Bilinguale Suche mit der SENTRAX-Technologie

Myrja Marx, Suriya Na nhongkai

Universität Hildesheim
Institut für Mathematik und Angewandte Informatik
Marienburger Platz 22
31141 Hildesheim
myrja@gmx.de, iamsuriya@yahoo.com

Zusammenfassung

Bei der krosslingualen Suche vermindert eine ungenügende Übereinstimmung mit den Formulierungen im gesuchten Dokument oft die Leistungsfähigkeit der Suche. Hinter der SENTRAX (Essence Extractor Engine) liegen zwei Container (indexierte Dokumente), die für die bilinguale Suche zusammenwirken. Sie entstehen aus der Verarbeitung von nahe zusammenstehenden, bedeutungstragenden Begriffen (Kookkurrenzen) in den zu durchsuchenden Dokumenten und erlauben eine Definition sowie Übertragung von "Konzepten", die zwar durch Worte ausgedrückt oder beschrieben werden, aber eine gewisse Unabhängigkeit von der spezifischen Wortwahl haben. Hierbei kann die Übertragung eines Konzeptes – statt der wortweisen Übersetzung der Anfrage – die Mehrdeutigkeiten entscheidend vermindern, da das Konzept den assoziierten Zusammenhang mit den übersetzten Begriffe bewahrt und die Verbindung zu den Umgebungen in den Texten herstellt. Somit kann sichergestellt werden, dass die dahinter liegenden Dokumente von den gleichen bzw. ähnlichen Themen handeln. Durch grafische Darstellung sind die mit den Suchwörtern assoziierten Begriffe in Ausgangs- und Zielsprache vergleichbar.

Abstract:

A insufficient match of keywords on crosslingual search mostly reduces the capability of information retrieval. For a bilingual search with SENTRAX (Essence Extractor Engine) two containers (“indexes”) are being used, which accrue of word cooccurrences and allow a definition as well as the transmission of concepts. Hereby, the transfer of a concept can minimize the ambiguity of terms because the associated correlation of terms is preserved. Thus, it is ensured that the documents are dealing of similar topics. Due to the graphic display the associated keywords are comparable to source and target language.

1 Einleitung

Die fortschreitende Globalisierung stellt viele Menschen beinahe täglich vor die Herausforderung sich nicht nur mit Informationen in der Muttersprache, sondern in verschiedenen Sprachen auseinanderzusetzen, um sich umfassend zu informieren.

Mit der Zunahme des Datenumfangs erhöht sich jedoch der Bedarf an geeigneter Suchtechnologie. Die meisten Nutzer von Suchmaschinen wissen wenig über Retrievalmethoden und kennen häufig den Gesamtbestand der Dokumentsammlung nicht, in dem sie suchen. Infolgedessen fällt die richtige Anfrage schwer, obgleich sie als wichtiger Schlüssel zur Lösung betrachtet werden muss.

Grootjen und van der Weide haben auf die Schwierigkeit der Anfrageformulierung hingewiesen (Grootjen, F. A.; van der Weide, Th. P., 2002). Zum einen haben sie sich mit der Fragestellung beschäftigt, ob der Benutzer konkret weiß, was er sucht und ob er weiß wie eine optimale Anfrage formuliert werden kann, um die gesuchte Information zu erhalten. Eine gute Anfrage erfordert, dass der Nutzer vorhersieht, welche Ausdrücke in den gesuchten Dokumenten stehen. Das heisst, er muss sich innerhalb von kürzester Zeit einen Überblick über die Dokumentensammlung verschaffen. Unerfahrene Nutzer beispielsweise, haben Probleme herauszufinden, um welche Themen es im entsprechenden Korpus geht. Anhand der Interaktion zwischen dem Nutzer und der Maschine kann der Suchende sein Verständnis für das zu durchsuchende Material langsam aufbauen. Dieser Prozess wird von Spink untersucht (Spink, A., Saracevic, T., 1997). Obwohl die Mensch-Maschine-Interaktion für ungeübte Benutzer am Anfang doppelt schwer ist (bedingt durch die sog. Nutzerbezogene Komplexität), hat sie auch positive Seiten, denn Wahlfreiheiten und Eigengestaltung stellen hohe Anreize und wirken oft motivationsfördernd (vgl. Bentz, H-J., 2005).

