur Erkennung herangezogen.

3.3 Morphologie

Der Parameter Morphologie spielt vor allem für wortbasierte Sprachenidentifizierer eine entscheidende Rolle. Für Sprachen mit vielen Wortformen pro Wort (hochflektierende oder agglutinierende Sprachen) muss die Liste der Wörter wesentlich größer sein, um dieselbe Abdeckung zu erreichen. So ist die Wortliste für das Russische zur Erreichung einer 70%igen Abdeckung des Trainingskorpus ca. 10000 Wörter, für das Norwegische (Bokmål und Nynorsk) nur etwa 4000 Wörter. Ähnliches gilt in geringerem Umfang für Sprachen mit einem hohen Anteil an Komposita im laufenden Text, wie das Deutsche.
Für N-Gramm-Sprachenidentifizierer spielt dieser Parameter keine entscheidende Rolle.

4 Andere Parameter

4.1 Mehrsprachige Dokumente

Mehrsprachige Dokumente sind dann ein Problem für Sprachenerkenner, wenn man nicht von einer (quantitativ) primären Sprache des Dokuments ausgehen kann - und dann, wenn das gesamte Dokument en bloc verarbeitet wird, ohne auf Formatierung, Abschnitts- und Satzgrenzen zu achten.
Mehrsprachige Dokumente mit einer starken quantitativen Differenz der einzelnen Sprachanteile, etwa ein deutsches Dokument mit einer englischen Fußnote oder Copyrightverweis oder mit einigen englischen Zitaten, bereiten bei der Sprachenidentifizierung keine Probleme - hier wird die primäre Sprache erkannt.

In Dokumenten mit einer quantitativ ausgewogenen Verteilung der verschiedenen Sprachen ist es notwendig, die Struktur des Dokuments zu analysieren. In den meisten Fällen sind die Sprachen relativ klar voneinander abgesetzt - d.h. ein Sprachenerkenner, der für mehrsprachige Dokumente ausgelegt ist, muss einzelne Textabschnitte voneinander trennen, und die Sprache einzelner Abschnitte bzw. Sätze erkennen, um daraus die Sprache(n) des Gesamtdokuments zu ermitteln. Hierfür ist allerdings ein wesentlich höherer Aufwand nötig als für einfache Sprachenerkennung ohne Strukturanalyse.

4.2 Unterscheidung eng verwandter Sprachen

Problematisch ist die Sprachenerkennung auch für sehr nah verwandte Sprachen - wobei hinzukommen muss, dass die verwandten Sprachen auch den gleichen Zeichensatz verwenden. So ist die Unterscheidung von Serbisch (in kyrillischem Zeichensatz) und Kroatisch (in lateinischem Zeichensatz) natürlich problemlos möglich.
Ein Beispiel für relativ nahe verwandte Sprachen mit identischem Zeichensatz sind Dänisch und Norwegisch (Bokmål): In diesen beiden Sprachen können ganze Sätze homograph sein - entsprechend schwierig ist die Unterscheidung bei kurzen Texten. Bei einem wortbasierten Algorithmus ist darauf zu achten, dass sich einerseits möglichst viele derjenigen Wörter in den Lexika befinden, die sich in den beiden Sprachen unterscheiden und andererseits alle Wortformen, die sich in beiden Sprachen gleichen, auch in den Wörterbüchern für beide Sprachen auftauchen, um nicht akzidentielle Klassifikationen kurzer Texte zu erzeugen, weil ein Wort zufällig nur im Trainingskorpus für eine Sprache war.
In noch extremerem Maße gilt dies für Sprachpaare wie Indonesisch-Malaiisch, die sich nur in einem Bruchteil des Vokabulars unterscheiden.
Bei der Unterscheidung nahe verwandte Sprachen bietet das wortbasierte Vorgehen gegenüber N-Gramm-Ansätzen den Vorteil, dass durch Veränderungen der Wörterbücher die Diskriminierungsrate manuell verbessert werden kann.
Bei eng verwandten Sprachen ist die korrekte Sprachenerkennung allerdings auch aus Benutzersicht nur eingeschränkt wichtig, da hier evt. falsch identifizierte Texte noch lesbar sind. Allerdings kann hier eine (eher emotionale) Anforderung hinzukommen - gerade Sprecher nahe verwandter Sprachen legen häufig großen Wert auf eine Abgrenzung zur jeweils anderen Sprachgruppe.
Aus der Perspektive der Dokumentenindizierung bereitet eine gewisse Fehlerrate bei der Diskriminierung nahe verwandter Sprachen kein Problem, da die Zeichensatzkodierung auch bei falscher Sprachenzuordnung richtig ist. Für eine Weiterverarbeitung (etwa Grundformenreduzierung) kann eine falsche Zuordnung jedoch negative Konsequenzen haben.

