Verfahren zur Terminologieprüfung im Vergleich

Eine maschinelle Terminologieprüfung unterstützt Autoren bei der korrekten Verwendung ihrer Unternehmensterminologie. Das vorliegende Dokument beschreibt gängige Verfahren zur Terminologieprüfung und beleuchtet sowie vergleicht die Vorteile und Nachteile der einzelnen Verfahren.

Inhalt

1 Einleitung

Die korrekte und einheitliche Verwendung von Benennungen bietet zahlreiche Vorteile für Unternehmen. Eine maschinelle Terminologieprüfung unterstützt die Autoren bei der korrekten Verwendung von Terminologie.

Auf technischer Ebene geht es bei der Terminologieprüfung darum, die in einem Text verwendeten Benennungen mit der Terminologiedatenbank abzugleichen. Der Benutzer soll insbesondere dann informiert werden, wenn Negativbenennungen verwendet wurden.

Es gibt verschiedene Ansätze zur Terminologieprüfung. Im Rahmen dieses Whitepapers werden drei Methoden beschrieben und verglichen:

- Fuzzy-Matching
- Stemming
- Linguistische Terminologieprüfung

Dabei geht es v. a. um die Vorteile und Nachteile der drei Methoden.

2 Übliche Verfahren für die Terminologieprüfung

2.1 Fuzzy-Matching

Fuzzy-Matching-Mechanismen sind über viele freie Software-Bibliotheken verfügbar. Daher ist es nicht verwunderlich, dass das Verfahren in vielen Translation Memory Systemen und Redaktionssystemen zum Einsatz kommt.

Die grundlegende Funktionsweise des Fuzzy-Matchings besteht darin, dass die Ähnlichkeit von zwei Strings auf Zeichenbasis bestimmt wird. Für die Ähnlichkeit wird ein bestimmter Schwellenwert festgelegt. Mit einem niedrigen Schwellenwert erhält man viele Treffer, dafür auch viele Falschmeldungen. Mit einem hohen Schwellenwert reduziert man die Falschmeldungen, bekommt aber auch weniger Treffer.

Im Detail gibt es viele Varianten des Fuzzy-Matchings, die sich v. a. im Messen der Ähnlichkeit von Strings unterscheiden.

Es kommen keine Kenntnisse über Sprache zum Tragen. Dies hat den Vorteil, dass Fuzzy-Matching für alle Sprachen funktioniert.

Eine Besonderheit: Fuzzy-Matching ist robust gegen Tippfehler. Deshalb verwenden auch viele Rechtschreibprüfungen Fuzzy-Matching.

Beispiele[1]:

- Förderwerk**e** ↔ Förderwerk ⇒ 95.65 % Ähnlichkeit und damit wahrscheinlich ein Treffer
- Baum ↔ B**ä**um**e** ⇒ 60 % Ähnlichkeit und damit wahrscheinlich kein Treffer

2.2 Stemming

Die Grundidee hinter Stemming ist, dass Wörter auf einen Wortstamm zurückgeführt und die Wortstämme miteinander verglichen werden. Ein Stemming-Verfahren muss für jede Sprache entwickelt werden. Dabei ist zu beachten, dass Stemming bei manchen Sprachen besser, bei anderen schlechter bzw. gar nicht funktioniert. Für viele Sprachen sind bereits Software-Bibliotheken vorhanden; für bestimmte Stemmer sind sogar die Algorithmen frei zugänglich, z. B. der Algorithmus des bekannten Snowball Stemmers[2].

Beispiele[3]:

- Vergleich Änderungen ↔ Änderung
  - Großbuchstaben normalisieren ⇒ änderungen ↔ änderung
  - Umlaute normalisieren ⇒ anderungen ↔ anderung
  - Endung "en" entfernen ("stemmen") ⇒ anderung ↔ anderung
  - Endung "ung" entfernen ("stemmen") ⇒ ander ↔ ander

Beide Wörter werden als Matches erkannt. "Sicherung" und "Sicherheit" werden auf "sich" abgebildet und gelten gegenseitig als Treffer.

1 Zum Selbst-Ausprobieren des Verfahrens siehe: https://www.tools4noobs.com/online_tools/string_similarity/ (letzter Abruf: 13.11.2017, 13.56 Uhr)

2 http://snowball.tartarus.org/algorithms/german/stemmer.html (letzter Abruf: 13.11.2017, 14.11 Uhr)

3 Ergebnisse auf Basis dieses Tools: http://text-processing.com/demo/stem/ (letzter Abruf: 13.11.2017, 14.15 Uhr)

2.3 Linguistische Terminologieprüfung

Die Entwicklung einer linguistischen Terminologieprüfung erfordert tiefes Verständnis der jeweiligen Sprache. Im Zuge dieses Verfahrens wird Sprache in Morpheme (kleinste bedeutungstragende Einheiten) zerlegt und dann einer linguistischen Analyse unter Beachtung der Wortbildungsregeln unterzogen.