2 Bilinguale Suche

Einhergehend mit der Zunahme des Datenumfangs und der Globalisierung kommt die Idee der bilingualen Suche ins Spiel. Die Grundidee des herkömmlichen Ansatzes ist es, zwei oder mehrere Information Retrieval (IR)-Systeme zu verbinden. Statt einen Korpus zu übersetzen, wird die Anfrage ins jeweilige System übertragen und dort separat bearbeitet. Die Ergebnisse der Systeme werden gesammelt und in einer Liste sortiert ausgegeben (siehe Abbildung 1).

Abbildung 1 Context-Liste der SENTRAX (aus: Na nhongkai, S., 2006)

Es wird angenommen, dass das IR-System von genügender Qualität ist. Wenn ein IR-System mit unterschiedlichen sprachlichen Korpora verbunden werden soll, ergeben sich zwei Fragen:

Zum einen stellt sich die Frage, ob die übertragene Anfrage für die fremde Sprache gut geeignet ist. Da eine Übersetzung oft nicht eindeutig ist, kann nicht sicher festgestellt werden, ob die originale Anfrage die gleiche semantische Bedeutung mit der übersetzten Anfrage hat. Falls die Anfragen nicht gleich sind, erhält man möglicherweise nicht das gewünschte Ergebnis.

Hierbei handelt es sich um das Problem der Mehrdeutigkeit von Wörtern. Mehrdeutigkeit betrifft in diesem Kontext Homonymie, Polysemie und Synonymie.
Die Homonymie nimmt Bezug auf Lexeme, die gleich geschrieben sind, aber unterschiedliche Bedeutungen haben („Tau“ = „Seil“ und „Tau“ = „morgendlicher Niederschlag“). Die Polysemie hingegen nimmt Bezug auf die Idee eines einzelnen Lexems mit mehreren relevanten Bedeutungen (z. B. kann "Pferd" für ein Tier, ein Turngerät oder eine Schachfigur stehen).

Das Synonym weist auf die Beziehung zwischen unterschiedlichen Lexemen mit gleicher Bedeutung hin (beispielsweise senkrecht – vertikal, Orange – Apfelsine).
Aufgrund dieser Vielschichtigkeit kann die übersetzte Anfrage den ursprünglichen Sinngehalt der formulierten Anfrage oft nicht bewahren. Eine Möglichkeit die ursprüngliche Bedeutung der formulierten Anfrage nicht zu verlieren, ist bei der SENTRAX dadurch gegeben, dass das gesamte Konzept übertragen wird und nicht nur einzelne Ausdrücke.

Die zweite Frage ist, ob das IR-System auf den beispielsweise englischen Korpus und den Korpus der deutschen Sprache gleich gut passt. Dabei geht es um die Vergleichbarkeit der sprachlichen Eigenschaften. Die englische und die deutsche Sprache zum Beispiel entstammen gleichen Ursprungs, haben sich inzwischen unabhängig voneinander entwickelt. Aus diesem Grund kann nicht eins zu eins übersetzt werden. So etwa ist die deutsche Sprache eine stark flektierende Sprache, die englische hingegen nicht (vgl. Rapp R., 1999). Im Deutschen tauchen oft Komposita auf, z. B. "Finanzamt" oder „Sprachforschung“, im Englischen weniger, z. B. „finance office“ oder „linguistic research“. Aus diesem Grund bedarf es einer gewissen Vorarbeit, um die Symmetrien der Sprachen kenntlich und somit verwertbar zu machen. Eine weitere Möglichkeit besteht darin, ein Wort innerhalb seiner relevanten Umgebung zu betrachten.

Ein neuartiger Ansatz wird wird durch die Suchanwendung SENTRAX aufgezeigt. Hierbei handelt es sich um ein duales IR-Modell, bei dem anhand von Hilfstechniken, wie z. B. Relevanz-Feedback, die Suchanfrage erweitert werden kann, indem Zusatzwörter zu den Suchwörtern anteilig hinzukommen und dem Nutzer eine Interaktion während der Suche anbieten. Besondere Vorteile der SENTRAX sind durch die Lernmöglichkeiten während des Suchprozesses und die Ideen- bzw. Begrifferweiterungen mittels eines Konzeptnetzes gegeben, was bei einer konventionellen Suchmethode nicht vorliegt. Weil der Suchende mit der SENTRAX auf der Konzeptschicht anstatt auf der Wortschicht arbeitet ist es einfacher, weiterführende Zusatzbegriffe auszuwählen, um so das Suchkonzept zu verschärfen und dabei nicht von der Suchrichtung abzulenken (vgl. Na nhongkai, S., 2006).

