Ursula Reuther, Saarbrücken

Linguistisch basierte Extraktion von Termen und die Visualisierung terminologischer Relationen

Eine sorgfältige und umfassende Aufbereitung von terminologischem Wissen ist stets eine Voraussetzung für eine zielgerichtete, effiziente und leicht zugängliche Nutzung eines Fachwortschatzes. Dies gilt sowohl für den Bereich der grammatischen Fachterminologie als auch für jede andere Domäne, die über einen eigenen Fachwortschatz verfügt. Der Vortrag beschreibt hier fachgebietsunabhängig linguistisch basierte, automatische Verfahren und Werkzeuge, die sowohl bei der Erstellung einer Terminologie als auch bei deren Bereitstellung und Nutzung sowie zur Darstellung terminologischer Zusammenhänge sinnvoll eingesetzt werden können.

Zunächst wird beschrieben, wie man morphosyntaktische Analysemethoden der maschinellen Sprachverarbeitung nutzen kann, um eine deskriptive, systematische und fachgebietsbezogene Terminologie aufzubauen und wie auf Basis solch einer Analyse-komponente für das Deutsche aus einem Korpus terminologische Benennungen automatisch extrahiert werden können. Die verwendete Analysekomponente basiert auf MPRO (vgl. Maas/Rösener/Theofilidis 2009), einem Programm, das auf Basis eines morphologischen Lexikons mit ca. 140.000 Einträgen und eines Lemma-Lexikons mit ca. 755.000 Einträgen ein gegebenes Korpus verarbeitet und analysiert. Als Ergebnis werden Merkmalsbündel mit morphosyntaktischen (Flexion, Komposition und Derivation) und semantischen Informationen ausgegeben:

> {ori=Absorptionskälteanlage,lu=absorptionskälteanlage,c=noun,ehead={case=no
> m;gen;dat;acc,nb=sg,g=f,infl=weak;strong;null},s=anlage,lng=germ,w=3,cs=n#n#
> n,gs=f#f#f,t=absorption#kälte#anlage,ds=absorbieren~ation#kalt~e#anlage,ls=ab
> sorbieren#kalt#anlage,ss=process#state#loc,lngs=lat#germ#germ,compparts=abs
> orptionskälte;kälteanlage }

Diese Merkmale werden dann für die Bestimmung der Termhaftigkeit eines Worts herangezogen. So können folgende Kriterien u.a. als Hinweis auf die Termhaftigkeit eines Worts genutzt werden.

- Kompositum
- Simplex mit bestimmten (semantischen) Eigenschaften
- Fremdwörter
- Bindestrichkonstrukte
- Adjektiv-Nomen-Konstrukte
- Toponyme
- Namen

Im zweiten Teil des Vortrags wird aufgezeigt, wie vorgenannte morphosyntaktische Analyseverfahren kombiniert mit semantischen Techniken und mit probabilistischen Verfahren zur Gewinnung von Begriffsrelationen und deren Darstellung genutzt werden können.

Die zuvor extrahierten Benennungen werden unter Verwendung von statistischen Verfahren, wie z.B. der tf-idf-Gewichtung bewertet. Diese Gewichtung kann nur dann sinnvoll berechnet werden, wenn die extrahierten Terme in lemmatisierter Form vorliegen, was durch die MPRO-Analyse gewährleistet ist. Die Begriffsrelationen können dann in Form von Wortwolken visualisiert werden und zum einen für die Anzeige von Suchbegriffen und zur Query-Erweiterung in Information-Retrieval-Anwendungen genutzt werden, zum anderen aber auch zur Dismabiguierung von mehrdeutigen Benennungen (*Virus* – Computervirus vs. Virus als Krankheitserreger). Auf Basis der morphologischen Wortstruktur können zudem Ober- und Unterbegriffsrelationen erkannt werden (*Absorptionskälteanlage > Kälteanlage > Anlage*).

Literatur

Drewer, Petra (2016): Terminologiemanagement: Methodische Grundlagen. In: Hennig, Jörg/Tjarks-Sobhani, Marita (Hrsg.): Terminologiearbeit für Technische Dokumentation, Lübeck: Schmidt-Römhild (tekom-Schriften zur Technischen Kommunikation, 21), 50-62.

Gindiyeh, Mahmoud (2013): Anwendung wahrscheinlichkeitstheoretischer Methoden in der linguistischen Informationsverarbeitung, Berlin: Logos Verlag.

Maas, Heinz Dieter/Rösener, Christoph/Theofilidis, Axel (2009): Morphosyntactic and semantic analysis of text: The MPRO tagging procedure. In: Mahlow, Cerstin/Piotrowski, Michael (Hrsg.): State of the art in computational morphology. Workshop on systems and frameworks for computational morphology, SFCM 2009, Zurich, Switzerland, September 4, 2009. Proceedings. 1st ed. New York: Springer (Communications in computer and information science, 41), 76–87.