Harald H. Zimmermann F-100-23 05.10.83
Krishna Sharma Zim/AL

Mehrsprachige Automatische Indexierung mit CTX

Es gibt - grob gerechnet - zwei große Problemkreise im Bereich von Indexierung und Retrieval. Einmal ist es die Frage nach der "richtigen" Textkondensation. Im "einfachen" Fall der Freitextindexierung geht es um die Ermittlung von Textwörtern als Deskriptoren: Hierzu ist mit CTX ein Lösungsansatz gegeben. Mit der Einführung der "gewichteten“ Indexierung sollen hier im Rahmen von IFES weitere Forschungen und Entwicklungen erfolgen. Zum anderen kann die Sprache eines Dokuments bzw. die Dokumentations- oder Indexierungssprache selbst eine Informations- oder Kommunikationsbarriere für der Zugang darstellen,, wenn der Recherchierende diese Sprache (sei es nun eine Fach- oder Fremdsprache) nicht oder nicht genügend beherrscht.

Es existiert inzwischen eine Reihe grundsätzlicher Möglichkeiten, dieses zweite Problem zu bewältigen. Konzentriert man sich auf den Fachexperten als Rechercheur bzw. Informationssuchenden, so könnte für Indexierung und Retrieval beispielsweise eine international abgestimmte Notation (z.B. im Patentbereich die IPC) zugrundegelegt werden. Dies bedeutet jedoch, dass neben der natürlichsprachigen Kommunikations(fach)sprache eine weitere 'Sprache' beherrscht werden muss. Ein Vorteil, der sich daraus ableitet, ist die Möglichkeit der Vereindeutigung (soweit das Thema bzw. der Gegenstand selbst auf derartige Klassen bzw. Kategorien eindeutig abbildbar ist). Über Konkordanzen müssen aber in jedem Fall Zuordnungen zwischen natürlich- oder fachsprachlichen Begriffen oder Begriffsfeldern einerseits und der jeweiligen Notation andererseits hergestellt werden. Zudem sind Notationen bzw. Klassifikationssysteme erfahrungsgemäß weniger flexibel und ergeben darüber hinaus eine geringere "Precision" (bei andererseits hohem Recall), da möglichst Ad-hoc-Verästelungen im Notationensystem vermieden werden sollen.

Eine Alternative zur Notation (oder Klassifikation) stellt die Verwendung von Begriffsbeziehungen (semantischen Relationen) dar. Hierzu liegen sowohl von sprachwissenschaftlicher Seite als auch aus der praktischen Information und Dokumentation genügend Modelle und Verfahren vor. Man muß allerdings von vornherein festhalten, dass die "klassischen" Thesaurus-Konzepte in der Dokumentation, z.B. "kristallisiert" in der Norm DIN 1463, allzu sehr auf die intellektuellen Verfahren der Indexierung und des Retrievals bzw. gedruckte Informationsdienste ausgerichtet sind. Ein Beispiel dafür ist die Einführung so genannter "Vorzugsbenennungen", die ihren wesentlichen Sinn (neben einer gewissen Disambiguierungsfunktion und der - meist vergeblichen - Hoffnung auf eine normierende Wirkung) im "platzsparenden" Druck gefunden zu haben scheinen.

Im hier vorzustellenden Ansatz wird der Begriffsrelationierung grundsätzlich der Vorzug vor einem Notationssystem gegeben. Dies erscheint einsichtig aufgrund der Überlegung, dass dem Benutzer die Verwendung seines "gewohnten" (natürlichsprachig-fachsprachigen) Benennungssystems ermöglicht werden soll. Etwaige Differenzen zwischen dem Benennungssystem des Benutzers und dem Benennungsinventar des IR-Systems können durch geeignete Begriffsrelationierungen' (z.B. der Quasi-Synonymie und Synonymie) behandelt werden. Voraussetzung dazu ist, dass dem Systementwickler und -pfleger solche Differenzen bekannt sind. (Hier zeigt sich bereits, wie wichtig eine Rückkopplung Benutzer/System bei Zweifelsfällen sein kann). Neben der Begriffsrelationierung kann eine Begriffsdefinition in ein IR-System integriert werden. Sie macht zumindest in Zweifelsfällen eine Klärung bezüglich der Bedeutung von Benennungen im Indexierungsinventar möglich. In CTX-II ist diese Funktionsvariante modellhaft bereits realisiert. Insbesondere Online-Systeme bieten hier die Möglichkeit, lexikalische Hilfen unmittelbar in den Suchprozess einzubinden.