Das klassische Vorgehen, eine bilinguale Suche aufzubauen, ist die Verknüpfung von zwei IR-Systemen durch eine Übertragungsmethode. Das größte Problem dabei ist die Mehrdeutigkeit der Übersetzung. Ein weiteres Hindernis liegt im Mangel an Ressourcen für die Entwicklung, z. B. der Mangel an elektronisch lesbaren Wörterbüchern für die entsprechende Sprache. Obwohl die Umsetzungssprache solche Mängel umgehen kann, erhöht sich bei wiederholten Übersetzungen trotzdem die Möglichkeit der Mehrdeutigkeit. Bei einer Suchanfrage wird oft ein Konzept (Na nhongkai, S., Bentz, H.-J., 2005) benutzt, um etwas zu definieren, damit die beabsichtigte eindeutige Bedeutung erkannt und verstanden werden kann. So ein Konzept ist in der Regel aus mehreren Eigenschaften zusammengesetzt. Dadurch wird es möglich, dass das gesamte Konzept begriffen wird, obwohl einige Eigenschaften in der Beschreibung fehlen. Anhand dieser Vorgaben kann mittels einer toleranten Konzeptübertragung die Mehrdeutigkeiten bei der bilingualen Suche vermieden werden, weil sich der Kern des Konzeptes nach der Übertragung noch erhält.

Um beim bilingualen Suchschritt die Begriffe der Ausgangsprache in die Zielsprache zu übertragen, stehen zwei Möglichkeiten zur Verfügung. Das elektronisch lesbare deutsch-englische Wörterbuch von der TU-Chemnitz[1] und die Transfermatrix, die der Methode von R. Rapp (1999) entspricht. Dabei wird bei der Übertragung der korpusbasierten Begriffe das gesamte (jeweilige) Suchkonzept bewahrt. Daraufhin werden die Konzeptvergleichsmaße entworfen. Die im Hintergrund stehende Theorie für konzeptionelle Graphabgleichung ist hier an die Vorgaben von M. Montes-y-Gómez, A. López-López und A. F. Gelbukh (2000) angelehnt (vgl. Na nhongkai, S., 2006). Diese Vergleichsmaße dienen zur besseren Auswahl von Suchkonzepten. Die Bausteine des bilingualen Systems werden mit der SENTRAX passend zusammengestellt (siehe Abbildung 2).

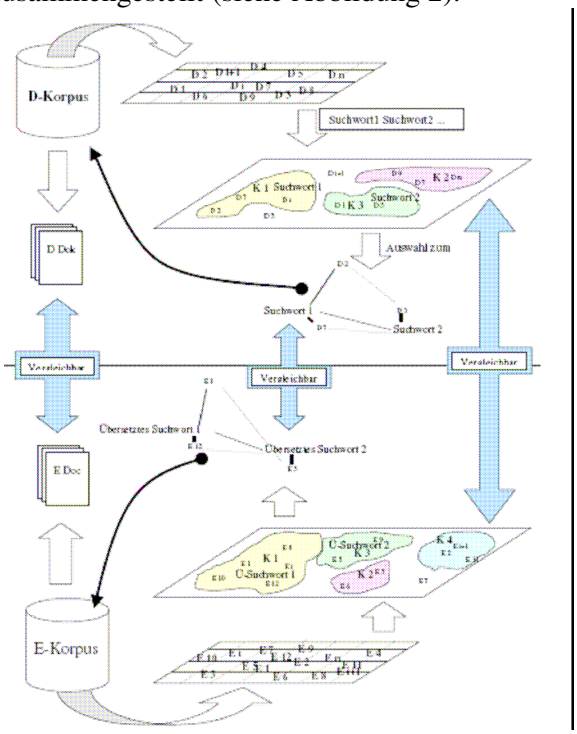

Abbildung 2 Grundidee der bilingualen Suche mit der SENTRAX-Technologie

[1] ftp://ftp.tu-chemnitz.de/pub/Local/urz/ding/de-en/ ,Verifizierungsdatum am 10.10.2006.

