Terminologie extrahieren mit

extraTerm - en

Linguistic Content AG

Termextraktion Englisch

Hinweise zum Extraktionsverfahren
Erläuterung der Ergebnisse

Stand: 13. März 2017

1. Einführung

Mit **extraTerm** gehen Sie den ersten Schritt in Richtung Terminologieaufbau oder Glossarerstellung, indem Sie aus Ihren Dokumentbeständen alle Termkandidaten extrahieren und mit linguistischen Merkmalen sowie statistischen Kennzahlen auszeichnen lassen.

In einem erweiterten Anwendungsszenario können Sie **extraTerm** auch nutzen, um eine **Bestandsterminologie** um neue Terme zu erweitern. In diesem Fall wird bei der Termextraktion die Bestandsterminologie berücksichtigt und die darin erfassten Terme werden separat ausgezeichnet. Auch können den Termextraktionsergebnissen gemäß Ihren speziellen Anforderungen **Metadaten** zugeordnet werden (z. B. Dokumentart, Baugruppe, Abteilung).

Sprechen Sie uns an, wenn Sie spezielle Anforderungen an die Termextraktion mit **extraTerm** haben: **kontakt@iailc.de**.

Bei einer Termextraktion mit **extraTerm** werden Ihnen standardmäßig vier Ergebnisdateien bereitgestellt:

- eine tabellarische Aufbereitung der extrahierten Terme in Form einer Excel-Tabelle
- eine Auflistung aller ermittelten Vorkommenskontexte pro Termkandidaten
- eine Auflistung aller Belegdokumente pro Termkandidaten
- eine Datei mit grundlegenden Kennzahlen zur durchgeführten Termextraktion.

Struktur und Inhalte der vier Ergebnisdateien werden in Kapitel 4 näher erläutert. In Kapitel 2 gehen wir auf die Aufbereitung der Eingabedaten ein. In Kapitel 3 geben wir Hintergrundinformationen zu den Verfahren, die bei der Termextraktion mit **extraTerm** zur Anwendung kommen.

2. Datenaufbereitung

Bei Durchführung einer Termextraktion mit **extraTerm** werden nur die Textanteile Ihrer Dokumente berücksichtigt – und berechnet! Auszeichnungselemente, Grafiken und dergleichen werden vorab ausgeblendet.

Die in einer Datei enthaltene Textmenge kann somit deutlich geringer ausfallen als die für die Datei angegebene Dateigröße. Durch den Einsatz von Datenkompressionsverfahren kann aber auch umgekehrt bei bestimmten Dateiformaten die Textmenge größer ausfallen als die angegebene Dateigröße.

Wenn Sie für eine Termextraktion mit **extraTerm** unseren Webshop nutzen, wird Ihnen nach dem Hochladen Ihrer Dateien die berechnete Textmenge pro Datei angezeigt.

Einen Näherungswert für die Textmenge, die beispielsweise in einer MS-Word-Datei enthalten ist, erhalten Sie, wenn Sie die Anzahl der „Zeichen (mit Leerzeichen)“ (Menüoption *Überprüfen* > *Wörter zählen*) ermitteln.

Zur Extraktion der Daten und Metadaten aus unterschiedlichen Dateitypen wird das Apache Tika™ Toolkit eingesetzt. Von den Dateiformaten, die das Apache Tika™ Toolkit unterstützt, sind folgende Formate erfahrungsgemäß für eine Termextraktion mit **extraTerm** geeignet:

Format	**Anmerkungen**
html, xml	
doc, docx, xls, xlsx, ppt, pptx	Microsoft Office
odt, ods	OpenOffice, LibreOffice
epub	

Wenn Ihre Daten in einem anderen Format vorliegen, prüfen Sie, ob ein verlustfreier Export in eines der von **extraTerm** unterstützten Formate möglich ist, oder kontaktieren Sie uns mit einer entsprechenden Anfrage: **kontakt@iailc.de**.

Wenn Sie uns Beispieldaten zukommen lassen wollen, prüfen wir gern, ob das jeweilige Format von Apache Tika™ unterstützt wird und die extrahierten Daten verwertbar sind.

