Seminar „Multimediale Informationssysteme“
AG Datenbanken und Informationssysteme
SS 2002 Universität Kaiserslautern

Multilinguales Information Retrieval

Seminarausarbeitung von
Erik Wagner
Matr.Nr.: 340553
E-Mail : e_wagner@informatik.uni-kl.de

Inhaltsverzeichnis :

1. Einleitung

Der Ausdruck Multilinguales Information Retrieval (MLIR) ist trotz einer eigentlich langen Geschichte keineswegs eindeutig. In [1] werden fünf Begriffsbestimmungen angeboten:

1. IR in einer anderen Sprache als Englisch
2. IR auf einer Sammlung von Dokumenten, wobei jedes Dokument in mehreren Sprachen vorliegt. Die Suche beschränkt sich auf die Versionen der Dokumente in der Anfragesprache.
3. IR auf einer einsprachigen Dokumentensammlung, die in mehreren Sprachen befragt werden kann.
4. IR auf einer Sammlung von Dokumenten in vielen Sprachen, die in vielen Sprachen befragt werden kann.
5. IR auf einer Sammlung von Dokumenten, die jeweils mehr als eine Sprache enthalten können.

In [2, 3] wird unter der Bezeichnung Cross-language Information Retrieval (CLIR) der Kernaspekt des MLIR eingegrenzt: Informationsgewinnung bei Überschreitung der Sprachgrenze. Daneben gibt es weitere Bezeichnungen, die sich auf diesen Themenkomplex beziehen.
Hier soll vor allem 4) interessieren. Der Benutzer soll sich eine Anfragesprache frei wählen und Dokumente in beliebigen Sprachen erhalten. Deshalb behalten wir die Bezeichnung MLIR bei. Die Sprachen, die dabei betrachtet werden, sollen vornehmlich auf einem Alphabet basieren, das sich nach der Lautlehre der jeweiligen Sprache orientiert. Dies sind zum Beispiel alle Sprachen, die mit lateinischen Buchstaben geschrieben werden. In [4] wird erwähnt, dass sich beim Retrieval mit Ideogramm (Bildzeichen) - basierten Sprachen besondere Schwierigkeiten auftun. Solche Sprachen, allem voran Chinesisch, erfordern zum Teil ganz andere Retrieval-Techniken und sollen hier nicht weiter betrachtet werden.
Den eigentlichen Retrieval Vorgang kann man folgendermaßen zusammenfassen (siehe Abb.1) : eine Anfrage Q, die aus einer Menge von Termen besteht, wird in eine Repräsentation QR umgewandelt. Analog dazu wird ein Dokument D in eine Repräsentation DR transformiert. Die Repräsentationen werden in einem Prozess C verglichen.

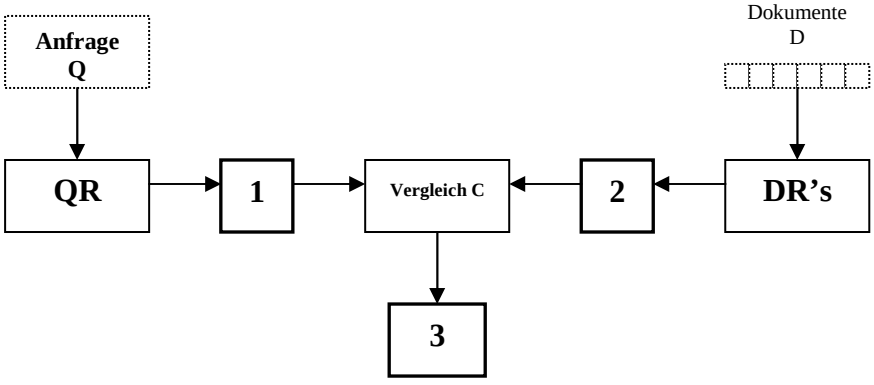


Abbildung 1: Ansatzpunkte für die Übersetzung

Das Problem ist, dass die Sprache der Anfrage nicht immer dieselbe ist, wie die Sprache der Dokumente, die gesucht werden. Wenn wir Repräsentationen von beiden schaffen, können sie nicht sinnvoll miteinander verglichen werden. Man muss Anfrage und Dokument auf dieselbe Vergleichsebene bringen, Quell- oder Zielsprache. Für das multilinguale Information Retrieval gibt es daher zwei *(drei)* wesentliche logische Ansatzpunkte, unter die man die bisher geleisteten Arbeiten einordnen kann (Abbildung 1). Zum einen

1. die Übersetzung der Anfrage in die Sprache der Dokumente oder
2. die Übersetzung der Dokumente in die Anfragesprache

(3. der Vergleich von Anfrage und Dokument auf einer sprachneutralen Ebene.)

Der erste Ansatz ist der Naheliegendste, da die Anfrage in der Regel merklich kleiner ist als ein einzelnes Dokument, geschweige denn eine Sammlung mit tausenden. Dabei reicht es, die Schlüsselwörter oder Phrasen, mit denen der Suchbedarf festgelegt wurde, in akzeptable Entsprechungen in der Zielsprache zu überführen. Eine solche Abbildung ohne Sinnverlust zu gewährleisten stellt das Hauptproblem dar, weil eine 1:1 Wortzuordnung nur selten möglich ist. So stehen sich oft vage Konzepte gegenüber: z. B. ‘image, picture, mapping, ...’ = ‘Bild, Abbildung, Darstellung, ...’. Die Übersetzung des Anfrageterms muss nicht nur dieselbe Bedeutung haben, sondern auch noch ein guter Suchterm sein.
Der zweite Ansatzpunkt ist bei weitem nicht einfacher. Hätte man eine effiziente Möglichkeit, Dokumente der Zielmenge zur Laufzeit zu übersetzen, so gut wie durch einen menschlichen Dolmetscher, würde das Problem auf das monolinguale IR reduziert. Aber die Maschinelle Übersetzung (MÜ, engl. MT : machine translation) stellt ein immer noch nicht befriedigend gelöstes Problem dar. Der Aufwand, den die MÜ erfordert, ist immens. Die Übersetzungen der Dokumente in alle Systemsprachen ist sehr aufwändig. Vergleicht man die Anzahl der Terme in der Anfrage mit denen in der Dokumentensammlung, ist dies offensichtlich eine unpraktikable Lösung. Man kann jedes Dokument auch in alle Sprachen des Systems übersetzen, wenn es in die Sammlung eingefügt wird. Dies erfordert aber auch einen hohen Aufwand. Wir werden trotzdem noch im letzten Kapitel eine kurze Übersicht über Ansätze der MÜ geben.
Ein möglicher dritter Ansatz wäre Dokument und Anfrage in einer sprachneutralen Art zu repräsentieren. Sie sind unabhängig von irgendeiner Sprache und können dann direkt abgeglichen werden. Dazu muss man aber erst Anfrage und Dokumente auf diese neutrale Ebene transferieren. Da dieser Ansatz in gewisser Weise die beiden anderen beinhaltet, werden wir hierauf nicht genauer eingehen.
Das Hauptproblem beim MLIR ist die möglichst exakte Disambiguierung der Wortbedeutung. Disambiguierung ist die Auflösung der Ambiguität, der Doppeldeutigkeit. Beim MLIR werden die Probleme des IR dupliziert. Erst muss nach einer guten Übersetzung gesucht werden, danach erfolgt die eigentliche Suche.