3 Mehrsprachige Suche mit der SENTRAX-Technologie

Die SENTRAX ist eine IR-Anwendung mit einer grafischen Mensch-Maschine-Schnittstelle. Sie bietet vier nützliche Funktionen an (vgl. Na Nhongkai, S., Bentz, H.-J., 2005). Anhand der Mensch-Maschine-Schnittstelle kann der Nutzer flexibler arbeiten, als mit einer herkömmlichen IR-Anwendung. Zwei von vier Funktionen werden für Grafikdarstellungen verwendet, wobei Tipp- bzw. Schreibfehler berücksichtigt (LexicoMap, Abbildung 3) und eine andere für die Begriffe zur Erweiterung der Suche verwendet werden (Context-Map).

Abbildung 3 Eingabemaske der SENTRAX mit fehlerhafter Suchanfrage („Konflik“) und korrekter Ausgabe („Nahost-Konflikt“)

Zwei weitere Funktionen, „Treffer“- und „SimilarDoc“ (vgl. Bentz, H.-J., 2006) liefern die Dokumente entsprechend den Suchwörtern als Liste zurück, wobei die jeweilige Prozentangabe einen Zusammenhang zwischen den Dokumenten und den Suchwörtern darstellt. Die Funktion SimilarDoc, kann erst auf der Basis einer erzeugten Trefferliste aktiviert werden. Der Nutzer wählt irgendein Dokument aus und erhält so eine neue Trefferliste, die nun alle ähnlichen Texte im Zusammenhang zum ausgewählten Dokument zeigt. So lassen sich z. B. auch Dubletten oder Fast-Dubletten auffinden.

Die auffälligen Merkmale der SENTRAX sind die grafischen Funktionen. Eine erste Grafikfunktion (LexicoMap) zeigt dem Suchenden diverse Schreibweise als Empfehlung (siehe Abbildung 3). Die im Hintergrund wirksame Technik fußt auf einer SpaCAM. Die SpaCAM-Technik wird in Heitland, M. (1994) beschrieben sowie in M. Hagström, M. (1996). Eine zweite Grafikfunktion ist die zweidimensionale grafische Mensch-Maschine-Schnittstelle, die ContextMap (siehe Abbildung 4, aus: Na nhongkai, S., 2006), die durch die assoziierten Wörter gefüllt wird. Die auf der Grafik-Oberfläche ausgegebenen Wörter haben nicht nur Beziehungen zu der Suchanfrage, sondern auch untereinander. Diese Beziehungen werden in einer kleinen Gruppe klassifiziert, in der eng verwandte Begriffe erfasst werden.

