67th IFLA Council and General Conference
August 16-25, 2001

Code Number: **099-183(WS)-G**
Division Number: **VI**
Professional Group: Information Technology **Workshop**
Joint Meeting with: -
Meeting Number: 183
Simultaneous Interpretation: **-**

Mehrsprachiger Zugang bei Informationssystemen

Carol Peters
Istituto di Elaborazione della Informazione, CNR
Pisa, Italien
E-mail: carol@iei.pi.cnr.it

Páraic Sheridan
MNIS-TextWise Labs, Syracuse
New York, NY, USA
E-mail: paraic@textwise.com

Kurzfassung:

Mit dem schnellen Anwachsen der weltweiten Informationsgesellschaft hat der Aufgabenbereich der Bibliotheken sich ausgedehnt auf alle Arten von Informationssammlungen, alle Arten von Speichermedien unter Verwendung vieler verschiedener Zugangsmöglichkeiten. Die Nutzer heutiger Informationsnetzwerke und digitaler Bibliotheken sind nicht länger durch geographische und räumliche Grenzen eingeschränkt und möchten einschlägige Informationen finden, abrufen und auswerten, gleich in welcher Sprache sie gespeichert wurden. Deshalb hat man in den letzten Jahren viel Mühe aufgewendet auf die Erforschung und Entwicklung von Werkzeugen und Verfahren für mehrsprachigen Informationszugang (Multilingual Information Access = MLIA). Dieser Kurs wird den Teilnehmern einen Überblick über die interessantesten Entwicklungen in diesem Bereich geben. Zu den behandelten Themen gehören Zeichenverschlüsselung, spezielle Anforderungen bei einzelnen Sprachen und Schriften, Wege zur Auffindung und Darstellung sowie Techniken zur sprachübergreifenden Informationsgewinnung und die Bedeutung der Quellen.

1. Einführung

Die weltweite Informationsgesellschaft hat die Methoden der Wissensaneignung, -verbreitung und des Wissensaustauschs grundlegend verändert und bewirkt eine schnelle Revolution in der Welt der

Bibliotheken. Nutzer der international im Netz angebotenen Sammlungen müssen in der Lage sein, einschlägige Informationen aufzufinden, abzurufen und zu verstehen unabhängig davon, in welcher Sprache und Form sie gespeichert wurden. Viele Nutzer haben einige Fremdsprachenkenntnisse, aber deren Umfang reicht vielleicht nicht aus um Fragen zu stellen, die ihre Informationsbedürfnisse hinreichend ausdrücken. Solche Nutzer zögen großen Vorteil daraus, könnten sie ihre Fragen in der Muttersprache stellen, denn dann könnten sie Informationen aus einschlägigen Dokumenten beurteilen und abrufen, auch wenn diese nicht übersetzt sind. Andererseits können einsprachige Nutzer Übersetzungshilfen benutzen als Hilfe beim Verständnis ihrer Suchergebnisse in einer anderen Sprache.

Deshalb hat man in den letzten Jahren viel Mühe aufgewendet auf die Erforschung und Entwicklung von Werkzeugen und Verfahren für mehrsprachigen Informationszugang (Multilingual Information Access = MLIA) und sprachlich übergreifende Informationsgewinnung (CrossLanguage Information Retrieval = CLIR). Dies ist ein komplexer, viele Wissensgebiete umfassender Bereich, in dem die Verarbeitung natürlicher Sprachen und Informationsgewinnungstechniken zusammenlaufen. Ziel des Kurses ist die bessere Kenntnis der damit verbundenen Fragen und der verschiedenen Bestandteile, die für wirksame mehrsprachige Schnittstellen und sprachenübergreifende Werkzeuge in digitalen Bibliothekssystemen nötig sind. Dieser Vortrag bietet einen kurzen Abriß der behandelten Hauptthemen. Für eine eingehendere Erörterung sei der Leser verwiesen auf die Nr. 1 im Literaturverzeichnis.

2. Mehrsprachige Textverarbeitung

Bei der Informationsgewinnung erhält man gewöhnlich einen Eindruck von dem gesuchten Text durch `IndizierungsmerkmaleA, gewonnen aus der Dokumentensammlung oder dem Anfragetext eines Nutzers. Vereinfacht gesprochen besteht dieser Suchprozess aus vier grundlegenden Schritten: Zeichenumwandlung, Kennzeichnung von Wörtern, Entfernen von `StopwörternA und Normalisierung der verbleibenden sinntragenden Begriffe. Während diese Verarbeitungsschritte schon ausgiebig untersucht wurden im Zusammenhang mit Suchanfragen in englischsprachigen Texten, ergeben sich doch neue Herausforderungen beim Zugang zu Informationen in mehreren Sprachen.