2. Sprachverarbeitung im MLIR

Verschiedene Techniken aus dem Bereich der Computerlinguistik können Anwendung finden, um den Suchprozess im IR zu unterstützen. Es handelt sich hierbei vor allem um Methoden und Werkzeuge für die Behandlung natürlicher Sprache (engl. Natural Language Processing oder NLP). Sie konzentrieren sich auf verschiedene Ebenen linguistischer Analyse. Ausgehend von den beiden o. g. Ansätzen lässt sich das weitere Vorgehen wie folgt unterteilen. Zunächst werden wir auf Möglichkeiten der Anfrageerweiterung mit Hilfe von Thesauri und Korpusbenutzung eingehen. Anschließend werden wir uns mit der Erkennung von Wortarten und –formen mittels morphologischer Analyse und Tagging beschäftigen,

danach einige Vorgehensweisen zur Erkennung einer Sprache kennenlernen und zum Abschluss drei Modelle maschineller Übersetzung behandeln.

2.1 Anfrageerweiterung

2.1.1 Thesauri

Die Schlüsselwörter, die der Benutzer in seiner Anfrage angibt, definieren die von ihm gewünschten Konzepte. Falls diese Terme nicht spezifisch genug sind, muss man damit rechnen mit einer unübersichtlich großen Menge an Dokumenten konfrontiert zu werden. Dann kann in Betracht gezogen werden, die Anfrage um einige Terme zu erweitern. Im boolschen Ansatz engen weitere Terme das gesuchte Konzept immer weiter ein und reduzieren die Ergebnismenge. So hat eine Anfrage ‘Technik AND Import’ wahrscheinlich mehr Ergebnisse als ‘Computer AND Software AND Technik AND Import’. Dies setzt voraus, dass in Ergebnissen alle Terme enthalten sein müssen. Im Gegenzug kann die Anfrage beim Vektorraumansatz auch um Synonyme erweitert werden. Da die Bedeutung eines Anfrageterms durch mehrere Begriffe angegeben wird, können hier auch mehr Ergebnisse gefunden werden. Dabei dürfen aber nicht alle Terme im Ergebnis gefordert sein, so wie beim boolschen Modell. Für beide Modelle ist Anfrageerweiterung sinnvoll. Um diese Erweiterung automatisch oder manuell zu unterstützen, kann man einen Thesaurus, ein Begriffswörterbuch, heranziehen (vgl. [6, 5]).
Ein Thesaurus ist eine Ontologie, eine Wissenssammlung, in der konzeptuell das Wesen eines einzelnen Begriffs beschrieben wird. Es ist also im Grunde eine meist strukturierte Konzeptliste, mit deren Hilfe der terminologische Hintergrund eines Suchterms über seine Beziehung zu anderen Konzepten erhellt werden soll. Statt wie in einem normalen Wörterbuch kann der Thesaurus deshalb auch aus Deskriptoren bestehen. Deskriptoren sind in diesem Fall Bezeichner, die jeweils für ein ganzes Konzept stehen und daher auch für eine Gruppe von Termen. Beispielsweise ließe sich durch ‘AVIATION’ jeder Begriff ausdrücken, der mit Luftfahrt zu tun hat, etwa ‘Flugzeug’ oder ‘Tower’. Mit Deskriptoren kann man also ganze Domänen und Kategorien bezeichnen. Wenn man Texte ausschließlich mit solchen Deskriptoren kategorisiert, spricht man dabei auch von Klassifikatoren. Man rechnet sie zum kontrollierten Vokabular. Der Thesaurus soll möglichst alle Konzepte enthalten, die der Benutzer bei seiner Suche nach Dokumenten vielleicht finden will. Dabei kann ein Dokument relevant sein, obwohl es oberflächlich gesehen nicht thematisch von dem gesuchten Konzept handelt. Erst durch die Beziehungen, die möglicherweise zwischen Suchtermen und Dokumententermen bestehen, kann in so einem Fall über Relevanz entschieden werden.
Nach [7] lassen sich diese Beziehungen in der Regel auf folgende drei verteilen:
- Äquivalenzrelationen
- Hierarchierelationen
- Nichthierarchische Relationen

Äquivalenz wird durch Angabe von Synonymen ausgedrückt. Beziehungen, die häufig in IR - tauglichen Thesauri auftreten können, sind (siehe [5]):

- BT (broader term) - OB (Oberbegriff), Hypernym
- NT (narrower term) - UB (Unterbegriff), Hyponym
- RT (related term) - VB (Verwandter) assoziierter Begriff
- WT (whole-term) - Ganzheit, Holonym (z. B. Rad - Speiche)
- PT (part-term) - Fragment, Meronym (s.o. oder Sand-korn)
- TT(top term) - Kopf der Hierarchie
- UF (used for) - BF (benutze für) Alternative, quasi-Synonym

- USE - BS (benutze Synonym) bevorzugt
- Antonyme, Gegensätze.

Oberbegriff und Unterbegriff stehen in hierarchischer Relation, während Ganzes und Teil nicht als Hierarchierelation betrachtet werden kann. Hierarchien erweitern Anfragekonzepte und verschaffen dem Benutzer eine größere Übersicht. Sie unterstützen auch die Suche nach allgemeineren Konzepten. In der terminologischen Struktur des Thesaurus werden Homonyme, Wörter mit gleicher Gestalt, aber verschiedener Bedeutung, dadurch eindeutig bestimmt, indem ihnen Referenzworte zugeordnet werden. So kann eng. 'to bow' (beugen) von 'bow of an archer' (Bogen eines Schützen) unterschieden werden. Es werden oft Beziehungen zwischen Synonymen aufgespannt. Die Disambiguierung heißt auch Polysemkontrolle (Trennung verschiedener Bedeutungen) [8].
Ein primitiver Thesaurus enthält nur solche Wortfelder wie 'Tempel, Kirche, Glaube, glauben, Religion, ...', bei welchem die Begriffe durch Assoziationen verknüpft sind. Man kann dann alle Wortfelder suchen, in denen ein Term vorkommt. Diese Wortfelder stellen Konzepte dar. Welche Konzepte in einem Thesaurus aufgeführt werden müssen, ist ein Hauptproblem bei der Schaffung des Thesaurus. Die Form eines Eintrags im Thesaurus richtet sich nach dem Zweck. Deshalb kann der Eintragsterm verschiedene äußere Formen annehmen [8]. Der Eintrag kann als Lexem auftreten, der grammatischen Grundform, oder als Wortstamm. Der Thesaurus mit Grundformen enthält den Infinitiv 'sagen', einer mit Wortstämmen dagegen 'sag'. Indem man jede grammatische Beugungsendung unterlässt kann man beim Nachschlagen durch einen rohen Abgleich die flektierten Wörter aus Dokumenten sicher in eine repräsentative Form bringen. Dagegen steht in einem Wortformenthesaurus jede gebeugte Form. Diese gebeugten Formen wären zwar in einer Suche als Zeichenkette leichter zu finden, aber das Volumen des Thesaurus würde so anschwellen, dass er nicht mehr praktisch verwaltet werden kann. Sowohl die Einträge als auch ihre Synonyme und Referenzen können Phrasen sein. Phrasen sind Sequenzen von Wörtern, die nicht ohne Sinnverlust aufgeteilt werden können. Besonders idiomatische Redewendungen kann man so durch ein Wort ersetzen, das den gleichen Sinn hat , etwa ‚sterben' für ‚ins Gras beißen'. Dem führenden Term eines Eintrags können zusätzliche Informationen zugefügt werden. Diese Informationen geben Eigenschaften des Wortes und Verwendungen wieder. Ihn um Informationen zu erweitern verbessert die Möglichkeiten, bei einer Suche im Thesaurus den richtigen Eintrag zu finden. So können Referenzwörter bei einem Eintrag stehen, die ein Wort disambiguieren. Beispielsweise kann man dadurch 'Rat (Personen)=Ausschuss' und 'Rat(Äußerung)= Hinweis' voneinander trennen. Die Einträge im Thesaurus können sonst noch Schreibvarianten, grammatische Varianten (Verb-Substantiv) und Ähnliches beinhalten. An zusätzlichen Informationen kann man die Wortklasse, Anwendungsfelder und Domänen aufführen. In einigen Fällen kann die Beifügung der Wortklasse zu einem Eintrag dessen Bedeutung eindeutig festlegen. Das Adjektiv 'verlegen (adj)=beschämt' ist dadurch gut von dem Verb 'verlegen(v)=positionieren' zu unterscheiden. Je nach Sprache kann dies aber auch überflüssig sein. In einem englischen Thesaurus ist so eine Möglichkeit durchaus sinnvoll, da Wörter ihre exakte Bedeutung erst aus der Stellung im Satz gewinnen. Im Deutschen werden Substantive schon durch die Großschreibung einer Wortart zugeordnet. Hier kann diese Markierung eher unterbleiben. Wie man Worten ihre Wortart zuordnet, werden wir später noch sehen.
Ein Thesaurus ist oft auf eine Domäne begrenzt. Durch die Begrenzung auf einen kleineren Kontext erreicht man eine größere Einengung der Konzeptbedeutung. Für die 'Fliege' kann man dann etwa Vokabular aus der BIOLOGIE-ENTOMOLOGIE heranziehen, für 'fliegen' Worte bzgl. PHYSIK-AERODYNAMIK. Besonders auf eine Domäne bezogen ist es schwer zu sagen, welche Einträge absolut notwendig sind. Falls der Thesaurus für allgemeinere Zwecke dienen soll, kann man die Einträge zur Unterscheidung der Wortbedeutung mit