Bereits in einzelsprachlichen IR-Systemen wird mit der Verwendung von Begriffsrelationen die Möglichkeit geschaffen, unter Verwendung unterschiedlicher Benennungen zu gleichen Suchergebnissen zu gelangen. Dabei muss, besonders bei der maschinellen Freitextanalyse, in Rechnung gestellt werden, dass - aufgrund der Vagheit natürlichsprachiger Begriffe (das Problem der Mehrdeutigkeit sei einmal ausgeklammert) - in den seltensten Fällen strikte Synonymie-Beziehungen (im streng mathematisch-formalen Sinn) vorliegen: Vielleicht ist dies bei der Verwendung von Abkürzungen anstelle von Langformen (E.V. - EINGETRAGENER VEREIN) noch der Fall, vielleicht auch noch für unterschiedliche Bezeichnungen von Gegenständen (FAHRSTUHL/AUFZUG) oder Tieren (GIMPEL/BUCHFINK); problematischer wird es schon bei Nahrungsmitteln (WECK/SEMMEL). Dennoch können mit der Einführung spezifischer Relationen (z.B. der Assoziation und der Quasisynonymie) Hilfsmittel geschaffen werden, die den Recherchierenden - im Bewusstsein der Unterschiede - Möglichkeiten einer Verkürzung der Suchanfrage und in diesem Sinne einer Erweiterung des Selektionsfeldes geben.

Die Einbeziehung von Texten/Dokumenten in verschiedenen Sprachen und deren Indexierung stellt einen Spezialfall dieser Begriffsrelationierung dar. Setzt man voraus, dass Texte mittels einer Freitextindexierung in Analogie zum CTX-Verfahren auf natürlichsprachige Benennungen abgebildet wurden, so kann eine Vergabe sprachpaarspezifischer Synonymie- oder Quasisynonymierelationen die Möglichkeit eröffnen, verschiedensprachige Datensammlungen in den Retrievalprozess mit einzuschließen.

Die damit verbundene Problematik ist allenfalls quantitativ, keinesfalls qualitativ anders als bei monolingualen Systemen. in den monolingualen Relationen Synonymie, Quasisynonymie und

Assoziation sind identische Funktionen bereits vorgegeben. Es liegt also nahe, derartige Relationstypen auch für eine multilinguale Indexierung heranzuziehen.

Bereits die lexikalische Zuordnung von ausgangs- und zielsprachlichen Benennungen schafft somit die Möglichkeit, in einer anderen Indexierungssprache erschlossene Dokumente in den Retrievalprozess einzubeziehen. Geht man davon aus, dass der Recherchierende die Originalsprache des Dokuments beherrscht, so handelt es sich um eine "reine" Verkürzung des Retrievalprozesses. Kennt er die Originalsprache jedoch nicht, so ist es immerhin möglich, bereits über die einfache systemseitige Zuordnung von Deskriptoren an mögliche Quellen heranzuführen, die dann ggf. auf anderem Wege weiter verarbeitet werden müssen.

Als Ergänzung und Alternative bietet sich die Möglichkeit, dass bei der Indexierung gleichzeitig der Titel, das Abstract und eventuell auch der Volltext in die Zielsprache maschinell übersetzt und miterschlossen werden. Damit ergeben sich weitere Möglichkeiten für die CTX-Entwicklung. Dementsprechend muss das CTX-Verfahren dahingehend erweitert bzw. ergänzt werden, dass es nicht mehr allein eine (oder mehrere) einzelsprachliche maschinelle Analysekomponenten, sondern auch Transfer- und Synthesekomponenten und somit ein MÜ-System in sich inkorporiert. Das nächstliegende in diesem Zusammenhang ist, eine Synthese der Indexierungsfunktion von CTX und der Übersetzungsfunktion des Saarbrücker Übersetzungssystems SUSY zu bilden.

Von einer ausführlichen Beschreibung von CTX und SUSY wird hier Abstand genommen, da dies mehrfach und hinlänglich geschehen ist (z.B. Zimmermann, Maas, Luckhardt)*. Jedoch dürfte es für das Verständnis der folgenden Ausführung hilfreich sein, das Funktionsprinzip von beiden an Hand vereinfachter Skizzen zu vergegenwärtigen:

Nach der abgeschlossenen Analyse werden im CTX-Verfahren die formal inhaltlichen Stichwörter (Deskriptoren) bestimmt. Sodann werden diese zusammen mit dem präformatierten Text in GOLEM formal aufbereitet und anschließend daran in die Datenbank eingespeichert.

Der SUSY-Ablauf nach der Analyse sieht wie folgt aus: Die Analyseergebnisse bilden den Input für den Transfer, in dem dann die eigentliche Überführung in die Zielsprache stattfindet. Die Übersetzungen werden dann semantisch vereindeutigt. Danach erfolgt die syntaktische und abschließend die morphologische Synthese der vereindeutigten Übersetzung..

* Zimmermann, H., Kroupa, E., et al.: CTX Ein Verfahren zur computergestützten Texterschließung (FB-ID-8300 b), Karlsruhe, 1983

Maas, Heinz-Dieter: Das Saarbrücker Übersetzungssystem SUSY. In: Sprache und Datenverarbeitung 1. (1978)

Luckhardt, Heinz-Dirk: Weshalb SUSY mehrsprachig ist? In Linguistische Arbeiten, neue Folge, Heft 3, Sonderforschungsbereich 100. Saarbrücken 1982.