Spracherkennung: Da Verarbeitung von Texten zwecks Gewinnung von Indexmerkmalen häufig Schritte erfordert, die sprachspezifische Kenntnisse voraussetzen, ist es erst einmal wichtig, die Sprache des Textes zu identifizieren, falls diese noch unbekannt ist. Bisher gab es viele verschiedene Zugangsweisen für das Problem der Sprachidentifizierung bei allgemeinen Texten. Die Verfahren reichen von der Erkennung des Vorhandenseins spezifischer Zeichen in den Texten [2] über das Auffinden spezieller Buchstabenmarkierungen [3] bis zum Aufspüren bestimmter Wörter [4]. Ein allgemein gebräuchliches Verfahren der Sprachidentifizierung bei Zugang zu mehrsprachigen Texten ist die Verwendung sprachspezifischer Stopwörter, mit deren Hilfe die Sprache der Texte festgestellt wird [5].

Zeichenverschlüsselung: Während die meisten westeuropäischen Sprachen von dem Standardcodeschema ISO-8859-1 (Latin-1) erfaßt werden, erfordert der mehrsprachige Zugang zu Sprachen mit nicht-lateinischen Schriften die Befassung mit dem Thema der Dokumentkodierung. Die Kodierung einer Sprache, insbesondere die Kodierung des Zeichensatzes, der verwendet wird zur Darstellung des Alphabets der Schrift einer gegebenen Sprache, bestimmt die graphische Darstellung eines Schriftstücks und seiner binären Wiedergabe. Eine Zeichenkodierung ist daher spezifisch für ein jeweils gegebenes Alphabet, und in vielen Fällen gibt es mehrere Darstellungsformen für ein Alphabet (z.B. die Kyrilliza im Russischen). Abhängig von der Anzahl der Zeichen, die zur Darstellung einer Sprache benötigt werden, kann ein Kodierschema auf einem einzigen Byte beruhen (z.B. im Deutschen) oder erfordert eine Zwei-Byte-Kodierung (z.B. im Chinesischen).

Bei dem Versuch, ein einziges Kodierschema für die graphische Darstellung aller Sprachen der Welt zu schaffen, hat das UNICODE-Konsortium (www.unicode.org) den UNICODE-Standard entwickelt. Dieser bietet ein Zeichenkodierungssystem für den Austausch, die Verarbeitung und die Darstellung

geschriebener Texte der verschiedenen Sprachen der modernen Welt. Bei der Textverarbeitung für mehrsprachigen Zugang benutzen die mit UNICODE arbeitenden Systeme standardisierte Befehlssätze, um die ursprüngliche Kodierung von Texten (z.B. Shift-JIS für Japanisch) in ein UNICODE-Format als einzige Standarddarstellung zu übertragen (z.B. UTF-8).

Sprachspezifische Kennzeichnung: Nach der Identifizierung der Sprache eines Textes und der Vereinheitlichung des Zeichenkodes besteht der nächste Schritt in der Identifizierung der einzelnen Wörter. Zwar geht dies bei vielen Sprachen problemlos wegen der Leerstellen zum Zweck der Abgrenzung, doch verbinden und verketten viele Sprachen die Wörter, um neue zu bilden. Im Extremfall werden keine Spatien zwischen den Wörtern im Text benutzt (z.B. im Chinesischen und Japanischen), so daß der Kennzeichnungsvorgang alle Wortgrenzen bestimmen muß. In solch einem Fall wird üblicherweise ein Wörterbuch der in dieser Sprache vorkommenden Wörter benutzt, um korrekte Wörter zu bestimmen. Es wird ein Verarbeitungsprozess durchgeführt, bei dem ein Satz des Textes gescannt wird, um jene Wörter aus dem Wörterbuch zu finden, die den im Text gefundenen Zeichen gänzlich entsprechen. Während dieser Phase der Herausarbeitung wird die Zeichensetzung entfernt, und Bindestriche zwischen Wortbestandteilen werden eliminiert.