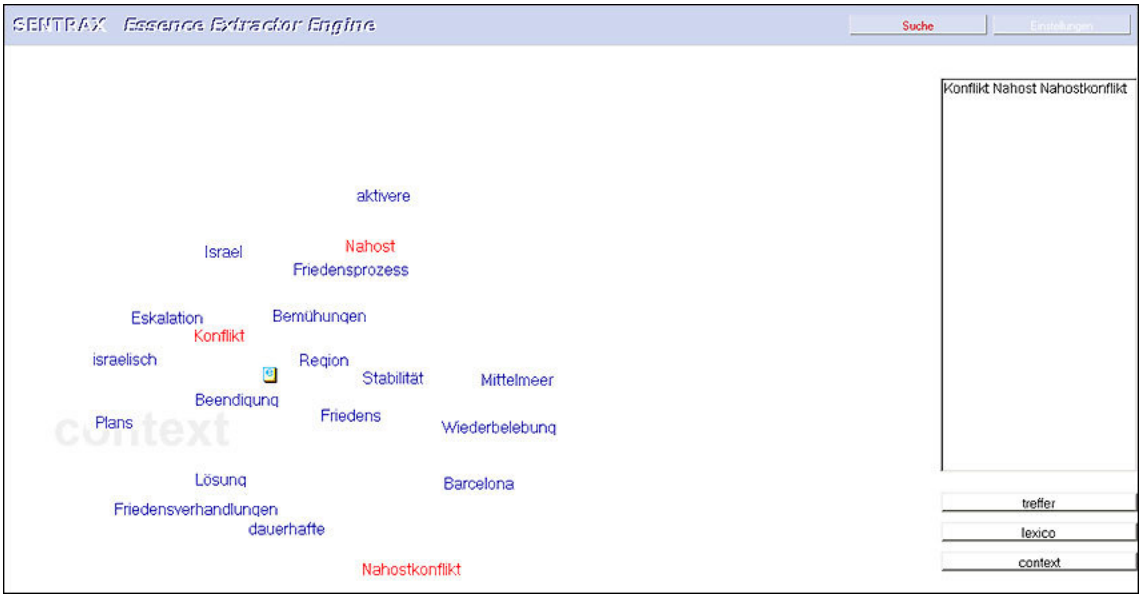

Abbildung 4 ContextMap mit Vorschlägen für eine weiterführende Suche

Der Hintergrund dieser IR-Strategie basiert auf der statistischen Wörterhäufigkeit und der Relation zwischen den Wörtern. Die erste Ordnung bzw. direkte Assoziation wird aus der Häufigkeit berechnet, mit der die Wörter miteinander auftreten. Die Beziehungen der Wörter wie bei Synonymen, Analogien und Antonymen werden durch die zweite Ordnung bzw. indirekte Assoziation ermittelt (vgl. Ackermann, M. 2000). Die Clusterstrategie wird dazu verwendet, um die deutliche Beziehung zwischen den Wörtern auf dem Bildschirm zu zeigen und gleichzeitig in der Gruppe zu klassifizieren. Durch die Mensch-Maschine-Schnittstelle können standardisierte Precision- und Recallwerte nicht gemessen werden.

3.1 Hypothesen

Ausgangspunkt sind zwei Datensammlungen "Deutsch" (D) und "Englisch" (E), die parallel seien. (Eine durch die Parallelität bereits mitgegebene Zuordnung dient lediglich zur späteren Überprüfung unserer Entscheidung, ob das von der Maschine mittels des neuen, schon skizzierten Vorgehens gefundene Dokument das gesuchte Zieldokument in der anderen Sprache ist.). Für D und E werden zunächst unabhängig die SENTRAX-Container erzeugt (vgl. Na Nhongkai, S., Bentz, H.-J., 2005).

Die Vermutung ist, dass bei dieser Datenlage die beiden internen Konzeptnetze eine ähnliche Struktur haben. "Ähnlich" im Sinne der parallelen Dokumentenpaare:

Die Umgebung eines Dokuments in seinem Index "entspricht" der Umgebung seiner Übersetzung im anderen Index. Sollte diese Vorstellung zutreffen, dann müsste die (automatische) Übertragung der einem Dokument hier zugeordneten Wortgruppen ein Cluster von Wörtern dort erzeugen, zu denen das Zieldokument unter allen am besten passt. Als Vorteil bei dieser Methode ist zu erwarten, dass Mehrdeutigkeiten durch die (später vollautomatische) Übersetzung nicht stören, da Bestandteile, die keine Korrespondenzen in den Dokumenten haben, durch den SENTRAX Automatismus unwirksam bleiben. Hierdurch

entsteht eine enorme Reduktion der kombinatorischen Möglichkeiten. Der Umstand, der hier ausgenutzt wird, vergleicht sich mit folgender, "natürlicher" Situation: Wenn man mit einem Menschen spricht, der unsere Sprache nicht besonders gut kann, dann versteht er Gesprächspassagen nicht, in denen Wörter oder Phrasen vorkommen, die ihm fremd sind. Er versteht eben nur das, was in seinem Gehirn eine Korrespondenz in seiner Muttersprache besitzt. Insofern bleibt zuweilen eine große Ausdrucksvielfalt auf unserer Seite nutz- und wirkungslos, da das Verstehenspotenzial auf der anderen Seite sehr eingeschränkt ist.

Die Suchwörter und ihre umgebungsbedingten assoziierten Begriffe, die in der "ContextMap" als einziger Treffer im Container "D" auftreten, werden als Schlüsselwörter bzw. vorgeschlagene Zusatzbegriffe für die Suche im Container "E" verwendet. Eventuelle Mehrdeutigkeiten im Wörterbuch werden unbesorgt übernommen. Die in den Korpora existierende Übersetzung der Schlüsselwörter sollte zum parallelen englischen Dokument entsprechend des vorherigen deutschen Dokuments führen. Diese Vermutung und Methode ist symmetrisch, lässt sich also auch von "E" nach "D" verwenden.