semantischen Kategorien markieren. Durch Markierung mit den Domänenbezeichnern <FINANZ> und <NAUTIK> kann man dann ‘Bank<FINANZ>=Geldinstitut’ und ‘Bank<NAUTIK>=Untiefe’ auseinanderhalten [9]. Eine solche Möglichkeit entspricht in etwa auch der Suche in einem Thesaurus, der auf eine Domäne spezialisiert ist.
Multilinguales Information Retrieval erfordert ein Wörterbuch, das einen Anfrageterm auf seine Übersetzung abbildet. Für die Struktur eines solchen bilingualen Wörterbuches gilt mehr oder weniger dasselbe, wie für einen monolingualen Thesaurus. Hier muss aber nicht disambiguiert werden, um die Bedeutung eines Anfrageterms zu präzisieren oder um Synonyme zu bekommen, sondern um seine konzeptuelle Bedeutung zu konservieren. Dies soll zum Beispiel verhindern, dass die mathematische ‚Abbildung’ im Englischen durch ‚picture=Bild’ übersetzt wird, sondern sinngemäß etwa als ‚mapping’.
Die Grundprobleme bei der Benutzung eines Thesaurus sind Aktualität und Vollständigkeit. Ein Term muss immer richtig beschrieben sein, die Beziehungen zwischen Wörtern müssen stimmen. Thesauri für bestimmte Anwendungen müssen alle passenden Konzepte enthalten. Neue Ideologie, neues Vokabular, muss regelmäßig eingeführt werden.

2.1.2 Korpusbenutzung

Um die Hypothesen über die Eigenschaften von Wortformen und Wortbenutzungen zu erstellen, braucht man ein geeignetes Medium, wo man dies beobachten kann. Ein Korpus oder Textkörper [10] ist eine Sammlung von Dokumenten, die dazu dient, sprachliche Phänomene wie z. B. die Häufigkeit, mit der ein Wort benutzt wird, über statistische Analysen zu ermitteln. Dokumente in einem solchen Korpus sind über inhaltliche Aspekte verknüpft. Sie beziehen sich daher vielleicht auf ein bestimmtes Sachgebiet oder stammen vom selben Autor.
Ein großes Problem im IR ist Vokabular, das zu neu ist, um schon im Thesaurus gefunden zu werden [3]. Vor allem in Hinsicht auf die Entwicklung des Internet, dem weltweiten Netz zum Austausch von Information, tauchen ständig neue Begriffe auf, die in überhaupt keinem Thesaurus anzutreffen sind. Wörter aus einer Anfrage, die nicht darin aufzufinden sind, können auch nicht erweitert werden. Bei einem Übersetzungswörterbuch heißt das, der Term kann gar nicht übersetzt werden. Die Vollständigkeit eines Thesaurus erfordert häufige Wartung. Wir werden daher gleich auf die Möglichkeiten eingehen, thesaurus-artige Strukturen über statistische Auswertung eines Korpus zu erstellen. Ein anderes Anwendungsgebiet für einen Korpus ist die automatische Anfrageerweiterung.

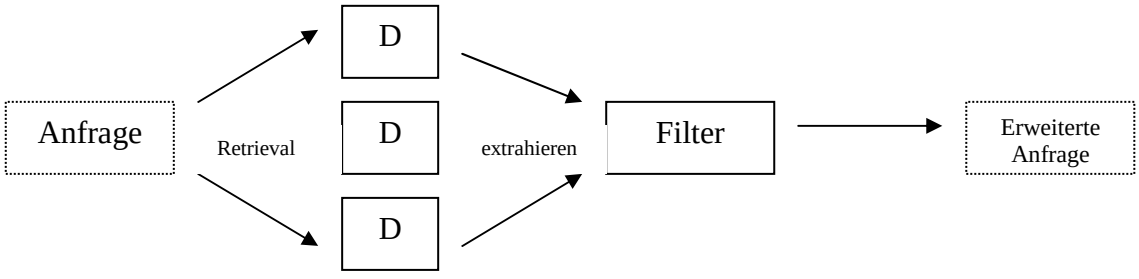


Abbildung 2: Anfrageerweiterung mit Korpus

Ein monolingualer Korpus kann für die Anfrageerweiterung benutzt werden, genauso wie ein Thesaurus. Der Unterschied besteht darin, dass keine beständigen Konzeptfelder verarbeitet

werden. Wortbeziehungen werden ad hoc durch die statistische Analyse ermittelt. Ein solcher Korpus soll in der Regel Referenzvokabular zu der Anfrage beisteuern, welches spezifisch für eine Domäne ist. Automatische Anfrageerweiterung gewinnt durch Retrieval eine Menge von Dokumenten, aus denen in einigen Schritten eine neue Anfrage gebildet wird.
Diese Schritte, die in Abb. 2 schematisch dargestellt sind, kann man so unterteilen :

- Suche: Die Anfrage findet im Korpus eine Menge von Dokumenten D (retrieval),indem die Ähnlichkeit von Dokument und Anfrage ermittelt wird, z. B. über Termvektoren.
- Einstufung: Die Menge der Dokumente (Ranking), die bei den Vergleich mit der Anfrage hoch eingestuft wurde, wird als relevant betrachtet (z. B. die besten 100).
- Extraktion: Die Terme werden daraus extrahiert. Dabei entfallen Stoppworte.
- Filter: Durch (meist) einfache Kriterien werden einige Terme ausgewählt. Zum Beispiel aus den besten Dokumenten die 100 Terme nach den besten (häufigsten) 300 Termen.