Um die Anzahl der Indexierungsmerkmale zu reduzieren, die in die Wiedergabe eines Textes aufgenommen werden müssen, werden Wörter mit geringem Informationsgehalt oft ausgeschlossen - sogenannte `StopwörterA wie *the* und *at* im Englischen. Da zwischen 30% und 50% der Wörter eines Textes zu solch einer Stopwortliste gehören, hat ihre Nichtberücksichtigung einen bedeutenden Einfluß auf den Index der Suchwörter. Stopwörter können gewöhnlich in jeder Sprache leicht aufgefunden werden aufgrund ihrer Zugehörigkeit zu bestimmten Wortarten (z.B. Bestimmungswörter, Präpositionen) oder ihres häufigen Vorkommens in einer Textprobe.

Wortnormalisierung: Der letzte Schritt der Textverarbeitung zwecks Indexierung für die Suche erfordert die Normalisierung der sinntragenden Wörter, die nach Entfernung der Stopwörter übrigbleiben. Die gebräuchlichste Normalisierungsmethode besteht in der Reduzierung der Wörter auf eine Stammform durch Weglassen von Suffixen und Flexionsformen. Im einfachsten Fall entfernt der Stammbildungsalgorithmus einfach die Standardendungen (z.B. _-s>, _-es>, _-ation> im Englischen) solange, bis die kürzeste Form übrigbleibt. Der bekannteste Algorithmus dieser Art wurde für das Englische von Porter entwickelt [6], ähnliche Algorithmen gibt es für andere Sprachen [5]. Eine Alternative ist die stärker linguistisch bestimmte morphologische Analyse des Textes zur Bestimmung der Wurzelformen der Wörter. Wortnormalisierung führt zu größerer Wirksamkeit sowohl bei der Textverarbeitung als auch dem Suchvorgang. Dies gilt besonders für europäische Sprachen, die eine weitaus stärkere Flexionsmorphologie aufweisen als das Englische.

In der Normalisierungsphase ist es üblich, besonders wenn es um mehrsprachigen Zugang geht, aus mehreren Wörtern bestehende Redewendungen als einzelne Indexierungsmerkmale herauszulösen, so daß Redewendungen eher als Einheiten übersetzt werden können denn als einzelne Wörter. In vielen Fällen liefert eine Wort-für-Wort-Wiedergabe keine korrekte Übersetzung (versuchen Sie beispielsweise *fast food* ins Französische oder Deutsche zu übersetzen). Redewendungen können in einem Text erkannt werden durch Abgleich mit einem Wörterbuch bzw. phraseologischen Lexikon oder durch statistische Verfahren, die häufig zusammen vorkommende Wörter als mögliche Redewendungen erkennen.

Zusammenfassung: Die bei mehrsprachiger Textverarbeitung gemachten Schritte hängen ab von den beteiligten Sprachen und dem Zugang im Allgemeinen, der bei einem vorgegebenen System zu mehrsprachigem Informationszugang gewählt wird. Die aus der Textverarbeitungsphase resultierenden Indexmerkmale müssen übereinstimmen mit dem Verfahren, das zur Formulierung von Fragen in einer Sprache gebraucht wird bei Anwendung auf Texte in vielen anderen Sprachen. Daher braucht man unbedingt Kenntnisse der verschiedenen Zugänge bei sprachenübergreifender Informationsgewinnung und der verschiedenen Verfahren, die bei jedem Zugang gewählt werden.

3. Zugänge zur sprachübergreifenden Informationsgewinnung

Im Wesentlichen müssen bei sprachübergreifender Informationsgewinnung Methoden entwickelt werden, die erfolgreiches Durchsuchen von Texten in mehreren Sprachen ermöglichen und die untersuchten Dokumente entsprechend ihrer Relevanz bewerten. Bei nur einer Sprache geschieht dies traditionellerweise durch irgendeine Art von Wortvergleich und -gewichtung, bei Texten in mehreren Sprachen haben wir das zusätzliche Problem, die Wörter über mehrere Sprachen hinweg zu vergleichen und zu gewichten. Dies erfordert die Einbeziehung irgendeines Hilfsmittels zur Übersetzung der Sprache, in der die Anfrage erfolgt, in die Sprache der Dokumente (oder umgekehrt), dazu kommt das Problem der Mehrdeutigkeit, das schon bei der Suche in nur einer Sprache besteht, aber bei mehreren Sprachen sehr viel größer wird. Drei Hauptzugangsmethoden wurden ausprobiert: maschinelle Übersetzung; wissensbasierte Techniken (z.B. Thesauri oder Wörterbücher); korpusbasierte Verfahren. Jede dieser Methoden brachte ermutigende Ergebnisse, hatte aber auch jeweils Nachteile aufzuweisen.