Bei einem großen Container, sollte die Durchmischung der Begriffe auf dem Konzeptnetz gemäß derselben Anfrage möglichst gering werden. Sollte dieses geschehen, könnte das korpusbasierte Semantiknetz mit der ContextMap- Funktion erschaffen werden.

Gibt es kein paralleles Dokument in der Zielsprache entsprechend der Anfrage, sollte das Konzeptnetz zu anderen, ähnlichen Dokumenten führen.

In der Situation, dass der Zielcontainer viel größer oder viel kleiner als der Ausgangscontainer ist, wird die Kookkurrenzhäufigkeit nur gering abweichen, aber die Antwort, nämlich das parallele Dokumentenpaar, sollte noch in dem Dokumententreffer auftauchen. Die bilinguale Suche durch das Konzeptnetz könnte also dennoch funktionieren.

Hat der Zielcontainer mehrere Sprachen als zwei, sollte die bilinguale Suche durch die SENTRAX trotzdem gut funktionieren, weil ein Zusammenhang zwischen Wörtern unterschiedlicher Sprachen nur selten zustande kommen wird.

Für eine nicht-parallele Textsammlung werden die deutschen und englischen Dokumente durch die SENTRAX unabhängig verwaltet. Befänden sich die relevanten Dokumente in beiden Containern, sollten die Konzeptnetze miteinander vergleichbar sein. Wäre der Zusammenhang zwischen den Konzeptnetzen zu schwach, hätten die entsprechenden Dokumente möglicherweise keine Relation zueinander.

3.2 Experimente

Durch die vorhergehend formulierten Hypothesen ergeben sich vier Fälle: (1) der Standardfall, (2) Sonderfälle, (3) die Konzeptnetzänderung und (4) Suche im nicht-parallelen Korpus. Die Verhältnisse im Suchprozess werden für jeden der vier Fälle beobachtet. Die Ergebnisse im Standardfall und in den Sonderfällen bestätigen, dass die bilinguale Suche mittels Konzeptnetzen nicht nur das gesamte Such-Konzept bewahren kann, sondern auch stabil ist. Die Übersetzung wird stets mit dem Online-Wörterbuch http://www.leo.org/ manuell vorgenommen; in der später kommenden Ausbaustufe soll hier auch eine automa-

tische Übersetzung eingeschaltet werden können (vgl. Na Nhongkai, S., Bentz, H.-J., 2005).

Der Suchbedarf beschränkt sich jedoch nicht nur auf eine Richtung wie beim Standardfall Die bilinguale Suche mittels Konzeptnetz funktioniert auch in der Gegenrichtung Englisch → Deutsch (siehe Abbildung 5, aus: Na nhongkai, S., 2006).

Die englische Anfrage in diesem Beispiel ist aus den Wörtern „energy“, „saving“, „ecology“, „environment“ und „research“ zusammengesetzt. Dieses Konzept führt zu dem E-Dokument „ep-01-06-13.txt11“. Die Übersetzung der Anfrage mit Hilfe des Online-Wörterbuchs ist „Energie“, „sparend“, „Ökologie“, „Umwelt“ und „Forschung“. Obwohl das Wort „sparend“ nicht gefunden werden kann, taucht das Wort „einzusparen“ in der deutschen Umgebung auf. Nach der Auswahl des zusätzlichen Attributs „einzusparen“ wird das deutsche parallele Dokument „ep-01-06-13.txt11“ getroffen.

