



ConceptManager: Pflege multilingualer Ontologien im crosslingualen Retrieval

A. Jackson, M. Lewandowski, Gr. Thurmair, J. Zwickl

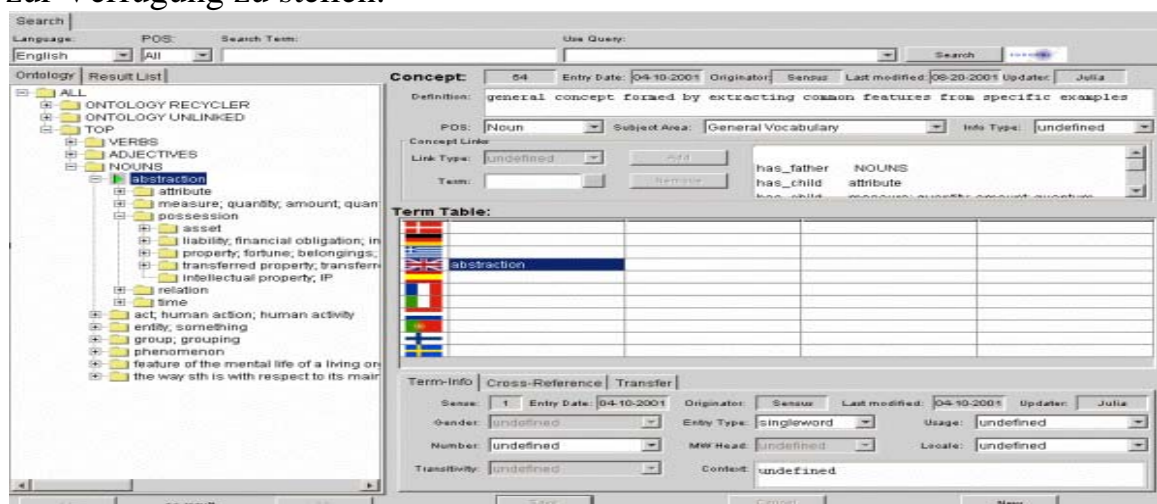
Compendium GmbH
Balanstr. 54
81541 München

Zusammenfassung

Der ConceptManager ist ein Tool zur Pflege multilingualer Ontologien, wie sie beim crosslingualen Retrieval benötigt werden. Er unterstützt die Erstellung von Begriffshierarchien und das Definieren von Begriffen; er unterstützt die Zuordnung von Termen zu diesen Begriffen in verschiedenen Sprachen; und er unterstützt die Zuordnung von linguistischen Annotationen zu diesen Termen, wie sie für die maschinelle Verarbeitung relevant sind.

1 Ontologien

Ontologien haben sich als nützliches Mittel bei der Suche entwickelt, weil sie eine Expansion der Suchbegriffe gestatten, die die Ergebnisse verbessern hilft. Trotz einiger Ansätze zur automatischen Erzeugung von Ontologien aus Korpora ist es in der Praxis erforderlich, eine interaktive Pflegekomponente zur Verfügung zu stellen.



Existierende Tools zur Pflege von Ontologien unterstützen üblicherweise den hierarchischen Aspekt, jedoch nicht den linguistischen Aspekt: Während Thesauri eine Normierung der Begrifflichkeit anstreben, muss im Retrieval die



Vielfalt der sprachlichen Realisierungen eines Begriffs erfasst und repräsentiert werden.

Der ConceptManager unterstützt die Erstellung von Hierarchiebäumen (*is_a*-Relationen) und die Verknüpfung der Begriffe mittels der bei EuroWordNet definierten Relationen. Die Begriffe, die die einzelnen Knoten bilden, werden durch Definition, Wortart und Fachgebiet näher beschrieben.

2 Terme, Multilingualität

Jeder Begriff kann durch Terme in verschiedenen Sprachen ausgedrückt werden. Der ConceptManager gestattet die Eingabe von Termen und ihrer Synonyme in den verschiedenen Sprachen; auf diese Weise entsteht ein multilinguales ConceptNet. Die Strategie dabei ist, eine universelle Ontologie mit mehrsprachigen Termen anzureichern, statt mehrere einzelsprachliche Ontologien einander zuzuordnen. Homographe Terme werden über ihre Beziehung zu den Begriffen disambiguiert.

3 Linguistische Beschreibung

Jeder Term einer Sprache kann um linguistische Annotationen angereichert werden: Wortart, Genus, Transitivität usw.; Querverweise wie „*ist_Abkürzung von*“; Übersetzungshinweise wie „*präferierte Übersetzung, Teiläquivalenz*“ können angegeben werden. Ziel ist es, das multilinguale ConceptNet möglichst vielen linguistischen Komponenten, von Indexing über Information Extraction bis zur maschinellen Übersetzung, zugänglich zu machen. Die linguistischen Features stellen ein Subset von *OLIF* dar, einem offenen Lexikonstandard, den der ConceptManager als Schnittstelle zu Fremdanwendungen bei Import und Export unterstützt.

4 Einbettung

Der ConceptManager ist Teil eines Workflows, in dem mittels Term-Extraktionsprogrammen aus einem repräsentativen Korpus die relevanten Suchterme gefunden werden. Sie werden in den ConceptManager importiert, dort interaktiv in ihre begrifflichen Relationen gebracht, ggf. um Übersetzungen angereichert und mit linguistischen Annotationen versehen. Das Ergebnis wird in die Datenstrukturen der weiterverarbeitenden Komponenten (z.B. einer Suchfragen-Expansions-Komponente) kompiliert. Das Tool ist erfolgreich zum Aufbau eines elfsprachigen ConceptNets mit ca. 15000 Begriffsknoten eingesetzt worden.