Maschinelle Übersetzung: Vollständige maschinelle Übersetzung (Machine Translation = MT) wird nicht als realistische Lösung angesehen für das Problem des Abgleichs von Dokumenten und Suchvorgängen in verschiedenen Sprachen. Ziel eines MT-Systems ist es, eine lesbare und verläßliche Version eines Textes in der Zielsprache zu erstellen, während sprachübergreifende Suche ausreichende Übereinstimmungen zu finden sucht zwischen der Anfrage in einer Ausgangssprache und dem Dokument in einer Zielsprache, so daß eine mehr oder weniger große Relevanz des Dokuments festgestellt werden kann hinsichtlich des Informationsbedürfnisses, das in der Anfrage formuliert wird. Das Übersetzen ganzer Sammlungen von Dokumenten in eine andere Sprache (die der Anfrage) ist daher nicht nur sehr teuer, sondern erfordert auch eine Reihe von Schritten, die vom rein fragetechnischen Standpunkt aus überflüssig sind, z.B. die Verschlüsselung linguistischer, semantischer und pragmatischer Informationen.

Beim Einsatz von MT-Systemen hat man sich eher darauf konzentriert, die Suchformulierungen als die Dokumente zu übersetzen. Allerdings bestehen Anfragen gewöhnlich aus Wortfolgen mit geringer oder fehlender syntaktischer Struktur. Die Eingabe kann daher durch ein MT-System nicht grammatikalisch zergliedert und traditionelle Methoden zur Vermeidung von Mehrdeutigkeit können nicht angewandt werden, weil es keinen semantisch zusammenhängenden Text gibt. Genaue Übersetzung ist daher nicht möglich, aber auch nicht nötig. Es besteht kein Bedarf für ein linear kohärentes und eindeutiges Ergebnis, tatsächlich können mehrere Übersetzungsmöglichkeiten zu einer Ausweitung der Anfrage führen, die das Ergebnis verbessern. Es konnte gezeigt werden, daß einfache und weniger kostenintensive Abfragetechniken zumindest genauso wirksam funktionieren und daß auf Wörterbüchern beruhende Verfahren kommerzielle MT-Systeme beim Übersetzen der Anfragen übertreffen [17].

Mehrsprachige Thesauri: Frühe Versuche haben gezeigt, daß mehrsprachige Thesauri brauchbare Ergebnisse bei sprachübergreifender Suche liefern können, und heute gibt es eine ganze Reihe kommerzieller thesaurusbasierter Systeme. Ein mehrsprachiger Thesaurus für Indexierung und Suche mit kontrolliertem Wortschatz kann als Sortiment einsprachiger Thesauri angesehen werden, die alle einem gemeinsamen System von Begriffen folgen. Bei einem kontrollierten Wortschatz gibt es eine festgelegte Anzahl von Begriffen, die bei der Indexierung und Suche angewandt werden. Dadurch wird das Problem der Mehrdeutigkeit ausgeschlossen. Die Nutzer können eine Bezeichnung in ihrer Muttersprache verwenden und damit den entsprechenden Begriff finden, um Dokumente in einer anderen Sprache zu überprüfen. In der einfachsten Form kann dies erreicht werden durch Nachschlagen in einem Thesaurus, der für jeden Begriff entsprechende Termini in verschiedenen Sprachen nachweist und ein Register für jede Sprache hat. Bei komplizierteren Systemen würde die Suche nach Übereinstimmung von Terminus und Deskriptor intern erfolgen.

Mit dem Zugang über einen kontrollierten Wortschatz müssen passende Wörter des Wortschatzes jedem Dokument der Sammlung zugeteilt werden. Üblicherweise wurde dies manuell von Fachleuten auf dem jeweiligen Gebiet erledigt. Dies ist teuer. Zur Zeit werden Methoden entwickelt für die (halb)automatische

Bestimmung dieser Indikatoren. Es bleibt das Problem, daß Thesauri teuer sind in der Entwicklung, kostspielig im Unterhalt und schwer auf aktuellem Stand zu halten. Zusätzlich hat es sich als recht schwierig erwiesen, Nutzer im wirksamen Gebrauch der Thesauri zu unterweisen.