Das Resultat ist eine neue Anfrage. Im Gegensatz zum Relevanz Feedback(wird unten erläutert) läuft dieser Prozess ohne Benutzerbewertung ab. Dieses Verfahren liefert dabei in der Regel immer noch so viele Terme, dass ihre Anzahl reduziert werden muss. Sie sollen in die Anfrage eingefügt werden, ohne dass sie zu bedeutungsschwach wird. Deshalb benutzt man dafür Gewichtungen über das statistische Vorkommen. Diejenigen Terme, die in Dokumenten zu häufig oder zu selten auftreten, sollen eine schlechte Gewichtung bekommen. Sie beschreiben den Inhalt der gefundenen Dokumente nicht gut. Es sind zum einen wahrscheinlich Wörter, die in jedem beliebigen Text häufig auftreten. Sie haben nur eine schwache Bedeutung und erweichen das Anfragekonzept zu sehr. Ebenso kann man Terme ignorieren, die nur jeweils einmal in einem Text vorkommen. Sie sind kaum signifikant für den Inhalt des Textes und somit auch nicht von Bedeutung für die Anfrage. Solche und ähnliche Filtermechanismen kann man einsetzen, um aus einem Dokument weitere Anfrageterme zu extrahieren. Die Anfrage wird um diese neuen Terme erweitert und die richtige Literaturdatenbank wird damit befragt. Wenn die Dokumente für die Erweiterung von einem Benutzer aus einer Ergebnismenge gewählt werden, ist das Relevanz Feedback.
Auf ähnliche Weise kann man über einen Korpus eine Thesaurus-Struktur erzeugen. Dazu wird die Wortbenutzung untersucht. Wörter, die immer gemeinsam verwendet werden, also zumeist in denselben Texten auftreten, legen nahe, dass sie inhaltlich verwandt sind. Sie haben eine hohe Kookkurenz. Also weist man Termen, die oft in denselben Dokumenten auftreten, eine hohe Ähnlichkeit zu. Die ähnliche Benutzung drückt sich in ähnlichen Dokumentvektoren aus. Durch den Vergleich der Dokumentvektoren wird ein Ähnlichkeitswert zwischen den Termen ermittelt. Jedem Term aus einer Anfrage werden alle Worte zugeordnet, die eine große Ähnlichkeit dazu hatten. Ähnlich sind sie sich zum Beispiel dann, wenn die Verknüpfung ihrer Vektoren einen Schwellwert überschreitet. Durch diesen Wert kann man den Grad der Verwandtschaft definieren. Listet man die Ergebnisse einer solchen Analyse auf, also zu jedem Wort bis zu einer unbestimmten Zahl seiner Verwandten, ergibt sich ein einfacher monolingualer Thesaurus. Man nennt es auch Ähnlichkeits-Thesaurus (similarity thesaurus) [11], da es über das Ähnlichkeitsmaß erstellt wird.
Auf einer Menge von Dokumenten in mehreren Sprachen kann man so die Übersetzungsbeziehung erstellen [12]. Gegeben seien zwei Texte in den beiden untersuchten Sprachen. Die Dokumente seien sich von ihrem Inhalt ziemlich gleich. Wenn zwei Wörter A und B oft nahe beieinander stehen, dann sollten die Übersetzungen der Wörter ebenfalls oft zusammenstehen. Kann man diese Beobachtung für den Term A und seine potenzielle Übersetzung A’ oft machen, lassen sie sich zuordnen. Vergleicht man so erzeugte

Beziehungen über einer größeren Textmenge, erhofft man sich, zuverlässige Wortzuordnungen zu bekommen.
Der kritische Punkt bei allen Korpusmethoden ist der Korpus selbst. Je weiter das spezifische Wissen in der Textsammlung sein soll, desto umfangreicher wird sie. Es ist daher sinnvoll, für jede eingegrenzte Domäne eine eher kleine, erlesene Menge von Dokumenten zu vereinigen. Diese Dokumente müssen aber dann auch repräsentativ sein. Es ist nicht vertretbar, dass etwa in einer Sammlung über IR nur MLIR Texte vertreten sind. Entsprechend muss man für jedes Wissensgebiet, über das die Suche laufen kann, einen solchen Korpus haben. Andernfalls tun sich Lücken auf. Der Korpus muss zumindest das Kernvokabular der Domäne beinhalten. Für die Anwendung linguistischer Methoden muss auch gelten, dass ein Korpus von einwandfreier Qualität ist. Wie schon erwähnt sollten Texte in einem Korpus durch ihren Inhalt verknüpft sein [10]. Für das multilinguale Retrieval sucht man daher Texte zusammen, die den Kontext von möglichen Anfragen in der Zielsprache erhellen können. Ideal ist es, wenn man darin zu einem Text in einer Sprache eine adäquate Übersetzungen in jeder anderen Sprache, die das System unterstützen soll, parallel zuordnen kann. In diesem Kontext spricht man von einem parallelen Korpus. Die Nützlichkeit eines parallelen Korpus ist an gewisse Bedingungen geknüpft. Nahezu ideal wäre es, einander Sätze zuzuordnen. Nächstbesser ist danach die Verbindung von Dokumentteilen und zuletzt ganze Texte. Ein paralleler Korpus muss aus Texten bestehen, in denen das Vokabular nach herkömmlichen Maßstäben verwendet wird. Vor allem wenn neueste wissenschaftliche Entwicklungen gesucht werden, kann man dies kaum garantieren. Wenn beispielsweise ein Autor in einer ihm nicht völlig vertrauten Sprache schreibt, kann ein solcher Text nur mit großen Vorbehalten einbezogen werden.
Die Verwendung eines (parallelen) multilingualen Korpus, in dem jedem Text in einer Sprache entsprechende Texte in anderen Sprachen zugewiesen werden, kann die Mittel für die Übersetzung liefern. Wie schon weiter oben erwähnt, kann das Maß an Ähnlichkeit (in der Benutzung) eines Wortes der Quellsprache mit einem in der Zielsprache als rohe Übersetzung dienen. Eine konsequente Auswertung eines bilingualen Korpus kann die manuelle Erstellung oder Aktualisierung eines Wörterbuches überflüssig machen. Zu jedem Auftreten eines Wortes in quellsprachlichen Texten werden in den dazu parallelen Texten der Zielsprache die Ausdrücke gesucht, die oft parallel benutzt werden. Für ein Wort in der quellsprachlichen Version eines Satzes(Abschnittes) steht meist eine direkte Übersetzung auf der Zielseite, umlagert von anderen Wörtern. Deshalb kommen zuerst alle Worte des Satzes (Abschnittes) der Zielsprache als Übersetzungen in Frage. Ein Wort der Quellsprache und ein Wort der Zielsprache sind sich um so ähnlicher, je öfter sie gemeinsam in parallelen Texten vorkommen. Hat man zum Beispiel zwei Sätze in Deutsch und Englisch:

There’s a cat on the tree. = Auf dem Baum sitzt eine Katze.
The cat has red spots. = Die Katze hat rote Flecken.

kann man schnell sehen, dass ‘cat’ und ‘Katze’ Übersetzungen sind. Diese Beziehung wird durch die Sätze zweimal gestützt. Eine Beziehung ‘tree’= ‘Baum’ kommt nur einmal vor und ist somit weniger wahrscheinlich. Alle diese potenziellen Übersetzungen können nach der Ähnlichkeit mit dem zu übersetzenden Wort in eine Rangfolge gebracht werden. Dabei steht die wahrscheinlich beste Übersetzung an erster Stelle. Diejenigen Ausdrücke, die am häufigsten als Entsprechung zum Wort in der Quellsprache benutzt werden, müssen folglich mit der größten Wahrscheinlichkeit gültige Übersetzungen sein.
Ein Wort von häufig damit verknüpften Referenzwörtern zu trennen erhöht seine Bedeutungsvielfalt. Kommt ein Wort wie ‘verlegen’ in einem Text vor, braucht man noch andere Referenzwörter, um die Bedeutung ‘peinlich’ zu erschließen. Word Sense Disambiguation (Disambiguierung der Wortbedeutung, vgl. [13]) versucht zu ermitteln,

