DISCO
Semantische Ähnlichkeit

White Paper
Version 1.0

linguatools – Peter Kolb und Petra Procházková GbR
Perleberger Str. 55
D-10559 Berlin

http://www.linguatools.de

Was ist DISCO?

DISCO ist eine Java-Klasse zur Abfrage der semantischen Ähnlichkeit zwischen Wörtern. Die Ähnlichkeiten basieren auf der statistischen Auswertung sehr großer Textmengen. Die Java-API stellt u.a. die folgenden Methoden bereit:

- **Semantisch ähnlichste Wörter** zu einem Eingabewort ausgeben: z.B. *Stuhl → Sessel Sofa Bett Tisch Stuhle Lehnstuhl Schreibtisch Schemel Stühlen Stühle ...*
- **Größe der semantischen Ähnlichkeit** zwischen zwei Eingabewörtern anzeigen: sim(*Karpfen*, *Hecht*) = 0,129; sim(*Karpfen*, *Gemüse*) = 0,016; sim(*Karpfen*, *Rakete*) = 0,0.
- **Kollokationen** zu einem Eingabewort ausgeben: z.B. *ergreifen → Flucht Gegenmaßnahmen Initiative Maßregeln Beruf Schutzmaßnahmen Eigeninitiative ...*

Welches **semantische Wissen** Ihnen mit DISCO zur Verfügung steht, lässt sich gut anhand des folgenden Beispiels veranschaulichen, in dem DISCO eine Anzahl Wörter nach ihrer Bedeutung gruppiert. Bei den Wörtern handelt es sich um diese willkürlich ausgewählte Liste:

Getränk Automobil warten Saft Pkw wartet Cola Auto Autos gewartet

Jetzt wird mittels DISCO die semantische Ähnlichkeit zwischen allen Wortpaaren bestimmt. Die so erhaltenen Ähnlichkeitswerte bilden die Eingabe für einen einfachen Clustering-Algorithmus, der die paarweisen Ähnlichkeiten als Graph visualisiert. Das Ergebnis zeigt Abbildung 1.

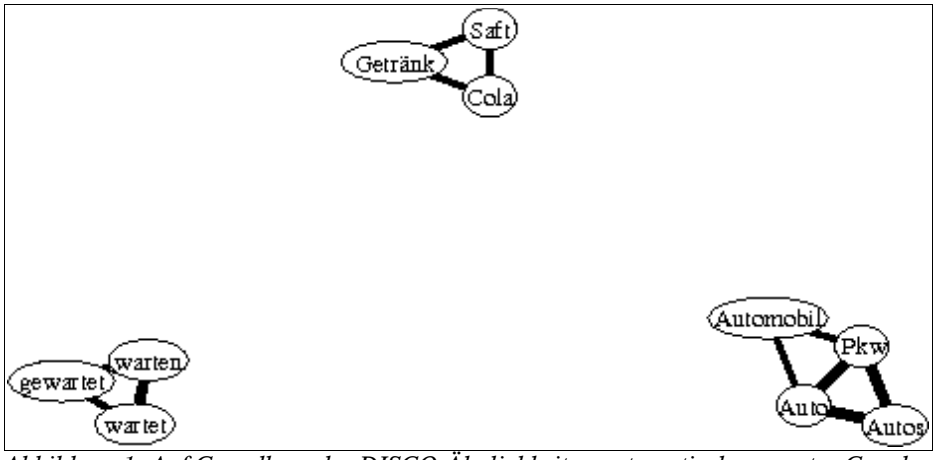

Abbildung 1: Auf Grundlage der DISCO-Ähnlichkeiten automatisch erzeugter Graph.

Zu erkennen ist, dass sich drei Cluster gebildet haben: links unten ein Cluster mit den drei Beugungsformen des Verbs *warten*, ein weiteres Cluster das die Wörter *Getränk*, *Saft* und *Cola* umfasst, sowie ein drittes Cluster mit den Synonymen *Automobil*, *Auto* und *Pkw* und der zu *Auto* gehörenden Pluralform *Autos*.