Auf jeden Fall bewegt sich die Forschung weg von der Suche mit kontrolliertem Wortschatz in Richtung der Freitextsuche, obwohl sprachübergreifende Freitextsuche in vielerlei Hinsicht eine kompliziertere Aufgabe ist. Sie erfordert, daß jedem Wort der Anfrage eine Gruppe von Suchwörtern in der Dokumentsprache zugewiesen wird, wobei möglicherweise jedes Suchwort spezifisch gewichtet wird, um den Grad der Relevanz auszudrücken, den das Vorkommen eines Suchworts im Text für dessen Bedeutung in Bezug zum Suchwort hat. Die größere Schwierigkeit der sprachübergreifenden Freitextsuche rührt daher, daß man mit der gemeinten Bedeutung arbeitet, während bei der Suche mit kontrolliertem Wortschatz die Bedeutung in gewisser Weise vorgeschrieben sein kann. Andererseits ist das Fragepotential größer als bei kontrolliertem Wortschatz.

Wörterbuchgebrauch: Viele Systeme mit sprachübergreifender Freitextsuche verwenden zweisprachige maschinenlesbare Wörterbücher (Machine-Readable Dictionaries = MRDs) als Übertragungsquelle. Solche Hilfsmittel werden kommerziell und im Internet immer mehr angeboten. Da sie gewöhnlich für menschlichen Gebrauch entwickelt wurden, benötigen sie eine gewisse Aufbereitung bis zur Verwendung in einem automatischen System. Dies bedeutet vor allem die Analyse der speziellen Markierungsinformationen, die der Erkennung der unterschiedlichen lexikalischen Informationen dienen: Leitbegriffe, Wortarten, Sinnabschnitte, Übersetzungsäquivalente usw.

Es konnte gezeigt werden, daß die wörterbuchgestützte direkte Übersetzung einer Anfrage, bei der jedes Wort oder jede Redewendung der Frage ersetzt wird durch eine Liste aller möglichen Übersetzungen, ein passables erstes Ergebnis für sprachübergreifende Suche erbringt, obwohl solche - verhältnismäßig einfache - Methoden weniger leistungsstark sind als einsprachiges Abfragen. Automatische MRD-Übersetzung einer Suche führt zu einem Effektivitätsverlust von 40 - 60% gegenüber einsprachiger Suche [9, 10]. Hierfür gibt es vor allem drei Gründe: 1) die für allgemeine Zwecke gemachten Wörterbücher enthalten gewöhnlich keine Spezialbegriffe, 2) Fehler bei der Übersetzung von Bezeichnungen aus mehreren Wörtern und 3) das Problem der Mehrdeutigkeit.

Vielleicht das größte Problem bei Verwendung von MDRs ist die Mehrdeutigkeit. Bei einer Wort-für-Wort-Übersetzung mit einem Wörterbuch wird jedes Wort ersetzt durch alle möglichen Übersetzungsäquivalente. Wenn der Ausdruck der Anfrage selbst schon mehrdeutig ist, kann dies zu einer großen Zahl von Ausdrücken in der untersuchten Sprache führen, unter denen viele falsch sind und zur Auswahl von wertlosen Dokumenten führen. Bei der Übersetzung von Sätzen oder ganzen Dokumenten liefert der Zusammenhang Informationen, die zur Einschränkung der Mehrdeutigkeit führen können, die Kürze der durchschnittlichen Anfrage bedeutet in diesem Rahmen einen Mangel an Kontext. Diesem Problem wird von der gegenwärtigen Forschung besondere Aufmerksamkeit gewidmet

Es wurde nachgewiesen, daß sowohl syntaktische als auch statistische Verfahren die Wirkung der Mehrdeutigkeit deutlich verringern und das Ergebnis der sprachübergreifenden Suche dem der der einsprachigen Suche annähern können. Gut formulierte Anfragen können mit Identifizierungskennzeichen an den Wortarten zur Eliminierung grammatischer Homonyme gebracht werden und dadurch die Anzahl der vom Wörterbuch gelieferten unkorrekten Zielbegriffe verringern. Speziell solche Techniken der Erweiterung der Anfrage haben stark zur Verringerung von Mehrdeutigkeit beigetragen. Eigentlich führen solche Techniken neue Begriffe ein, ausgewählt anhand eines bestimmten Merkmals, um die Anfrage zu präzisieren [vgl. z.B. 11, 12].