welches der Wörter die Übersetzung ist und welche Wörter als Referenzen dienen. Anhand der Referenzwörter soll die Übersetzung mit dem korrekten semantischen Sinn gefunden werden können. Zu diesem Zweck kann man zu jedem statistisch erstellten Eintrag eines Wörterbuches verschiedene Felder mit Referenzwörtern hinzufügen. So wird die Basis zur Disambiguierung geschaffen. Da solche ‘adäquaten’ Übersetzungen für einen parallelen Korpus nur selten erhältlich sind, kann man auf einen Korpus vergleichbarer Dokumente (wie in [14]) zurückgreifen. Hier beschränkt sich die inhaltliche Gemeinsamkeit auf ein Thema. Beispielsweise kann man Presseartikel, die sich auf dasselbe Ereignis beziehen, durch Übereinstimmung in Angaben bzgl. des Inhalts zuordnen. Sind zwei Texte aus der Rubrik ‚Wirtschaft’ vom 30.1.97 und der Ort des Ereignisses ist derselbe, kann man die Texte trotz unterschiedlicher Sprache einander zuordnen. Bezeichner für solche einfachen Kategorien sind den Texten manuell beigefügt worden und beschreiben z. B. Ort, Zeit und Rubrik einer Nachricht. In [14] werden Artikel in Italienisch und Deutsch über ihr Erscheinungsdatum und über Deskriptoren, wie sie in Tabelle 1 angegeben sind, einander zugeordnet. Dabei wurden Artikel der Schweizer Presse verwendet, die manuell kategorisiert wurden.
Ein Bezeichner ‘19970130c4gerfin’ steht für ein Ereignis vom 30.1.97 in Deutschland, das mit Finanzen zu tun hatte. Naturgemäß ist eine solche Zusammenfassung unzuverlässiger als ein paralleler Korpus.

fin	finance	zh	Zürich
kul	cultur	be	Bern
umw	environment	ge	Geneva
c1	United States	c4ger	Germany
c4	European Union	c4ire	Ireland
c7	Africa	c4ita	Italy

Tabelle 1: Deskriptoren für vergleichbaren Korpus

Negativ hervorgehoben werden sollte auch der Aspekt, dass der Benutzer an sich keinen Einfluss darauf hat, welche Texte genau zu der Korpusanalyse herangezogen werden. Die Themen oder Konzepte, die in der Sammlung vorkommen, werden durch die Summe aller Dokumente bestimmt. Diese Themen werden durch die Personen bestimmt, welche die Sammlung zusammenstellen. Das Gesamtkonzept der Sammlung drückt das aus, was der Ersteller sich darunter vorstellt. Wenn Texte ohne inhaltliche Bezüge zusammengestellt werden, passt eine solche Sammlung zu keinem Thema und keiner Domäne. Ein Benutzer muss genau über die Sammlung informiert sein, um effektive Anfragen stellen zu können. Zuletzt muss man das Hauptproblem für parallele Korpora betrachten, die mangelnde Verfügbarkeit von Übersetzungen. Eine nach menschlichen Maßstäben gute Übersetzung ist selten verfügbar, vor allem in technischen Ressorts. Besonders eine feinstrukturierte Zuordnung von Satz zu Satz dürfte am ehesten Mangelware sein, da manches in einer Sprache mit einem Satz, in einer anderen aber gewohnheitsmäßig in zwei ausgedrückt wird. Ein ungeeigneter Korpus wird die Anfrage nur unnötig verzerren.

2.2 Erkennung von Wortarten und –formen

2.2.1 morphologische Analyse

Wie schon oben erwähnt ist es nicht sinnvoll, jede gebeugte Wortform aus einem Dokument in dessen Repräsentation aufzunehmen. In der Regel gibt der Benutzer die Schlüsselwörter in

seiner Anfrage in ihrer Grundform an. In Texten treten die Begriffe aber eher in einer gebeugten Form auf. Deshalb bietet der eigentliche Wortstamm eine bessere Basis für eine Suche. Ein Lexem ist ein Wort, wie es im Wörterbuch zu finden ist, in seiner Grundform. Ist ein Lexem nicht aus mehreren Wörtern zusammengesetzt, wie z. B. das Kompositum ‘Haus-tür’, besitzt es einen Stamm, der die Grundbedeutung ausdrückt. Der Stamm von ‘Ver-walt-er’ ist also ‘walt’. Die Endung ‘er’ hat aber auch eine Bedeutung, sie weist auf eine Person hin. Die kleinsten bedeutungstragenden Einheiten, die in einer Sprache vorkommen sind die Morpheme, daher sind ‘ver’, ‘walt’ und ‘er’ Morpheme. Ein Wort im Wörterbuch zu finden ist oft nur eine Aneinanderreihung solcher Morpheme, z. B. ‚Ent-scheid-ung’. Bei der expliziten morphologischen Analyse wird eine Wortform in ihre Morpheme zerlegt. Ein Morphem wie ‘-ung’, dass nur zusammen mit anderen Morphemen eine Wortform bilden kann, nennt man abhängiges Morphem oder Affix.
Die Beugung erfolgt im Deutschen meist durch das Antreten von Einheiten an den Stamm. Dient ein Morphem nur dazu, ein Wort syntaktisch zu beugen, wie z. B. das Plural -n in ‘Eulen’ oder -t- in ‘sagte’, heißt es Flexionsmorphem. Wenn aus einem Wort durch das Antreten eines Affixes wie ‘-ung’ ein neues Wort entsteht, so nennt man das Affix Derivations- oder Ableitungsmorphem. Einige Derivationsmorpheme für Deutsch, Englisch und Spanisch sind in Tabelle 2 angegeben. Derivationsmorpheme sind in der Regel näher am Stamm als Flexionsmorpheme.

Deutsch	Englisch	Spanisch
-lich	-ly	-mente
-nis	-ness	-miento
-bar	-ble	-ble
-schaft	-ship	-dad
-ismus	-ism	-ismo
-er	-er	-ero/a
-ie	-y	-ia

Tabelle 2: Ableitungssuffixe

Entfernt man daher die Flexionsmorpheme, erhält man eine Form, die noch die Bedeutung des Wortes wiederspiegelt. Derivationssuffixe zu entfernen vereinfacht meistens die konzeptuelle Bedeutung der verbleibenden Form. So bezeichnet ‘sicht-bar’ die Möglichkeit der ‘Sicht’. Der Stamm ist meist aber auch vager. Für die Suche in einer Wortliste ist es günstig, viele der syntaktisch oder durch Derivation abgeleiteten Wortformen auf ihren Stamm oder eine Wortform mit Stammqualitäten zurückzuführen. Um zum Beispiel ein Verb zu suchen, muss man den Infinitiv bilden, der als Grundform betrachtet wird. Morpheme, die für den Suchprozess als weniger wertvoll angesehen werden wie Flexionsendungen, sollen daher entfernt werden. Eine vereinfachte Methode, um eine Wortform auf eine Grundform zu bringen, ist das Stemming.
Stemming ist eine Methode zur morphologischen Zerlegung einer Wortform. Beim Stemming findet eine implizite Erweiterung statt, wenn Derivationsaffixe entfernt werden. Das Konzept des Wortes wird weiter. Da sich die Forschung im Computer-Bereich hauptsächlich auf Englisch bezieht, in dem die Bildung von Wortformen durch Präfixe sich in Grenzen hält, ist das Abtrennen von Vorsilben nicht üblich. Also behandelt Stemming meist die Suffixentfernung. Durch die Anwendung von Regeln, die ein Suffix durch eine anderes ersetzen, soll in einem oder mehreren Schritten die gesuchte Form erstellt werden. Man würde ‘-tümer’ durch ‘tum’ ersetzen, da man weiß, Wörter auf ‘-tum’ wie ‘Reichtum’ haben häufig