Mit DISCO steht Ihnen ein Wissen über **Wortbedeutungen** zur Verfügung, das keine andere Ressource auf dem Markt bieten kann. Sie können mit Hilfe von DISCO die Bedeutungsähnlichkeit zwischen *beliebigen* Wörtern bestimmen, eingeschlossen Namen von Firmen, Organisationen und Personen sowie Abkürzungen (siehe Abbildung 2). Und das nicht nur für Deutsch, sondern auch für Englisch, Französisch, Spanisch und viele weitere Sprachen.

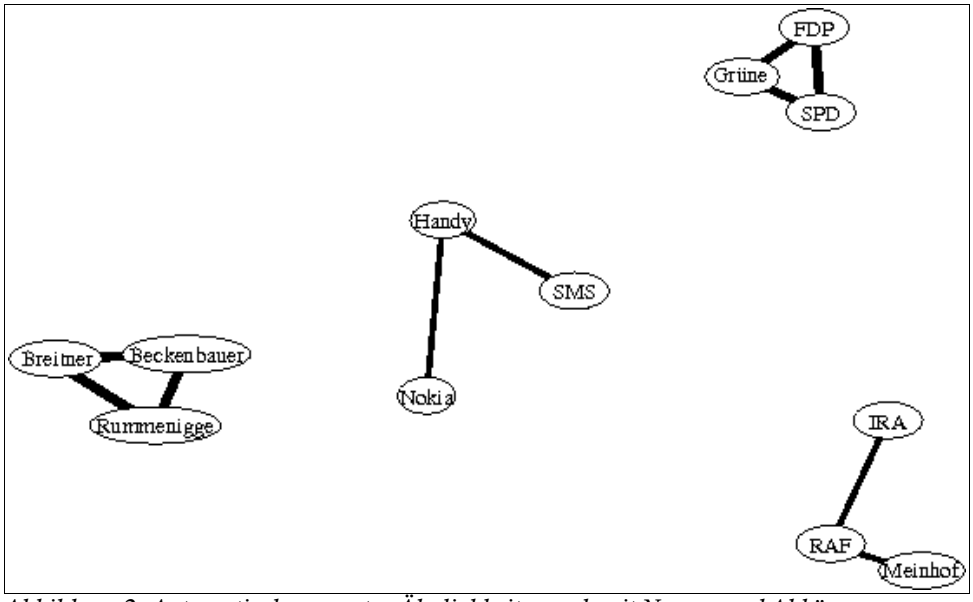

Abbildung 2: Automatisch erzeugter Ähnlichkeitsgraph mit Namen und Abkürzungen.

Anwendungsbeispiele

Wofür lässt sich das semantische Wissen von DISCO nutzen? Im folgenden wird eine Auswahl von Anwendungsfällen dargestellt. Weitere Anwendungsmöglichkeiten finden Sie auf der DISCO-Homepage unter http://www.linguatools.de/disco.html.

Suchmaschinen: Suchtermerweiterung und assoziative Suche
Im Unterschied zu herkömmlichen Volltextsuchmaschinen, die ein Wiederauffinden (engl. *retrieval*) bekannter Dokumente bzw. bereits gewusster Informationen ermöglichen, haben sich aktuelle Text-Mining-Lösungen zum Ziel gesetzt, dem Benutzer dabei zu helfen, Informationen zu finden, von denen er gar nicht wusste, dass es sie gibt. Ein Mittel, das zu erreichen, ist die **assoziative Suche**. Bei dieser Form der semantischen Suche werden dem Nutzer zu seiner Suchanfrage **verwandte Begriffe** angezeigt. Mit diesen kann er seine Anfrage erweitern oder ersetzen. Dadurch wird eine völlig neuartige Form der Suche möglich, nämlich ein intuitives Browsing durch die Begriffswelt der Dokumentensammlung. Auf diese Weise findet man erstens neue Informationen und kann sich zweitens einen Überblick über die Dokumentensammlung verschaffen.
Hat der Nutzer beispielsweise nach *Aspirin* gesucht, liefert DISCO die folgenden Begriffe als semantisch ähnlichste Wörter zu *Aspirin*:

> *Ibuprofen, Paracetamol, Acetylsalicylsäure, Ritalin, Tablette, Schmerzmittel, Diclofenac, ASS, ...*

Diese verwandten Begriffe können – u.U. nach Vorkommenshäufigkeit in der jeweiligen Dokumentensammlung sortiert – als Vorschläge zum Weitersuchen angezeigt werden.