Korpusbasierte Verfahren: Korpusbasierte Zugangsverfahren untersuchen große Textsammlungen nach statistischen Grundsätzen und ziehen automatisch jene Informationen heraus, die zur Konstruktion anwendungsspezifischer Übersetzungsmethoden benötigt werden. Die untersuchten Sammlungen können

aus parallelen (übersetzungsäquivalenten) oder vergleichbaren (bereichsspezifischen) Dokumenten bestehen. Die Verfahren, mit denen am häufigsten zur Untersuchung solcher Korpora experimentiert wurde, sind die des Vektorraums und der Wahrscheinlichkeitsrechnung.

Die ersten Versuche mit parallelen Korpora wurden mit statistischen Methoden unternommen zur Gewinnung von Daten mehrsprachig-äquivalenter Termini, die als Eingaben für die lexikalische Komponente von MT-Systemen benutzt werden konnten. Das Problem bei der Verwendung paralleler Texte als Übungsmaterial liegt darin, daß die Testsammlungen gewöhnlich bereichsspezifisch und teuer zu beschaffen sind - es ist schwierig, bereits vorhandene Übersetzungen der gesuchten Art von Dokumenten zu finden, und übersetzte Fassungen sind kostspielig in der Herstellung. Aus diesem Grunde gab es großes Interesse an den Möglichkeiten vergleichbarer Korpora.

Eine Sammlung vergleichbarer Dokumente enthält Texte, die zusammengefaßt sind aufgrund der Ähnlichkeit der behandelten Themen und nicht ihrer Übersetzungsäquivalenz. Voraussetzung ist, daß sie sich ähneln hinsichtlich der Gattung, der übereinstimmenden Paßform und der Zeitraums. Der Grundgedanke bei der Verwendung solcher Korpora besteht darin, daß die zur Beschreibung eines speziellen Themas verwendeten Wörter miteinander über die Sprachgrenzen hinweg semantisch verbunden sind.

Die bekannteste sprachübergreifende Strategie mit vergleichbaren Korpora ist der Zugang über einen mehrsprachigen Ähnlichkeitsthesaurus. [13] berichtet von Ergebnissen bei Verwendung eines Beispielkorpus, das geschaffen wurde durch die Zusammenfassung von Nachrichten der Schweizer Nachrichtenagentur (SDA) in Deutsch und Italienisch mit Hilfe eines thematischen Etiketts und des Datums, worauf sie zusammengeführt wurden zur Bildung eines `ÄhnlichkeitsthesaurusA. Deutsche Anfragen wurden dann ausprobiert anhand einer großen Sammlung italienischer Dokumente. Die Ergebnisse dieses Versuchs sind vielversprechend, vor allem bei Anwendung auf eine bereichsspezifische Sammlung.

Ein großer Nachteil korpusbasierter Verfahren besteht darin, daß sie sehr abhängig sind von dem jeweiligen Anwendungsbereich. Für neue Themenbereiche werden neue Referenzkorpora benötigt.

Zusammenfassung: Beim gegenwärtigen Entwicklungsstand können alle obengenannten Verfahren, wenn man sie in einem gut entwickelten, durchgeprüften und abgestimmten System verwendet, im allgemeinen Themenbereich etwa 80% der Wirkung einsprachiger Suche erzielen. Wie jedoch aus dieser kurzen Übersicht zu erkennen ist, hat jede einzelne Methode der sprachübergreifenden Suche ihre Grenzen. Welche Methode man auch wählt, die für die Anpassung der Frage an die Textsammlung verwendeten Mittel sind ein wichtiger Faktor für eine erfolgreiche Auswertung. Schon bestehende Methoden wie zweisprachige elektronische Wörterbücher reichen gewöhnlich nicht aus für diesen Zweck; die Entwicklung spezieller Hilfsmittel wie Thesauri und Übungskorpora ist teuer und solche Hilfsmittel meist nur beschränkt einsetzbar; ein neues mehrsprachiges Verfahren würde die Entwicklung neuer Quellen erfordern oder verhältnismäßig viel Arbeit zur Adaptierung der früher verwendeten.