noch die Endung ‘er’. So kann man bei ‘Kapazitäten’ die Endung schrittweise ‘ität’ und ‘en’, oder in einem Schritt ‘itäten’ entfernen. Negativ wirkt es sich aus, wenn durch die Entfernung eines Suffixes zwei Wörter zusammengefasst werden, die nur ursprünglich dieselbe Wurzel hatten, sich aber jetzt in ihrer Bedeutung deutlich unterscheiden. Wenn doing = ‘tuend’ mit doe = ‘Hirschkuh’ oder ‘Schicksal’ mit ‘schick’ zusammengeworfen wird, kann man schlechtere Ergebnisse erwarten. Für das Stemming empfiehlt es sich, die Verschmelzung nur in Übereinstimmung mit dem Nachschlagen in einem Wörterbuch der Anfragesprache durchzuführen [14]. Dadurch will man sichergehen, dass das durch den Stemmer erzeugte Stammwort tatsächlich existiert. Stemming wird überall inzwischen standardgemäß für die Bildung von Dokumentrepräsentationen angewendet.

2.2.2 Tagging

Die Entfernung eines Wortes aus seinem Kontext und das Abschneiden von Beugungsformen bei der Sprachverarbeitung zerstören gerade die inhaltlichen Beziehungen, welche die Begriffe verbinden. Losgelöst aus ihrer Umgebung gehen wichtige semantische Informationen über einzelne Terme verloren. Solche Beziehungen soll zum Beispiel Tagging erhellen. Beim Tagging soll die Wortart, die Rolle oder auch Part-of-Speech (POS), eines Terms innerhalb eines Satzes mit der entsprechenden Etikette (engl. tag) markiert werden. Die Anwendung von Tagging in einer Anfrage ist gebunden an die Benutzung natürlicher Sprache. Die Worte in vollständigen Texten können getaggt werden, losgelöste Schlüsselwörter in einer Anfrage dagegen kaum. Ein Verfahren für Tagging geht durch den Text und verteilt die Wortklassen-Marken.

How[WRB] has[VBZ] the[DT] threat[NN] of[IN] swine[NN]
fever[NN] affected[VBD] international[JJ] trade[NN]?

Abbildung 3: Satz mit POS-Tags [8]

In Abbildung 3, entnommen aus [15], steht ein englischer Fragesatz, wobei hinter jedem Wort ein Wortklassen-Tag angegeben ist. NN steht für ‘NOUN’ (Substantiv). Andere Wortarten wären etwa JJ für Adjektive, VB und VBZ für Verben. Wort und Wortart ergeben ein Token (z. B. ‘trade[NN]’), das dann in einem Thesaurus oder Wörterbuch gesucht werden kann. Das Wort wird nur gefunden, wenn die Wortklasse stimmt. Diese Unterscheidung muss natürlich von der Struktur des Thesaurus unterstützt werden. In Tabelle 4 stehen die Einträge zu den Worten aus Abbildung 3. Die Stopworte wurden daraus entfernt und die Terme sind mit einem Stemmer trunkiert.

Eintrag	**Übersetzung**
threat[NN]	achor| amag| bravat| conmin| disfuerz| espant| nublar| peligr| ret| ronc
swine[NN]	canall| cochin| galduf| jet| malaj| mam| marran| papalc| perr| puerc| sinvergonz| vergaj| villan
fever[NN]	calentur| chuch| fiebr| pasm
intern[JJ]	intern

Tabelle 4: 1:n Wörterbuch Englisch-Spanisch mit POS (Auszug)

Einige Wörter sind Homographe, d.h. zwei Wörter werden gleich geschrieben, obwohl sie verschiedene Wortklassen haben. Durch die Markierung eines Wortes mit seinem Wortart-Tag verringert sich die Verwechslungsgefahr bei der Suche im Thesaurus. Tagger arbeiten im Wesentlichen auf Sätzen von Regeln oder mit stochastischen Analysen.
Regelbasierte Tagger benutzen Gesetzmäßigkeiten und den kontextuellen Rahmen eines Wortes, um ihm sein Wortart-Tag zuzuteilen. Eine solche Regel könnte zum Beispiel lauten: Wenn vor X ein Artikel steht und nach X ein Substantiv, dann ist X ein Adjektiv. Oder, falls X die Endung '-heit' hat, ist es ein Substantiv. Diese Regeln werden angewendet, bis allen Wörtern konsistent eine Wortart zugewiesen wurde.
Beim stochastischen Tagging wird die Wahrscheinlichkeit, dass eine Wortform einer bestimmten Wortart angehört, und die Wahrscheinlichkeit, mit der bestimmte Wortarten aufeinander folgen, benutzt, um den Termen Tags zu verleihen. Diese Wahrscheinlichkeitsmaße können über Training auf einer Menge von Dokumenten gewonnen werden. Die Hypothese, die jedem Wort eines Satzes, also der Wortsequenz, eine Wortart zuordnet und am wahrscheinlichsten ist, wird dabei gewählt. Tagging nimmt insofern Einfluss auf das eigentliche Retrieval, weil es die Behandlung von Phrasen unterstützt. Bestimmte Folgen von Wortarten können vereinfacht als Phrasen gesehen werden und können dann beim Retrieval auch als solche behandelt werden, etwa zusammengefasst. Dabei werden zum Beispiel Ketten von Substantiven automatisch als potentielle Komposita aufgefasst. So könnte man Kombinationen von Dokumententermen automatisch verbinden und in den Index übernehmen. Analog kann man bei Anfragen in natürlicher Sprache verfahren.

2.3 Spracherkennung

Im Rahmen einer Anwendung im Internet muss man vielleicht vor allen anderen Arbeitsschritten sicherstellen, dass diese Maßnahmen nicht ins Leere laufen, weil die Kodierung, also deren Alphabet, nicht bekannt ist. Besonders linguistische Methoden können ineffektiv werden, wenn die Bedeutungseinheiten nicht erkannt werden. Im Bezug auf Stemmer wird dies wohl klar. Das kann auch so gesehen werden, dass man auf Texten, deren Sprache bekannt ist, explizit Wissen über die jeweilige Sprache anwenden kann. Sollte man beispielsweise wissen, es handelt sich um ein Dokument in Spanisch, hat man die Option, diesen Text an ein maschinelles Übersetzungssystem weiterzuleiten. Ein solches System ist spezialisiert und kann nur auf einem Paar bekannter Sprachen operieren.
Falls selbst die Kodierung des Dokuments nicht bekannt ist, muss die Erkennung derselben zuerst erfolgen. Die Kodierung, d.h. das Zeichensystem, gibt schon einige Anhaltspunkte bezüglich der Dokumentensprache. Während der Standard ISO-LATIN-1 (ISO-8859-1) jede westeuropäische Sprache zur Auswahl lässt, bedeutet JIS, es kann sich nur um Japanisch handeln. Für die Erkennung einer Sprachkodierung gibt es verschiedene Ansätze. Sie beruhen im Wesentlichen darauf, gewisse Charakteristika der jeweiligen Kodierung zu erkennen, also typische Sequenzen. Wenn das Kodesystem erst bekannt ist, kann die Sprache relativ schnell ermittelt werden. Zum einem kann man n-Gramm Statistiken heranziehen. Ein n-Gramm ist eine beliebige Teilzeichenkette der Länge n aus einem Wort. Je größer n, desto mehr Kombinationen von Zeichen gibt es. Lange Kombinationen sind aber eindeutiger einer Sprache zuzuordnen. Die Silbenstruktur einer Sprache ist meist so einmalig, dass man schon mit Trigrammen (3-Gramm) gute Ergebnisse erzielt. Die Zeichenkette 'sch' repräsentiert im Deutschen einen eigenen Laut, sie ist für Deutsch typisch. In spanischen Wörtern wird man 'sch' dagegen nie sehen. Wenn man die Sprache eines Textes sucht, sind solche Unterscheidungen daher sehr geeignet, den Kreis der zu berücksichtigenden Sprachen auszudünnen.