Beispiel 1: < Erkennung Fuge: Ölstandsanzeige ↔ Ölstandanzeige>
Beispiel 2: < Erkennung Ableitung: Bedienhebel ↔ Bedienungshebel >
Beispiel 3: < Syntaktische Varianten: Belüftungsfilter ↔ Filter für Belüftung >

Eine besondere Stärke zeigt das Verfahren der linguistischen Terminologieprüfung bei der Erkennung von Synonymen. Wenn grundlegende Synonymbeziehungen hinterlegt sind, wie etwa Gerät ↔ Maschine, kann z. B. auch Auswuchtgerät als Synonym zu Auswuchtmaschine erkannt werden.

2.4 Reguläre Ausdrücke

„Ein Regulärer Ausdruck (engl. regular expression, Abk. RegExp oder Regex) ist eine Zeichenkette, die der Beschreibung von Mengen beziehungsweise Untermengen von Zeichenketten mit Hilfe bestimmter syntaktischer Regeln dient. Reguläre Ausdrücke finden vor allem in der Softwareentwicklung Verwendung; für fast alle Programmiersprachen existieren Implementierungen.“[4]

Reguläre Ausdrücke ähneln einer Programmiersprache und so gibt es viele Möglichkeiten, eine Terminologieprüfung umzusetzen.

Beispiel:

- Explizite Listung: '[Baum|Bäume|Baume|Baums]
- B[alä]um[els+] ⇒ Würde aber auch "Bäums" erkennen

Das Beispiel zeigt, dass sich mit regulären Ausdrücken auch Texte auf Terminologie prüfen lassen. Dies erfordert aber viel Arbeit, Knowhow und führt zu einer enormen Komplexität.

Letztlich stellen reguläre Ausdrücke einen schwer vergleichbaren Sonderfall dar. Da sie sich abgesehen von speziellen Szenarien oder als Ergänzung nicht etabliert haben, wird dieser Ansatz nicht im Vergleich berücksichtigt.

4 http://www.regexe.de/hilfe.jsp (letzter Abruf: 13.11.2017, 14.25 Uhr)

3 Der Vergleich

In diesem Kapitel werden die Verfahren miteinander verglichen, die im vorherigen Kapitel vorgestellt wurden. Ausgenommen sind die regulären Ausdrücke zur Terminologieprüfung. Der Vergleich bezieht sich auf die folgenden Kategorien:

- **Rechtschreibfehler:** Wie robust ist das Verfahren gegenüber falsch geschriebenen Termen?
- **Flexion:** Erkennt das Verfahren auch grammatische Abwandlungen (z. B. in Kasus, Numerus und Genus)?
- **Bindestrichvariante:** Erkennt das Verfahren auch Bindestrichvarianten eines Terme?
- **Fugenvariante:** Erkennt das Verfahren auch Fugenvarianten eines Terme?
- **Ableitungsvariante:** Erkennt das Verfahren auch Ableitungsvarianten eines Terme?
- **Syntaktische Variante:** Erkennt das Verfahren auch syntaktische Varianten eines Terme?
- **Synonyme:** Erkennt das Verfahren auch Synonyme eines Terme?
- **Falschmeldungen:** Wie wahrscheinlich ist es, dass ein Term erkannt wird, der gar keiner ist?
- **Verfügbare Sprachen:** Für welche Sprachen ist das Verfahren verfügbar?
- **Kosten:** Welche Kosten bringt das Verfahren mit sich?

Erfüllungsgrad:

- kein Stern: nicht erfüllt
- ★: unzureichend erfüllt
- ★★: teilweise erfüllt
- ★★★: erfüllt

3.1 Rechtschreibfehler

Verfahren	Bewertung	Erfüllungsgrad
Fuzzy-Matching	Normalerweise sind Strings trotz Rechtschreibfehler noch sehr ähnlich. Letztlich hängt die Robustheit gegenüber Rechtschreibfehlern von der Länge des Worts ab. Bei kurzen Wörtern kann ein falscher Buchstabe schnell zu einer geringen Ähnlichkeit führen.	★★
Stemming	Terme mit Rechtschreibfehlern werden meist nicht erkannt. Dies wird aber in der Regel durch eine Rechtschreibprüfung mit Benutzerwörterbuch abgefangen.	★
Linguistische Terminologieprüfung	Terme mit Rechtschreibfehlern werden in der Regel nicht erkannt. Dies wird aber in der Regel durch eine Rechtschreibprüfung mit Benutzerwörterbuch abgefangen.	★

3.2 Flexion

Unter Flexion versteht man die grammatische Abwandlung eines Worts, z. B. in Hinblick auf Kasus, Numerus und Genus.