Semantisches Stemming
Ein wichtiges Hilfsmittel um bei Suchmaschinen die Trefferquote zu erhöhen, ist das sogenannte Stemming. Dabei werden nach bestimmten manuell erstellten Regeln die Endungen und Präfixe von den Wörtern abgetrennt. Das Ziel ist, die Wörter einer Wortfamilie, wie z.B. *Gift, Gifte, Giften, giftig, giftiges, vergiften, vergiftet* usw., auf den gleichen Stamm – *gift* – zurückzuführen. Werden

indexierte Dokumente und Suchanfragen auf diese Weise vorverarbeitet, findet die Suchanfrage *Gift* auch Dokumente, die nur die Wörter *Gifte* oder *giftigen* enthalten.
Nachteilig ist bei dieser Methode das Overstemming. Da die Stemming-Regeln auf alle Wörter ohne Rücksicht auf die Wortbedeutung angewandt werden, kommt es häufig zu Fehlern. So ist das Abtrennen der Endung *-ig* im Falle von *giftig* sinnvoll, im Falle von *wichtig* erhält man aber den *Wicht*, was das Suchergebnis nicht gerade verbessern dürfte.
Dieses Problem kann mit Hilfe von DISCO gelöst werden, indem nur Wörter auf den gleichen Stamm zurückgeführt werden, die ein Mindestmaß an semantischer Ähnlichkeit aufweisen. Fehler wie *wichtig* → *Wicht* können durch diese zusätzliche Berücksichtigung der Semantik vermieden werden.
Aber DISCO kann noch mehr. Lässt man sich z.B. die 100 semantisch ähnlichsten Wörter von *Gift* ausgeben, findet man darunter *Gifte*, *Giften*, *Giftes*, *Vergiftung*, *Vergiftungen*, *giftig* usw. Filtert man die Liste der semantisch ähnlichen Wörter mit einem herkömmlichen String-Ähnlichkeitsmaß (wie z.B. Levenshtein), so dass nur die orthografisch ähnlichen Wörter zum Ausgangswort übrigbleiben, erhält man eine Gruppe von morphologisch und semantisch ähnlichen Wörtern, die man zum Stemming einsetzen kann. Zum Aufbau eines Stemmers ist keine arbeitsaufwändige Analyse der einzelsprachlichen Grammatik und anschließende Aufstellung von Regeln mehr nötig.

Übersetzung
Eine häufige Schwierigkeit bei der Übersetzung ist das Problem der Wortwahl. Viele Wörter besitzen mehr als eine Bedeutung, und somit findet man im Wörterbuch mehrere mögliche Übersetzungen. Zur Auswahl der korrekten Übersetzung muss der Kontext, in dem das Wort vorkommt, berücksichtigt werden. Normalerweise werden dafür von Hand Regeln geschrieben, was mit großem personellen Aufwand verbunden ist. Mit DISCO geht das viel einfacher. Angenommen, wir wollen das mehrdeutige Wort *Aufzug* im folgenden Kontext ins Englische übersetzen:

> *Der Aufzug ist steckengeblieben.*

Für *Aufzug* findet man in einem Wörterbuch die Übersetzungen *lift*, *act* und *hoist*, für *steckengeblieben* die Übersetzung *stuck.* Nun bestimmt man mittels DISCO die semantische Ähnlichkeit zwischen den möglichen Übersetzungen und dem Kontextwort *stuck*:

sim(*lift*, *stuck*)	= 0,032
sim(*act*, *stuck*)	= 0,010
sim(*hoist*, *stuck*)	= 0,008

Die korrekte Übersetzung *lift* weist also die höchste semantische Ähnlichkeit mit dem Kontext (hier nur dem Wort *stuck*) auf.

