FV 5 Terminologieextraktion – multilingual, semantisch und mehrfach verwendbar

Melanie Siegel, Acrolinx GmbH, Berlin
Prof. Dr. Petra Drewer, Hochschule Karlsruhe

Grundlagen

Will man die Terminologie eines Unternehmens standardisieren und kontrolliert verwenden, so sind dazu – grob gesagt – folgende Arbeitsschritte erforderlich:

1. Sammlung von Terminologie
2. Systematisierung von Terminologie
3. Standardisierung von Terminologie
4. Kontrolle der Terminologieverwendung

Alle Arbeitsschritte können von geeigneten Tools unterstützt und begleitet werden. Von großer Bedeutung ist natürlich das System zur Terminologieverwaltung, also zur begriffsorientierten Speicherung und Weiterverarbeitung der terminologischen Daten. Doch auch andere Arbeitsschritte lassen sich automatisieren oder zumindest maschinell unterstützen. Der Vortrag widmet sich daher dem oben genannten Schritt 1 und betrachtet insbesondere die Möglichkeiten einer automatisierten Termextraktion.

Die Sammlung von Terminologie erfolgt auf verschiedenen Wegen: Einerseits müssen schon vorhandene Terminologiebestände im Unternehmen „aufgespürt" und verarbeitet werden. Dabei kann es sich um reine Wortlisten, um Glossare, um Vokabellisten, um Abkürzungsverzeichnisse und Ähnliches handeln. Andererseits reichen diese vorhandenen Bestände im Regelfall nicht aus, so dass weitere Quellen für Terminologie ausgewertet werden müssen. Bei diesen Quellen handelt es sich in erster Linie um bereits erstellte, ggf. schon übersetzte Technische Dokumentationen. Aber auch Translation Memorys sind eine wichtige Wissensquelle für multilinguale Terminologie.

Die folgenden Abschnitte erläutern, wie aus diesen Quellen automatisch ein- und mehrsprachig Terminologie extrahiert werden kann.

Extraktion von Terminologie in einer Sprache – linguistisch basiert

Die Texte werden zunächst mit linguistischen Informationen wie Tokenisierung, Morphologie und syntaktische Kategorien angereichert, dann werden Benennungen gezielt extrahiert: z. B. Nomen, Komposita, Akronyme, Produktnamen. Zusätzliche Information wird bei der Extraktion berücksichtigt, wie z. B. Information über mögliche Rechtschreibfehler, unbekannte Wörter, die Häufigkeit des Wortes in den Texten und auch Wörter, die schon in der Terminologiedatenbank stehen und daher nicht mehr extrahiert werden. Die Extraktion gibt Informationen, die für den Validierungsprozess wichtig sind, an den Anwender, wie Satzkontexte und Termvarianten, die in den Texten auftreten. Termvarianten können dabei Varianten in der Zeichenkette oder linguistische Varianten sein. Die extrahierten Benennungen werden im Anschluss von den Terminologie-Experten validiert und freigegeben.

Das Ergebnis einer solchen Termextraktion zeigt Abb. 1.

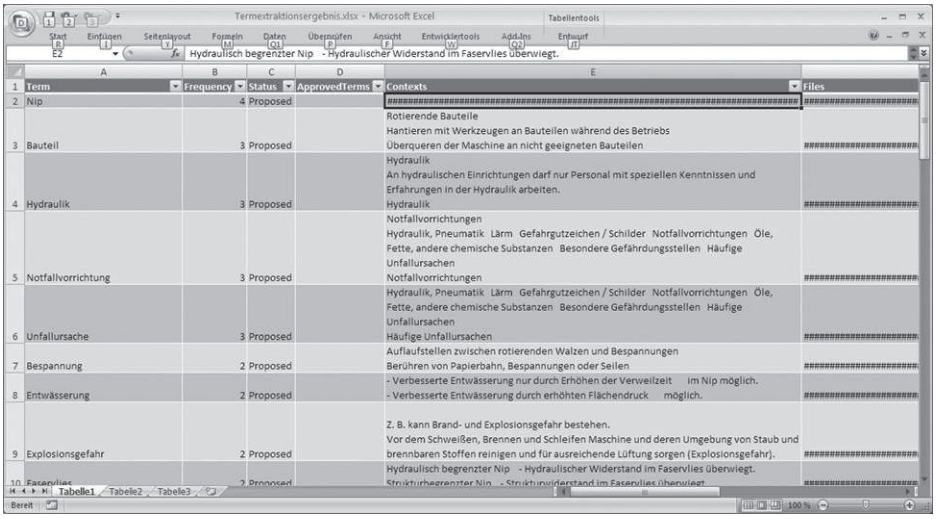

Abb. 1

Multilinguale Termextraktion – linguistisch basiert

Die multilinguale Termextraktion basiert auf der oben beschriebenen Termextraktion mit linguistischen Mitteln. Es handelt sich hier aber um ein statistisches Verfahren, das auch mit weniger linguistischen Informationen über die Zielsprache möglich ist. Wenn eine der beteiligten Sprachen eine bekannte Sprache – Englisch, Deutsch, Französisch, Chinesisch oder Japanisch – ist, dann können die anderen Sprachen beliebig sein. Es ist möglich, Benennungen in zwei, drei oder auch mehr Sprachen gleichzeitig zu extrahieren.