Eine andere Methode besteht darin, kleine Wörter zu erkennen. Diese kleinen Wörter sind so häufig in einem Text zu finden, dass sie ziemlich eindeutig auf eine bestimmte Sprache verweisen. Diese Wörter sind zum Beispiel Präpositionen und Artikel. In einem deutschen Text etwa sind die häufigen ‘der, die ‚das’ sehr auffällig. Diese kleinen Wörter sind also oft in Stoppwort-Listen. Für ein Dokument kann man daher zählen, wie oft ein Element aus einer länderspezifischen Stoppwort-Liste gefunden wurde. Die Sprache jener Liste, deren Elemente am häufigsten auftraten, wird angenommen. Ein Dokument wird dann mit dem passenden Sprachbezeichner markiert.
Spracherkennung erfordert einen hohen Aufwand, da man immer zwischen allen Sprachenpaaren unterscheiden muss.

2.4 Maschinelle Übersetzung

Maschinelle Übersetzung wurde als Werkzeug nicht für Information Retrieval konzipiert. Stattdessen wurden bestehende Übersetzungssysteme um Retrieval-Möglichkeiten erweitert. Prinzipiell erfordert MÜ einen sehr großen Aufwand und zeigt gerade an wichtigen Stellen Schwächen. Man unterscheidet drei Grundarten von MÜ-Systemen [16]: direkte, Interlingua- und Transfer Systeme.

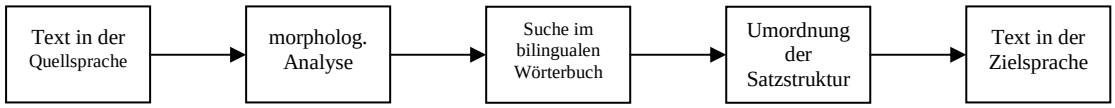


Abbildung 4: Modell eines direkten MÜ-Systems

Ein direktes MÜ-System (Abb. 4) wendet nur wenig Aufwand bei der Übersetzung an. Von einem Wort in einem Satz wird durch morphologische Analyse die Grundform ermittelt. Diese wird dann in einem bilingualen Wörterbuch nachgeschlagen, in dem nur eindeutige Wort-zu-Wort-Zuordnungen stehen. Anschließend wird die Satzstruktur der Quellsprache über einige Regeln grob in die der Zielsprache umgewandelt. Zum Beispiel wird ein Satz mit der Struktur Subjekt-Prädikat-Objekt in die Reihenfolge Subjekt-Objekt-Prädikat überführt. Da im Deutschen mehrere Satzstellungen möglich sind, die eine verschiedene Bedeutung haben, ist dieses Verfahren nur grob. Die semantische Bedeutung des ursprünglichen Satzes und syntaktische Beziehungen werden nicht herangezogen.
Auf Grund der Grobheit dieses Verfahrens ist der Übersetzungsvorgang in der Regel irreversibel. Beim Interlingua-Ansatz wird der Text in der Quellsprache zuerst in eine neutrale Repräsentation gebracht. Diese Repräsentation soll im gleichen Maße den Quelltext und die Bedeutung seiner Übersetzung darstellen. Sie enthält alle Informationen, die notwendig sind, um den Text in der Zielsprache zu generieren, sodass man nicht auf den Quelltext zurückschauen muss. Durch das zwischengeschaltete Medium Interlingua sollen Probleme, die wegen der strukturellen Unterschiede der zwei Sprachen auftreten, umgangen werden. Deswegen ist der Interlingua-Ansatz von besonderem Interesse für multilinguale Systeme (vgl. Abb. 5).
Ein einfaches Beispiel für eine Interlingua sind Termvektoren, die über kontrolliertem Vokabular erstellt werden. Man drückt den Inhalt aller Dokumente durch eine feste Menge vorgegebener Konzepte aus, auch wenn die Dokumente in verschiedenen Sprachen sind.

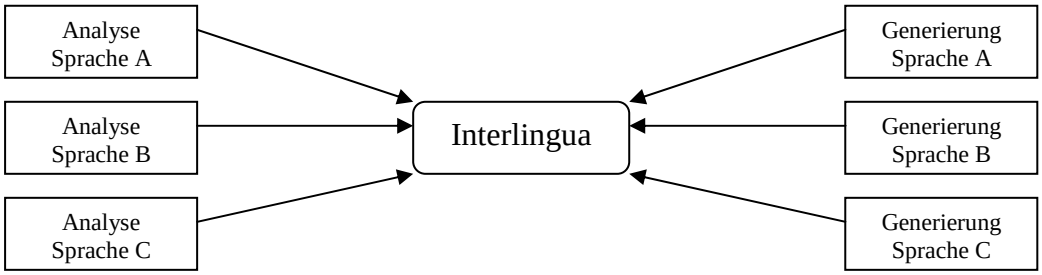


Abbildung 5: Modell eines Interlingua-Systems mit mehreren Sprachen

In der Theorie gehen die Möglichkeiten einer Interlingua viel weiter. Man kann von einer Sprache in eine beliebige andere übersetzen, wenn es ein Analysemodul für die Quellsprache und ein Generierungsmodul für die Zielsprache gibt. Bei einer Interlingua soll die Welt in eindeutiger Weise beschrieben sein. Jeder Begriff in einem Interlingua-Lexikon soll durch seine Form von jedem anderen zu unterscheiden sein. Im Lexikon steht nur eine Anzahl primitiver Konzepte, gerade groß genug, sodass man mit Verknüpfungen dieser Konzepte alles ausdrücken kann. Verknüpfungen sind z. B. ‘X macht Y’ oder ‘X ist Z’. Ein Wort wird in die Interlingua übersetzt, indem es als Kombination von einfacheren Konzepten ausgedrückt wird. So wird aus ‘Seher’ (Person,sehen). Wenn ein Wort nicht mehr vereinfacht werden kann, kann es ersetzt werden durch ein primitives Konzept der Interlingua. Eine Satz wird so von der Quellsprache übersetzt in eine Interlingua-Formel. In der Interlingua Formel werden alle syntaktischen und semantischen Wortbeziehungen ausgedrückt. Aus dieser Formel wird dann nach und nach eine Übersetzung erzeugt. Dieser Aufbau macht Wartung und Erweiterung eines Interlingua-Systems einfach. Das Hauptproblem dieses Ansatzes [16] ist aber, eine solche repräsentative Zwischensprache bereitzustellen. Dies ist ein Problem, welches noch nicht befriedigend gelöst wurde.
Die explizite Übersetzung zwischen genau zwei Sprachen leistet der letzte Ansatz, das Transfer-Modell. Im Gegensatz zum Interlingua-Ansatz werden hier Beziehungen, also Übereinstimmungen, konsequent ausgenutzt, um Texte zu übersetzen. Zwischen der Analyse in der Quellsprache und der Generierung in der Zielsprache ist eine Einheit geschaltet, die genau diese eine Abbildung leistet. Eine multilinguale Anwendung wird im Vergleich zum Interlingua-Modell entsprechend komplexer.