4.3 Nicht-berücksichtigte Sprachen und Zeichensätze

Dokumente in nicht berücksichtigen Sprachen sollten von einem Sprachenerkenner in den meisten Fällen als nicht-erkannt klassifiziert werden, und nicht fälschlicherweise einer der unterstützten Sprachen zugeordnet werden. Durch die höhere Präzision von wörterbuchbasierten Systemen, ist dieses Ziel hier leichter zu erreichen. Eine Ausnahme bilden eng verwandte Sprachpaare, bei denen eine Sprache nicht unterstützt wird; hier ist nicht ganz korrekte Zuordnung der nicht unterstützten Sprache relativ unproblematisch (etwa Bosnisch zu Kroatisch).
Eine unterstützte Sprache wird nur in unterstützten Zeichensätzen erkannt. Auch jede Transliteration, etwa eine Transkription in westlichen Text aus kyrillischer Quelle wird nur dann erkannt, wenn der Sprachenerkenner speziell hierfür trainiert wird. Liegen Algorithmen zur Transliteration vor, lässt sich eine solche Unterstützung von Transliterationen (und auch weiteren Zeichensätzen) in einem wortbasierten System relativ einfach durch simple Konversion der Wortliste vornehmen. Es ist darauf zu achten, dass bei nahe verwandten Sprachen, für die dieselben Zeichnsätze verwendet werden, jeweils dieselben Zeichensatzkodierungen unterstützt werden, da sich sonst Fehlklassifikationen ergeben.

5 Zusammenfassung

Sprachenidentifizierung für Standarddokumente kann als weitgehend gelöstes Problem ange-

sehen werden. Für Dokumente mit abweichende Eigenschaften gilt dies jedoch nicht. Fortschritte im Bereich der Sprachenidentifizierung aus der Sicht des Benutzers wie der Betreiber von Retrievalsystemen sind damit nach wie vor wünschenswert. Allerdings wird ein weiterer Fortschritt in der Klassifizierung bisher schlecht erkannter Dokumente durch einen hohen Aufwand für die Feinjustierung der Daten und Algorithmen erkauft, die für die Identifizierung herangezogen werden. Für jedes System ist das Verhältnis von Nutzen und Aufwand bezüglich der Anwendungsdomäne abzuwägen.

Im Falle von Internetsuchmaschinen ist die Notwendigkeit einer Weiterentwicklung klar. Hier tauchen verschiedenste Dokumententypen und sehr viele Dokumente mit ungewöhnlicher Struktur und ungewöhnlichem Wortschatz auf. Bei zahlreichen dieser Dokumente ist die richtige Identifizierung der Sprache wünschenswert - auch wenn dies einen höheren Aufwand bei der Entwicklung und Training der Software notwendig macht. Die teilweise im ELEXIR Sprachenerkenner implementierten, teilweise nur angesprochenen Lösungsvorschläge müssen vor dem Hintergrund einer Internet-Suchmaschine allerdings in hohem Maße nach ihrer algorithmischen Komplexität beurteilt werden. Da die Indizes größerer Suchmaschinen inzwischen auf mehr als 2 Milliarden Dokument verweisen (Stand Sommer 2002), müssen die Sprachenerkennungsmodule extrem schnell arbeiten - d.h. allzu aufwändige Analysen der Struktur von Dokumenten und zeitintensive Algorithmen können nicht eingesetzt werden.

Die qualitative Analyse hat gezeigt, dass wortbasierte und N-Gramm-basierte Sprachenerkenner unterschiedliche Stärken und Schwächen haben:

- wortbasierte Sprachenerkenner haben insgesamt einen höheren Trainingsaufwand; sie haben eine schlechtere Performanz bei Dokumenten mit einem hohen Anteil an seltenem Vokabular. Andererseits weisen sie den Vorteil auf, dass die Erkennungsrate ohne weiteres durch die Vergrößerung der Wörterbücher erhöht werden kann; manuelle Ergänzungen und Korrekturen sind relativ einfach möglich;
- N-Gramm basierte Sprachenerkenner sind einfacher zu trainieren und können mit unbekannten Vokabular umgehen, da sie nur Buchstabensequenzen zur Klassifizierung heranziehen. Fehlerquellen in der Erkennung lassen sich allerdings nicht ohne weiteres identifizieren und beseitigen; eine manuelle Verbesserung der Daten ist nicht oder kaum möglich.

Eine Kombination aus beiden Algorithmen wurde bisher meines Wissens nicht evaluiert. Will man an den Vorteilen des wortbasierten Algorithmus festhalten, liegt es jedoch nahe, nicht erkannte Wortformen im Text aufgrund eines N-Gramm-Algorithmus zu analysieren, und dies zur Optimierung der Erkennungsrate in den problematischen Fällen einzusetzen.

Bibliographie

Cavnar, W. und J. M. Trenkle (1994): *N-Gram-Based Text Categorizatio*n. In: Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, , 11-13 April 1994, S. 161-175.

Cowie, J, E. Ludovik und R. Zacharski (1999): *An Autonomous, Web-based, Multilingual Corpus Collection Tool*. Proceedings of the International Conference on Natural Language Processing and Industrial Applications, S. 142-148.

Grefenstette, G. (1995): *Comparing two Language Identification Schemes*. In: Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data, JADT 95, Rom.

Langer, Stefan (2001): *Sprachen auf dem WWW*. Proceedings der GLDV-Frühjahrstagung 2001, S. 85-91.

Van Noord, G. (ohne Jahr): TextCat language guesser. Internet-Dokument und Online-Demo: http://odur.let.rug.nl/~vannoord/TextCat/