Die Ausgangsbasis für die multilinguale Termextraktion sind häufig Translation-Memory-Dateien im TMX-Format, wie sie in vielen Unternehmen vorliegen. Es ist aber auch möglich, mit parallelen Textdateien zu arbeiten, bei denen die Sätze der Sprachen einander zugeordnet sind.

Wenn man nur parallele, also übersetzte Dokumente zur Verfügung hat, muss eine automatische Alignierung der Sätze vorgeschaltet werden. Allerdings müssen diese Dokumente sehr ähnliche Textstrukturen aufweisen, damit die übersetzten Sätze auch gefunden werden können.

Die linguistische Maschine analysiert die Sätze in der bekannten Sprache und extrahiert Terminologie. Dann wird die so extrahierte Terminologie mit den Benennungen der Zielsprache in Beziehung gesetzt. Dafür werden statistische Verfahren und zusätzliche linguistische Informationen genutzt, je nachdem wie viele Informationen über diese Sprache zur Verfügung stehen.

Das Ergebnis ist eine Excel-Tabelle mit Benennungen in den verschiedenen Sprachen, Frequenzinformationen und Informationen über die Satzkontexte (siehe Abb. 2).

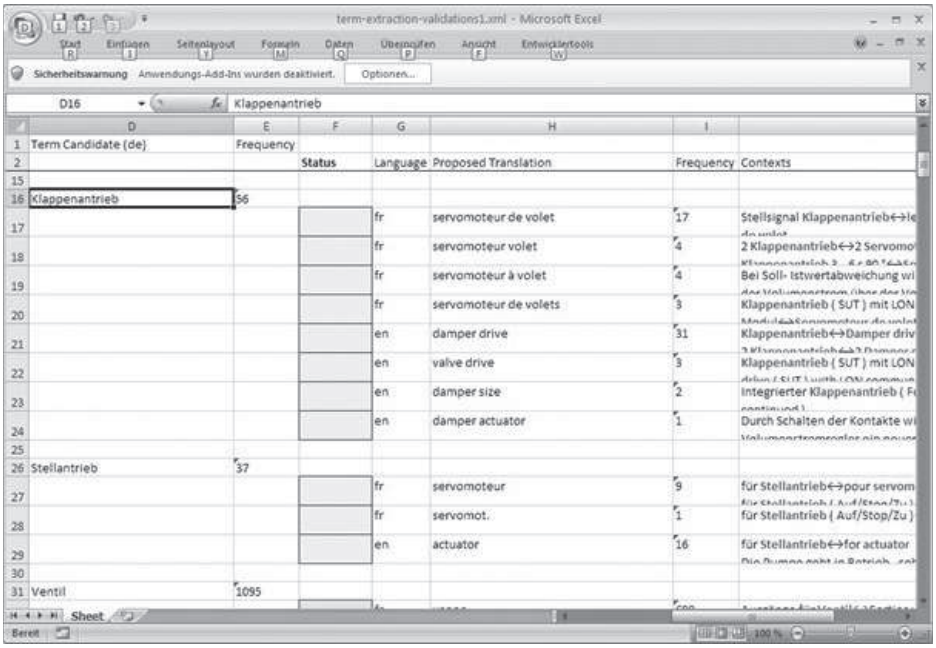

Abb. 2

Semantische Information aus der multilingualen Termextraktion

Diese Ergebnistabelle liefert spannende Informationen über die Inhalte der Dokumente in allen beteiligten Sprachen. Einige Beispiele für Benennungen, die in deutsch-englischen Translation Memorys gefunden wurden:

Grundstellung:
- starting position
- basic position
- home position
- basic setting
- initial position
- starting pos.
- normal position

adaptor:

- Zwischenstück
- Adaptor

operating unit:

- Bedieneinheit
- Bediengerät

Die Extraktion enthält Informationen über Varianten und Synonyme in allen beteiligten Sprachen. Aus den Beispielen kann man nicht nur schließen, dass es unterschiedliche deutsche Äquivalente für „operating unit" im Translation Memory gibt, sondern man kann auch feststellen, dass es eine semantische Beziehung zwischen „Bedieneinheit" und „Bediengerät" geben muss.

Arbeitsschritte im Anschluss an die Termextraktion

Im Anschluss an die Termextraktion müssen die terminologischen Daten validiert, systematisiert und weiterverarbeitet werden.

Dazu gehören u. a. folgende Tätigkeiten:

- Auswahl der relevanten Termini aus der Gesamtliste
- Definieren der relevanten Begriffe
  Durch das Definieren werden einerseits Begriffe geklärt und abgegrenzt, andererseits Synonyme erkannt. Weitere Synonyme sind bereits durch die Termextraktion erkennbar: einsprachig in Form von Termvarianten und mehrsprachig durch die verschiedenen zielsprachlichen Äquivalente.
- Festlegung von Vorzugsbenennungen
  Immer wenn Synonyme auftreten, muss festgelegt werden, welche Benennungen fortan verwendet und welche verboten werden sollen.

Diese zentralen Tätigkeiten, die im Anschluss an die Terminologiegewinnung oder -extraktion erfolgen, können nicht automatisiert werden, sondern bleiben rein menschliche Aufgaben. Sie bilden jedoch die Basis für eine anschließende kontrollierte, standardisierte Verwendung der Unternehmensterminologie.

für Rückfragen:
melanie.siegel@acrolinx.com
petra.drewer@hs-karlsruhe.de