Zu beachten ist auch, daß die meisten der heute verwendeten Systeme nur auf jeweils zwei Sprachen angesetzt werden können statt auf eine Vielzahl davon. Dies ist kaum überraschend. Die Situation ist viel verzwickter, wenn wir erfolgreiche Anfragen, gerichtet an eine größere Zahl von Sprachen, betreiben wollen und nicht nur über ein einziges Paar; dazu muß eine Art zwischensprachlicher Mechanismus untersucht werden - auf mehr oder weniger gedanklicher Ebene -, um den Transfer über mehrere Sprachen hinweg zu ermöglichen. In einer begrifflichen Mischsprache werden Begriffe und Redewendungen aus verschiedenen Sprachen, die demselben Bereich angehören, zu einem sprachunabhängigen System verbunden. So ist es möglich, gleichbedeutende Begriffe aller Sprachen zusammenzubringen und CLIR in jeder Sprachkombination zu erzielen, nicht nur bei paarweisen Zusammenstellungen. Doch die Entwicklung einer solchen Quelle ist keine leichte Aufgabe, und es bleibt noch viel zu tun, bevor wir von

wirklich vielsprachigen Suchverfahren sprechen können.

4. Bewertungen sprachübergreifender Systeme

Systembewertungen spielen eine bedeutende Rolle bei Anstößen zur Systementwicklung. Insbesondere gilt dies für sprachübergreifende Suchverfahren, die größtenteils noch in der Entwicklungsphase sind. Momentan gibt es verschiedene internationale Projekte auf diesem Gebiet.

- TREC - Text Retrieval Conferende Series (http://trec.nist.gov) bietet in diesem Jahr ein Übersetzungsverfahren vom Englischen und Französischen zum Arabischen.
- CLEF - Cross Language Evaluation Forum (http://www.clef-campaign.org/). CLEF, gefördert von der Europäischen Kommission als Teil des DELOS Network of Excellence for Digital Libraries, ist ein Bewertungsverfahren für europäische Sprachen und bildet die Fortsetzung des CLIR-Unternehmens, das 1997 von TREC begonnen wurde. CLEF 2001 hat vier Übersetzungsverfahren zur Bewertung von Suchmethoden im mehrsprachigen, zweisprachigen, bereichsspezifischen und einsprachigen (nicht-Englischen) Bereich. In diesem Jahr umfaßt die vielsprachige Dokumentensammlung vergleichbare Zeitungskorpora aus sechs Sprachen (Niederländisch, Englisch, Französisch, Deutsch, Italienisch, Spanisch) und Themenbereiche in 10 europäischen und 3 asiatischen Sprachen.
- NTCIR: Die NACSIS Test Collection for Information Retrieval (Testsammlung für Informationsgewinnung) (http://www.rd.nacsis.ac.jp/) wird betrieben vom National Institute for Informatics in Tokio. NTCIR enthält Übersetzungsverfahren für Chinesisch-Englisch und Japanisch-Englisch.

Diese Projekte bieten wichtige Foren für Systementwickler, die sich hier treffen, Ideen und Erfahrungen austauschen und Ergebnisse vergleichen können. Für Mitteilungen über die jüngsten Forschungsergebnisse auf MLIA-Gebiet sei der Leser verwiesen auf die neuesten Fortschrittsberichte dieser Initiativen [15, 16, 17].

5. Schlußfolgerungen

Wir haben einen sehr komprimierten Überblick über einige Themen gegeben, die zu berücksichtigen sind bei der Entwicklung eines Systems, das Zugang zu und funktionale Erschließung von Dokumentensammlungen in zahlreichen Sprachen bringt. In den letzten Jahren konnten auf diesem Gebiet viele Fortschritte gemacht werden und ist vieles angestoßen worden. Gegenwärtig konzentrieren sich die Bemühungen auf Probleme wie die Verbindung zahlreicher übersetzungsrelevanter Quellen zwecks Verbesserung der sprachübergreifenden Zusammenführung von Anfragen und Dokumenten, mehrsprachiger Zugang zu Sprachen mit `geringer DichteA - solche, für die linguistische Hilfsmittel nicht leicht in elektronischer Form erhältlich sind -, mehrsprachige Zugänge zu Multimediainhalten (vor allem gesprochene Dokumente und Darbietung der Ergebnisse aus vielsprachiger Suche) einschließlich inhaltlicher Zusammenfassungen zahlreicher Dokumente in verschiedenen Sprachen.