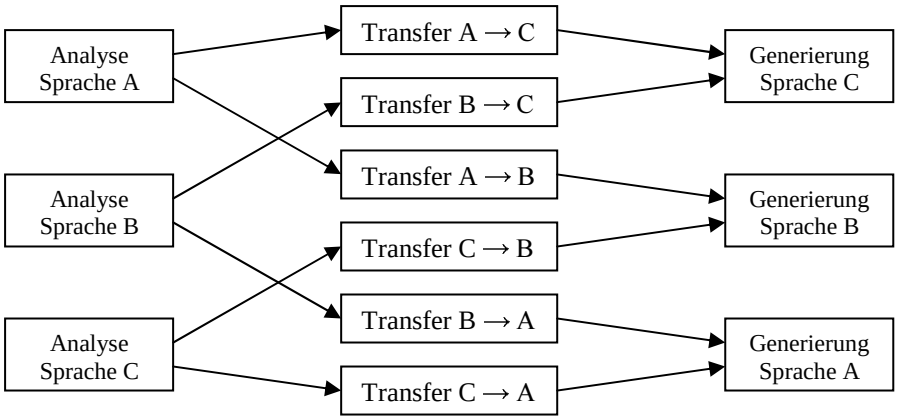
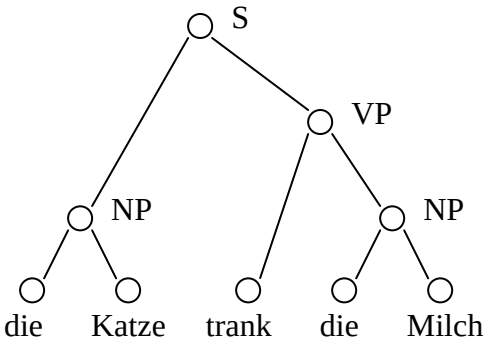


Abbildung 6: Modell eines Transfer-Systems mit mehreren Sprachen

Wie aus Abb. 6 ersichtlich wird, steigt der Aufwand quadratisch. Nichtsdestotrotz ist dies der häufigste Ansatz, vor allem wegen dem Fehlen einer angemessenen Interlingua.
Für die syntaktische Analyse und Synthese von Sprachen werden Grammatiken wie etwa die Phrasenstrukturgrammatik angewendet.



Grammatik :
S → NP + VP
NP → T + N
VP → V + NP
T → die
N → Katze | Milch
V → trank

S-Satz, NP-Nominalphrase, VP-Verbalphrase, N-Substantiv, V-Verb und T-Artike.

Abbildung 2.6.4: Beispiel einer Phrasenstrukturgrammatik mit Ableitungsbaum

In Abb. 2.6.4 steht ein Satz, alle Regeln die aufgerufen werden, um den Satz zu analysieren und der entsprechende Ableitungsbaum. Stellenweise werden solche Grammatiken auch schon im IR benutzt (vgl. [6]). Die Transformation der syntaktischen Strukturen von einer Sprache in eine andere stellt einen Großteil der Transfer-Systeme dar. Diese intensive Analyse wird vor allem dann notwendig, wenn allein die Stellung eines Wortes im Satz seine exakte Bedeutung bestimmt. Man kann zum Beispiel den Sinn von eng.'run' erst im Kontext gut bestimmen: 'gangsters on the run' und 'to run a business'. Gerade das komplexe Zusammenspiel von Morphologie, Syntax und Semantik macht maschinelle Übersetzung für das Information Retrieval unattraktiv, besonders im multilingualen Fall. Der Aufwand an Ressourcen und Arbeitszeit ist groß.

3. Zusammenfassung und Ausblick

Diese Ausarbeitung gab einen kleinen Einblick in heutige multilinguale Information Retrieval Techniken. Ausgehend von einer Einführung in die Thematik wurden einige Methoden zur Sprachverarbeitung wie Thesaurus-, Korpusbenutzung bei der Anfrageerweiterung, morphologische Analyse, Tagging, Stemmer bei der Wortformenerkennung, n-Gramm Statistiken, Stoppworterkennung bei der Spracherkennung, etc. vorgestellt. Abschließend wurden noch drei Grundarten von maschinellen Übersetzungssystemen angesprochen.
Aufgrund der Vieldeutigkeit von Vokabular und der Unvollständigkeit von Wörterbüchern ist MLIR immer noch sehr problematisch, ein enormer Forschungsaufwand wird (schon seit Jahren) betrieben und ist immer noch gerechtfertigt, da man doch noch weit von einem zufriedenstellenden Ziel entfernt ist.

Literaturverzeichnis

[1] Hull, D. A. und G. Grefenstette: Experiments in Multilingual Information Retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR. 1996.

[2] Oard, D. W.: Alternative Approaches for Cross-Language Text Retrieval. In: AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence, März 1997.
URL : http://www.glue.umd.edu/~oard/research.html

[3] Oard, D. W. und B. J. Dorr: A Survey of Multilingual Text Retrieval. Technischer Bericht CS-TR-3615, Institute for Advanced Computer Studies, Univ. of Maryland, College Park, MD, April 1996.
URL: http://www.ee.umd.edu/medlab/filter/papers/mlir.ps

[4] Croft, W. B., J. Broglio und H. Fujii: Applications of Multilingual Text Retrieval. In: Proceedings of the Twenty-Ninth Annual Hawaii International Conference on System Sciences, S. 98--107. Hawaii, 1995.
URL: http://www.cs.umass.edu/

[5] Frakes, W. B. und R. Baeza-Yates (Herausgeber): Information Retrieval --- Data Structures & Algorithms. Prentice Hall, Englewood Cliffs, New Jersey, 1992.

[6] Salton, G. und M. J. McGill: Information Retrieval: Grundlegendes für Informations-wissenschaftler. McGraw-Hill, Hamburg, 1987. original: Introduction to Modern Information Retrieval ,1983.

[7] Srinivasan, P.: Thesaurus Construction, Kapitel 9, S. 161--174. In: Frakes und Baeza-Yates

[8] Soergel, D.: Multilingual thesauri in cross-language text and speech retrieval. In: AAAI Spring Symposium on Cross-language Text and Speech Retrieval, AAAI. März 1997. URL: http://www.ee.umd.edu/medlab/filter/sss/papers/

[9] Loukachevitch, N. V.: Knowledge Representation for Multilingual Text Categorization. In: AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence, März 1997.

[10] Zunker, G. und R. Rapp: Maschinenlesbare deutsch- und englischsprachige Textkorpora, Band 2, S. 383--390. 1994. In: Sprache-Sprechen-Handeln. Akten des Linguistischen Kolloquiums, Graz 1994.

[11] Qiu, Y. und H. Frei: Concept Based Query Expansion. In: R. Korfhage (Herausgeber), Proc. o. t. Conf. on R & D in IR, Band 16 von ACM (SIGIR), S. 160--169. Springer, Pittsburgh, PA, USA, Juni 1993.

[12] Zunker, G. und R. Rapp: Wort-Kookkurenzen als Grundlage eines Algorithmus zur Terminologieextraktion. Niemeyer, Tübingen, 1996. In : Kognitive Aspekte der Sprache, Akten des 30. Linguistischen Kolloquiums, Gdansk, 1995.

[13] Rigau, G., J. Atserias und E. Agirre: Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. In: Proc. of joint EACL/ACL 97. Madrid, Spain 1997. Association of Computational Linguistics, ACL.

[14] Sheridan, P. und J. P. Ballerini: Experiments in multilingual information retrieval using the SPIDER system. In: G. Grefenstette (Herausgeber), Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva, August 1996.

[15] Davis, M.: New experiments in cross-language text retrieval at NMSU's Computing Research Lab. November 1996. approach of the New Mexico State University.
URL: http://crl.nmsu.edu/

[16] Hutchins, W. J. und H. L. Somers: An Introduction to Machine Translation. 1992.