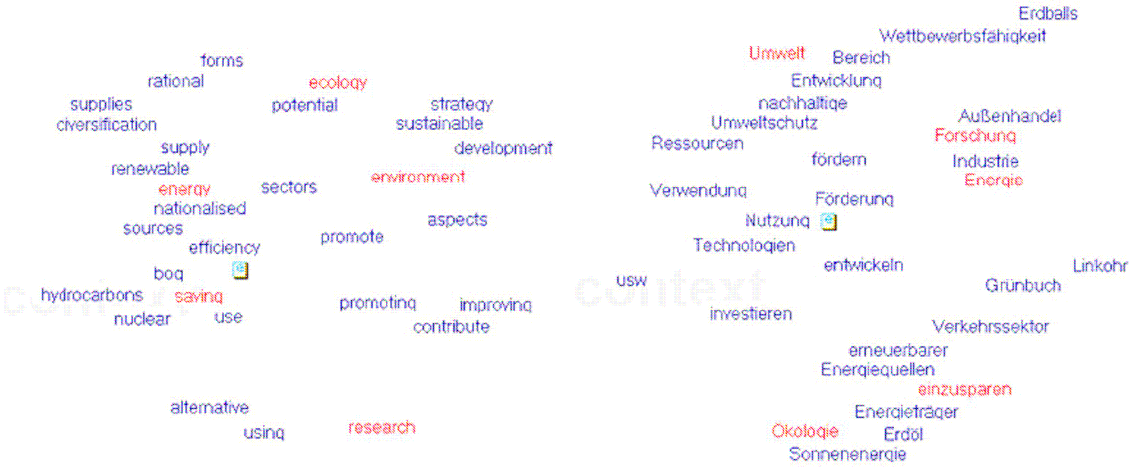

Abbildung 5 Links: ContextMap in der Ausgangsprache Englisch. Rechts: die ContextMap in der Zielsprache Deutsch

Anhand dieses Beispiels wird belegt, dass die Suche in der Gegenrichtung (E→D) ebenfalls funktioniert. Bemerkenswert dabei ist, dass der Zusammenhang der Attribute andere Attribute hervorbringt, wie in diesem Beispiel das Wort „einzusparen“. Die anderen Übersetzungspaare sind natürlich enthalten, z. B. „promote – fördern“, „using (use) – Nutzung (Verwendung)“, „renewable – erneuerbarer“ usw. Das Wort „sectors“ kann vielleicht dem Wort „Bereich“ oder dem Wort „Verkehrssektor“ entsprechen, weil es auf Englisch allein oder mit einem anderen Wort zusammen stehen kann.

Bei den Sonderfällen werden vier Fälle betrachtet: (1) der Zielcontainer ist viel größer als der Ausgangscontainer (2) der Zielcontainer ist kleiner als der Ausgangscontainer (3) das relevante Dokument wird im Zielcontainer entfernt (4) der Zielcontainer wird mit anderen, fremden Texten erweitert. In allen 4 Fällen liefert die SENTRAX zufrieden stellende Ergebnisse, selbst wenn relevante Dokumente aus dem Korpus entfernt werden (vgl. Na Nhongkai, S. 2006, 149 ff.).

Im Falle einer Konzeptnetzänderung wird untersucht, wie das Konzeptnetz bei Größenänderung des Containers verändert wird. Dabei werden zwei Fragestellungen berücksichtigt: Zum einen ob die Entwicklung des Konzeptnetzes bezüglich des deutschen Containers ähnlich zu der Entwicklung des Konzeptnetzes bezüglich des englischen Containers ist und zum anderen, ob sich durch die Vergrößerung des Containers eine Stabilisierung des Konzeptnetzes ergibt.

Die Untersuchungen haben ergeben, dass die Anzahl alter Begriffe bei der Vergrößerung des Containers deutlich höher ist, wenn die neue Textsammlung mit dem vorhergehenden Container zusammen gebildet wurde. Das Änderungsverhalten des deutschen und englischen Konzeptnetzes ist insofern ähnlich, da der inhaltliche Prozess unabhängig von der Sprache abläuft. Dabei ist es irrelevant, ob auf dem englischen oder deutschen Korpus gesucht wird. Wenn eine neue Textsammlung in den Container eingefügt wird, verändert sich die Wortliste hauptsächlich auf den hinteren Rangplätzen. Die neuen Begriffe tauchen in dem Konzeptnetz abhängig davon auf, wie viele Begriffe auf dem Konzeptnetz vom Nutzer eingestellt wurden. Weil das Konzeptnetz von den Termen in der Anfrage abhängig ist, kann man nicht feststellen, wann die Netze in den endgültigen Zustand übergehen. Aber es scheint, dass sich das Konzeptnetz bei stetiger Vergrößerung des Containers langsam entwickelt und stabilisiert (vgl. Na Nhongkai, S. 2006).