Verfahren	Bewertung	Erfüllungs-grad
Fuzzy-Matching	Je nach Wortlänge und Flexion des jeweiligen Worts und der Sprache. » Förderwerk und Förderwerke: Die flektierte Form wird erkannt (95 % Ähnlichkeit) » Ei und Eier: Die flektierte Form wird wahrscheinlich nicht erkannt (57 % Ähnlichkeit)	★
Stemming	Das Stemming weiß in der Regel, wie Wörter flektiert werden und erkennt entsprechende Terme zuverlässig. Probleme können jedoch bei Fremdwörtern auftreten. Lexikon ↔ Lexika würde z. B. nicht erkannt.	★★
Linguistische Terminologie-prüfung	Die Linguistische Terminologieprüfung weiß, wie Wörter flektiert werden und erkennt entsprechende Terme zuverlässig.	★★★

3.3 Bindestrichvarianten

Verfahren	Bewertung	Erfüllungs-grad
Fuzzy-Matching	Beispiel: Kabelfernsteuerung ↔ Kabel-Fernsteuerung Es hängt von der jeweiligen Implementierung ab, ob der Bindestrich beispielsweise als Worttrennung interpretiert wird. Es kann aber immer je nach Wortlänge und Schwellenwert zu Fehlern kommen.	★★
Stemming	Normalerweise berücksichtigen Stemmingverfahren keine Bindestrichvarianten.	
Linguistische Terminologie-prüfung	Die Linguistische Terminologieprüfung erkennt Bindestrichvarianten über Wortbildungsregeln.	★★★

3.4 Fugenvariante

Eine Fugenvariante liegt vor, wenn ein Wort sich nur hinsichtlich eines Fugenlauts von einem Term unterscheidet.

Verfahren	Bewertung	Erfüllungs-grad
Fuzzy-Matching	Fugenvarianten führen meist immer noch zu einer hohen Ähnlichkeit der Strings.	★★★
Stemming	Stemming scheitert an Fugenvarianten, da das Stemming bereits vor der Fuge endet: Ölstandsanzeige ↔ Ölstandanzeige ⇒ olstands-anzeig != olstandanzeig	
Linguistische Terminologie-prüfung	Die Linguistische Terminologieprüfung erkennt Fugenvarianten über Wortbildungsregeln.	★★★

3.5 Ableitungsvariante

Eine Ableitungsvariante verfügt bei gleichem Wortstamm über eine andere Ableitung als der ursprüngliche Term.

Verfahren	Bewertung	Erfüllungs-grad
Fuzzy-Matching	Ableitungsvarianten verändern das Wort so sehr, dass die Ähnlichkeit in der Regel unter den Schwellenwert fällt. Beispiel: Bedienhebel ↔ Bedienungshebel ⇒ 85 % Ähnlichkeit	★
Stemming	Stemming scheitert an der Ableitungsvarian-te, da das Stemming bereits vor der Ableitung endet: Bedienhebel ↔ Bedienungshebel ⇒ bedienhe-bel != bedienungshebel	
Linguistische Terminologie-prüfung	Die Linguistische Terminologieprüfung erkennt Ableitungsvarianten über Wortbildungsregeln.	★★★

3.6 Syntaktische Variante

Eine syntaktische Variante verfügt über eine abweichende Syntax, z. B. Bindestrich-Kompositum vs. Wortgruppe.

Verfahren	Bewertung	Erfüllungs-grad
Fuzzy-Matching	Für syntaktische Varianten erkennt das Fuzzy-Matching meist gar keine Ähnlichkeit Belüftungsfilter ↔ Filter für Belüftung ⇒ 51 % Ähnlichkeit	
Stemming	Stemming scheitert an der syntaktischen Variante: Belüftungsfilter ↔ Filter für Belüftung ⇒ be-luftungsfilt ↔ filt fur beluft	
Linguistische Terminologie-prüfung	Die Linguistische Terminologieprüfung erkennt syntaktische Varianten über Wortbildungsregeln.	★★★

3.7 Synonyme

Synonyme sind unterschiedliche Benennungen für einen Begriff.