Kontextsensitiver Thesaurus und kontextsensitive Rechtschreibkorrektur
Das im Abschnitt Übersetzung (s.o.) beschriebene Verfahren zur Ausnutzung des Kontextes kann auch eingesetzt werden, um aus den Vorschlägen eines Thesaurus das am Besten in den aktuellen Kontext passende Synonym auszuwählen. Ebenso kann aus den Korrekturvorschlägen eines Rechtschreibprogramms das Wort bestimmt werden, das am Besten zum Rest des Satzes passt. Textverarbeitungsprogramme können so intelligenter gemacht werden.

Lexikonerstellung, Aufbau von Thesauri und Ontologien
Bei der manuellen Erstellung von Lexika, Thesauri oder Ontologien kann mit DISCO enorm viel Zeit gespart werden. Aus den von DISCO zu einem Ausgangswort gelieferten Kollokationen und

semantisch ähnlichsten Wörtern werden einfach die gewünschten Beugungsformen, Schreibvarianten, Synonyme, Ober- und Unterbegriffe etc. ausgewählt und in das eigene Lexikon oder die Ontologie übernommen. Das zeitaufwändige Suchen nach sinnverwandten Wörtern wird so wesentlich verkürzt und die „Ausbeute“ steigt.

Verfügbare Sprachen

DISCO benötigt für jede Sprache einen Index mit Daten, ein sogenanntes **Sprachdatenpaket**. Momentan sind Sprachdatenpakete für Deutsch, Englisch, Französisch, Spanisch, Italienisch und Tschechisch verfügbar. Portugiesisch, Niederländisch, Polnisch, Russisch und Türkisch sind in Vorbereitung. Die aktuell verfügbaren Sprachen finden Sie auf der Seite http://www.linguatools.de/disco-download.html.

Verfahren

DISCO wäre vor 15 Jahren noch nicht machbar gewesen, da die nötigen Textmengen in elektronischer Form damals noch nicht verfügbar waren. Zur Berechnung der semantischen Ähnlichkeiten „liest“ DISCO enorme Mengen an Texten (in der Größenordnung von mehreren hundert Millionen laufenden Wörtern, was in etwa einem Dutzend Jahrgängen einer Tageszeitung oder tausend Büchern entspricht) und wertet sie mit eigens entwickelten statistischen Verfahren aus. Auf diese Weise kann der Gebrauch der Wörter bestimmt und miteinander verglichen werden. Die Idee dabei ist, dass Wörter, die ähnlich gebraucht werden, eine ähnliche Bedeutung haben.

Technische Daten

Bei DISCO handelt es sich um ein Java-Archiv. Die Sprachdatenpakete basieren auf dem Lucene[1]-Index. Sie können aufgrund der enthaltenen Datenmengen sehr groß werden. Die Größe der momentan verfügbaren Pakete (siehe http://www.linguatools.de/disco-download.html) bewegt sich zwischen 800 Megabyte und 8 Gigabyte.

Geschwindigkeit. Auf einem Athlon 64 mit 2,4 GHz und einer SATA-Festplatte berechnet DISCO die semantische Ähnlichkeit von ca. 50 Wortpaaren in einer Sekunde.

Unterstützte Betriebssysteme:	Windows Vista/XP/2000 Linux Solaris Mac OS X
Erforderliche Software:	Java Runtime Environment (JRE) ab Version 1.5
Hardware-Mindestanforderungen:	Pentium 4/Athlon XP oder vergleichbarer Prozessor mit 512 MB RAM und ausreichend Plattenkapazität für die Sprachdatenpakete.

1 http://lucene.apache.org

Weitere Informationen und Kontakt

DISCO-Homepage:	http://www.linguatools.de/disco.html
Liste der verfügbaren Sprachen:	http://www.linguatools.de/disco-download.html
DISCO FAQ:	http://www.linguatools.de/disco-faq.html
Dokumentation der Java-API:	http://www.linguatools.de/disco-api/

Für Fragen stehen wir Ihnen gerne zur Verfügung. Schreiben Sie einfach eine E-Mail an disco@linguatools.de.