Vergleichende Evaluierung von Verfahren zur Extraktion monolingualer Termkandidaten aus deutschen Texten

Tanja George*, Johannes Schäfer*, Ina Rösiger*, Ulrich Heid•, Michael Dorna◇

*Universität Stuttgart, •Universität Hildesheim, ◇Robert Bosch GmbH

Die Evaluierung von Ergebnissen der automatischen Termextraktion gestaltet sich als schwierig, da oftmals keine Goldstandard-Daten als Vergleichsmaterial verfügbar sind. Es ist auch nicht trivial, einen brauchbaren Goldstandard für Terme einer Domäne zu erstellen, da sich die Grenzen eines gegebenen Fachgebiets selbst durch Expertenmeinungen meist nicht genau spezifizieren lassen.
Wir berichten über Studien zum Vergleich verschiedener Ansätze zur Extraktion monolingualer deutscher Termkandidaten mit einem Schwerpunkt auf Mehrwort-Termen. Wir analysieren Texte aus dem Bereich des Heimwerkens (darunter ein Heimwerker-Handbuch, Forenbeiträge, FAQs, "Tipps and Tricks" für Heimwerker, ein Heimwerkerlexikon und Marketingtexte eines Unternehmens), aus denen ein Goldstandard abgeleitet wurde. Wir vergleichen verschiedene Stufen des Verhältnisses von Noise und Silence bei SDL Trados MultiTerm Extract (sog. Qualitätsfilter-Stufen) und verschiedene Varianten eines im EU-Projekt TTC ((Gojun et al., 2012), (Gojun & Heid, 2012), TTC Website[1]) entwickelten Forschungsprototypen. Dabei zeigen wir Ergebnisse der Nutzung verschiedener statistischer Ansätze zur Bestimmung von Mehrwort-Termen (C-value, Assoziationsmaße, statistische Tests (Pazienza et al., 2005)) und von Experimenten zur Extraktion aus dependenzgeparstem Text (Mate-Tools (Björkelund et al., 2010), evtl. weitere).
Für den Goldstandard wurden verschiedene POS-Sequenz-Muster der Frequenz $f > 3$ aus einem Korpus mit 2.689.383 Wörtern extrahiert; dies ergab insgesamt 18.611 Kandidaten. Diese Termkandidaten wurden anhand von vorab fixierten Richtlinien dreifach parallel annotiert. Aufnahme in den Goldstandard im engeren Sinne fanden Items, die von allen drei Annotatoren als terminologisch relevant angesehen wurden. In einem "liberaleren" Goldstandard sind auch Items enthalten, für die es eine 2:1-Mehrheit gab.
Das Poster stellt die Verfahren zur Entwicklung des Goldstandards (Richtlinien, Details zum Inter-Annotator-Agreement etc.), die untersuchten Extraktionsmethoden, die datenbankbasierte Evaluierungsinfrastruktur, die Ergebnisse und Beispiele aus der Fehleranalyse vor.

Literatur

Björkelund, A., B. Bohnet, L. Hafdell & P. Nugues. 2010. A high-performance syntactic and semantic dependency parser. demo session. In *Proceedings of the 23rd international conference on computational linguistics: Demonstrations (coling '10)*, 33 – 36.

Gojun, A. & U. Heid. 2012. Term candidate extraction for terminography and cat: and overview of ttc. In *Proceedings of the 15th euralex international congress*, 585 – 594.

Gojun, A., U. Heid, B. Weissbach, C. Loth & I. Mingers. 2012. Adapting and evaluating a generic term extraction tool. In *Proceedings of the eight international conference on language resources and evaluation (lrec 2012)*, 651 – 656.

Pazienza, M.T., M. Pennacchiotti & F.M. Zanzotto. 2005. Terminology extraction: An analysis of linguistic and statistical approaches, vol. 185, 255 – 279.

[1] www.ttc-project.eu