Trotz der Forschungsergebnisse von MLIA und CLIR mit ihren bemerkenswerten Fortschritten in den letzten Jahren muß festgehalten werden, daß die meisten in Alltagssituationen verwendeten Verfahren, die mit Dokumenten in zahlreichen Sprachen arbeiten, nur sehr einfache Zugangswerkzeuge bieten, die gewöhnlich nicht über einen kontrollierten Wortschatz auf einzelnen Gebieten hinausgehen und selten für mehr als nur zwei Sprachen. Eine neuere Umfrage bei einem Treffen der Teilnehmer an Projekten für Europäische Digitale Bibliotheken, gefördert von der Europäischen Kommission, ergab, daß zwar die meisten Projekte mit Dokumenten in mehreren Sprachen umgehen, aber nur sehr wenige bereits Verfahren anwandten, mit denen die Suche über mehr als nur eine Sprachsammlung gleichzeitig durchgeführt werden kann. Daraus wird klar, daß jetzt bedeutende Anstrengungen nötig sind, um die in der Welt der Wissenschaft erzielten Ergebnisse auf die Anwender zu übertragen. Wir hoffen, daß unser Kurs einen Schritt in diese Richtung bedeutet.

Literaturverzeichnis

1. Peters, C., Sheridan, P.: Multilingual Information Access. In M. Agosti, F. Crestani, G. Pasi (eds.). Lectures on Information Retrieval, Lecture Notes in Computer Science 1980, S. 51-80, Springer Verlag, 2001
2. Ziegler, D.: The Automatic Identification of Languages Using Linguistic Recognition Signals. PhD Thesis, State University of New York, Buffalo, 1991
3. Damashek, M.: Gauging Similarity with N-grams: Language-Independent Categorization of Text. Science, Vol. 267 (No. 10), 1995
4. Souter, C., Churcher, G., Hayes, J., Johnson, S.: Natural Language Identification Using Corpus-Based Models. Hermes Journal of Linguistics, Vol. 13, S. 183-203, Faculty of Modern Languages, Aarhus School of Business, Denmark, 1994
5. Wechsler, M., Sheridan, P., Schäuble, P.: Multi-Language Text Indexing for Internet Retrieval. In Proceedings of the 5th RIAO Conference, Computer-Assisted Information Searching on the Internet, Montreal, Canada, June 1997
6. Porter, M. F.: An Algorithm for Suffix Stripping. Program, Volume 14 (No. 3), S. 130-137, 1980
7. Ballesteros, L., Croft, W. B.: Resolving Ambiguity for Cross-Language Retrieval. In Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, S. 84-91, 1997
8. Soergel, D.: Multilingual Thesauri in Cross-Language Text and Speech Retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, S. 164-170, 1997
9. Hull, D. A., Grefenstette, G.: Querying Across Languages. A Dictionary-Based Approach to Multilingual Information Retrieval. In Proc. of 19th ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, S. 49-57, 1996
10. Ballesteros, L., Croft, W. B.: Dictionary-Based Methods for Cross-Lingual Information Retrieval. In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, S. 791-801, 1996
11. Ballesteros, L., Croft, W. B.: Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, CA, S. 1-8, 1997
12. Adriani, M., van Rijsbergen, C. J.: Term Similarity-Based Query Expansion for Cross-Language Information Retrieval. In Lecture Notes in Computer Science, 1696, 1999
13. Sheridan, P., Ballerini, J. P.: Experiments in Multilingual Information Retrieval Using the SPIDER System. In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, S. 58-65, 1996
14. Sheridan, P., Braschler, M., Schäuble, P.: Cross-Language Information Retrieval in a Multilingual Legal Domain. In ECDL =97 Proceedings, Pisa, Italy, S. 253-268, 1997
15. Voorhees, E. M., Harman, D. K. (eds.). The Eighth Text Retrieval Conference (TREC-8), US National Institute of Standards and Technology, 2000
16. Peters, C. (ed.). Cross-Language Information Retrieval and Evaluation. Proc. of the CLEF 2000 Workshop. Lecture Notes in Computer Science, 2069, Springer Verlag, 2001
17. Kando, N., Aihara, K., Eguchi, K., Kato, H.: Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, ISBN 4-924600-89-X, 2001