Verfahren	Bewertung	Erfüllungs-grad
Fuzzy-Matching	Synonyme stellen gänzlich andere Worte dar, sodass das Fuzzy-Matching keine Ähnlichkeit erkennt.	
Stemming	Synonyme stellen gänzlich andere Worte dar, sodass auch der Wortstamm ein anderer ist.	
Linguistische Terminologie-prüfung	Standardmäßig oder kundenspezifisch sind Synonyme wie „Rad“ ↔ „Reifen“ oder „Gerät“ ↔ „Maschine“ hinterlegt. Damit kann automatisch „Radauswuchtma-schine“ auf „Reifenauswuchtgerät“ abgebildet werden.	★★

3.8 Falschmeldungen

Verfahren	Bewertung	Erfüllungs-grad
Fuzzy-Matching	Hohes Risiko, dass die Terminologieprüfung ein ähnliches geschriebenes, aber gänzlich anderes Wort erkennt. Bauhäuser ↔ Brauhäuser ⇒ 95 % Ähnlichkeit	★
Stemming	Wörter mit unterschiedlicher Bedeutung können den gleichen Wortstamm haben: Sicherheit ⇒ sich Sicherung ⇒ sich	★
Linguistische Terminologie-prüfung	Falschmeldungen sind sehr selten und resultieren oftmals aus den Grenzen maschineller Sprachverarbeitung, z. B. wenn Weltwissen erforderlich ist, über das das Verfahren nicht verfügt.	★★★

3.9 Verfügbare Sprachen

Verfahren	Bewertung	Erfüllungs-grad
Fuzzy-Matching	Lässt sich letztlich auf alle Sprachen anwenden.	★★★
Stemming	Viele europäische Sprachen. Unterschiedliche Qualität.	★★
Linguistische Terminologie-prüfung	Weniger Sprachen, da sehr aufwändig. DE, EN, FR werden aber von den meisten Lösungen abgedeckt.	★

3.10 Kosten

Verfahren	Bewertung	Erfüllungs-grad
Fuzzy-Matching	Es gibt kostenfreie Lösungen	★★★
Stemming	Es gibt kostenfreie Lösungen	★★★
Linguistische Terminologie-prüfung	Fast nur über kommerzielle Anbieter	★

4 Zusammenfassung

	Fuzzy-Matching	Stemming	Linguistische Terminologieprüfung
Robustheit gegen Rechtschreibfehler	★★	★	★
Flexion	★	★★	★★★
Bindestrichvarianten	★★		★★★
Fugenvarianten	★★★		★★★
Ableitungsvarianten	★		★★★
Syntaktische Varianten			★★★
Synonyme			★★
Falschmeldungen	★	★	★★★
Verfügbare Sprachen	★★★	★★	★
Kosten	★★★	★★★	★

5 Fazit

Fuzzy-Matching und Stemming sind Ansätze, die sich v. a. durch überschaubare Kosten und eine hohe Sprachabdeckung auszeichnen. Es gibt bereits fertige Software-Bibliotheken, die übernommen werden können. Manche Lösungen sind sogar kostenfrei erhältlich. Die Entwicklung einer linguistischen Terminologieprüfung erfordert einen nicht zu vernachlässigenden Forschungs- und Entwicklungsaufwand. Daraus resultiert, dass entsprechende Produkte ihren Preis haben und für vergleichsweise wenige Sprachen erhältlich sind. Dafür bieten sie ein Maß an Qualität, besonders in der Erkennung diverser Termvarianten, an das andere Ansätze nicht herankommen.

Der Aufwand, Negativbenennungen explizit zu erfassen, ist bei preiswerten Verfahren wesentlich höher. Um von Fuzzy-Matching- oder Stemming-Verfahren erkannt zu werden, müssen alle Varianten einer Negativbenennung der Software explizit bekanntgemacht werden. Eine linguistische Terminologieprüfung erkennt automatisch auch linguistische Varianten hinterlegter Negativbenennungen.

Letztendlich ist die Akzeptanz der Anwender das Entscheidende für den Einsatz eines Verfahrens. Hierbei gilt es, dass das Verfahren möglichst viel Effizienz bietet und dabei den Anwender möglichst wenig stört. Eine linguistische Terminologieprüfung kann in beiden Aspekten überzeugen. Sie deckt

viele Arten von Termvarianten ab, ohne dass alles explizit hinterlegt werden muss, was das Arbeiten durch weniger Hinterlegungsnotwendigkeit effizienter macht. Dabei können sogar Varianten von Negativbenennungen maschinell ermittelt werden, ohne dass der Terminologieverantwortliche jede einzelne Variante kennen muss.

Nicht zuletzt generiert eine linguistische Terminologieprüfung weniger Falschmeldungen als Fuzzy-Matching oder Stemming – der Anwender kann somit tendenziell störungsfreier arbeiten.

Über uns:

Konsistenz, Verständlichkeit und übersetzungsgerechtes Schreiben sind die Schwerpunkte unserer Autorenunterstützung. Doch dies allein genügt uns nicht: Mit dem gebündelten Wissen aus Forschung und Praxis entwickeln wir Produkte, die auf den ersten Blick begeistern sollen. Diesen Anspruch verfolgen wir mit einem kompetenten Team, das weiß, worauf es bei anwenderfreundlicher Software ankommt.

Congree Language Technologies GmbH
Im Stoeckmaedle 13
76307 Karlsbad
www.congree.com

congree