Bei der Suche im nicht-parallelen Korpus lässt sich feststellen, dass Konzeptnetze die viele Übersetzungs-vergleichbare Begriffe haben, zu den entsprechenden bilingualen Dokumentenpaaren führen. Durch die sprachliche Vorverarbeitung kann das Problem der unterschiedlichen sprachlichen Nutzungsweise teilweise verhindert werden, indem die Stammform und das Vernachlässigen von unbenötigten Wortarten die Ablenkung durch ungeeignete Begriffe verhindern kann. Als Konsequenz ergibt sich daraus auch eine Änderung der Assoziationsstärke. Dies könnte die sprachliche Symmetrie bringen, die für den Vergleich der Konzeptnetze nötig ist.

5 Fazit

Die Nutzung der zusätzlichen Begriffe aus den relevanten Dokumenten der Ausgangssprache als zusätzliche Übertragungsbegriffe ist eine sinnvolle Methode. Dies kann manuell sowie automatisch erfolgen. Obwohl die manuelle Auswahl einfach und direkt ist, muss der Nutzer viel Zeit aufwenden, um die gefundenen Dokumente zu sichten. Bei der automatischen Auswahl können die ersten *n* gewonnenen Begriffe der Worthäufigkeit oder der Assoziationsstärke nach gemäß der Suchanfrage übernommen werden. Diese Methode kann als „Pseudo-Relevant-Feedback“ bezeichnet werden. Obwohl die Treffer-Funktion der SENTRAX die tolerante Suche mittels der SpaCAM-Technologie ermöglicht, können die getroffenen Dokumente eventuell von dem gesuchten Thema abweichen. Die SENTRAX ist eine Volltextsuchmaschine mittels Musterabgleichung. Wenn ein Dokument viele voneinaner unabhängige Themenbereiche abdeckt, kann die Suche abgelenkt werden. Bei Dokumenten dieser Art können einige Suchbegriffe in einem Thema und andere in einem anderen Themenkomplex vorkommen. Auch wenn die gesuchten Begriffe in einem Dokument weit von einander getrennt stehen, werden sie aufgrund von der Musterabgleichung bei der Volltextsuche als relevant gewertet. Aufgrund der Vielfältigkeit

der Themen und der schwachen Beziehungen durch weit auseinander stehende Wörter sollte ein solches Dokument als unrelevant angesehen werden.

Literaturverzeichnis

Ackermann, Martin (2000): Statistische Korpusanalyse zum Extrahieren von semantischen Wortrelationen. Dissertation. Hildesheimer Informatik-Berichte 1/2000. Hildesheim: Universität Hildesheim.

Bentz, Hans-Joachim (2006): Suchen und Problemlösen in komplexer Umgebung. In: Perspectives on Cognition: A Festschrift for Manfred Wettler. Rapp, Reinhard; Sedlmeier, Peter (Hrsg.). Lengerich: Pabst Science Publishers.

Bentz, Hans-Joachim (2006): Die Suchmaschine SENTRAX. Grundlagen und Anwendungen dieser Neuentwicklung. Universität Hildesheim. Unveröff. Manuskript.

Grootjen, F. A.; van der Weide. Th. P. (2002): Conceptual Relevance Feedback. In: *Proceedingof the 2002 IEEE International Conference on Systems*, Man and Cybernetics, (NLPKE 2002),Tunis, October 2002.

Na nhongkai, Suriya; Bentz, Hans-Joachim (2005): Bilinguale Suche mittels Konzeptnetzen. In: T. Mandl und C. Womser-Hacker (Hrsg.), Effektive Information Retrieval Verfahren in der Praxis: Ausgewählte und erweiterte Beiträge des Vierten Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2005) Hildesheim, 20. Juli 2005. Konstanz: Universitätsverlag [Reihe Schriften zur Informationswissenschaft 45], 2005, 203 – 218.

Na nhongkai, Suriya (2006): Untersuchungen zur sprachübergreifenden, bilingualen Suche mit Hilfe der Konzeptnetz-Technologie der SENTRAX-Engine. Universität Hildesheim. Dissertation.

R. Rapp: Automatic Identification of Word Translations from Unrelated English and German Corpora. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, College Park, Maryland, 519-526, 1999.

Spink, A. und T. Saracevic (1997): Interaction in Information Retrieval: Selection and Effectiveness of Search Terms – Journal of the American Society of Information Science and Technology. 48(8):741-761, 1997.