Folgende Formate werden von **extraTerm** mit Einschränkungen unterstützt:

Format	**Anmerkungen**
txt	Interpretation als Fließtext; s.u.
pdf	nicht immer verlustfreie Extraktion der Textanteile; s.u.

Bei **TXT-Dateien** werden aufeinanderfolgende Textzeilen als Fließtext interpretiert. So werden beispielsweise Überschriften, die nicht durch eine Leerzeile abgesetzt sind, mit unmittelbar folgenden Textzeilen zusammengezogen. Sofern die Überschriften nicht durch ein Satzendezeichen abgeschlossen sind, ist für die linguistische Analyse in der Folge nicht mehr ersichtlich, dass unter-

schiedliche Struktureinheiten vorliegen. Dies kann zu Fehlanalysen führen. Der gleiche Effekt wird auch bei Inhaltsverzeichnissen, Listen, Tabellen und ähnlichen Textelementen auftreten, in denen die einzelnen Strukturelemente weder durch Leerzeichen noch durch Satzendezeichen voneinander abgesetzt sind.

Aus **PDF-Dateien** lassen sich Textanteile in der Regel nicht völlig verlustfrei extrahieren. So werden Wortformen an Silbentrennpositionen oftmals auseinandergerissen, sodass die einzelnen Teile nur als unbekannte Wörter ermittelt werden können. Auch werden Lay-out- und Strukturelemente (z. B. Spaltensatz, Kopf- und Fußzeilen) nicht immer korrekt abgebildet, sodass es in der Folge vermehrt zu Fehlanalysen der Sprachverarbeitungskomponenten kommen kann.

Abhängig vom Erstellungsverfahren kann es auch vorkommen, dass aus PDF-Dateien keinerlei Textanteile extrahiert werden können. Entsprechende Dateien bleiben bei der Termextraktion unberücksichtigt – es fallen dafür auch keine Kosten an.

Unabhängig vom Dateiformat entscheidet es sich letztlich an den Inhalten, ob eine Datei für die Termextraktionsaufgabe geeignet ist. Eine Spreadsheet-Datei, die überwiegend Zahlenkonstrukte enthält, trägt selbstverständlich wenig zu den Termextraktionsergebnissen bei. Zwar werden in **extraTerm** zusätzliche Verfahren eingesetzt, um für die Termextraktionsaufgabe unplausible Datensätze auszufiltern, doch obliegt es dem Auftraggeber, eine inhaltlich sinnvolle Auswahl der Dateien für die Termextraktion vorzunehmen.

3. Extraktionsverfahren

Im Folgenden geben wir Hintergrundinformationen zu den Verfahren, die bei einer Termextraktion mit **extraTerm** zur Anwendung kommen.

Linguistische Analyse

Die Eingabedaten werden zunächst linguistisch analysiert. Dabei ermittelt die **morphologische Analyse** für jede einzelne Wortform die Wortart und zugehörige grammatische Informationen wie z. B. das Genus bei Substantiven. Die **grammatische Analyse** erkennt Wortgruppen und Satzbaumuster. Dabei werden mehrdeutige Wörter vereindeutigt. Durch den sehr hohen Ambiguitätsgrad des englischen Vokabulars kann es allerdings bei der grammatischen Analyse immer wieder zu Fehlanalysen kommen, sodass einzelne Wörter entweder nicht oder falsch vereindeutigt werden. Dies kann sich in den Ergebnissen der Termextraktion niederschlagen.

Auf Basis der linguistischen Analyse werden Termkandidaten nach bestimmten Kriterien ermittelt. Das grundlegende Kriterium ist dabei das Bildungsmuster:

- mehrwortiges Kompositum bestehend aus einem Substantiv und Adjektiven oder weiteren Substantiven sowie Konjunktionen oder anderen Funktionswörtern, z. B.:
  - *acoustic absorber*, *active navigation system*
- einwortiges zusammengesetztes Substantiv, z. B.:
  - *acrylester, handgrip, microsurgery*
- einfaches Substantiv (inklusive abgeleiteter Substantive), z. B.:
  - *absorption, manoeuvrability, rubber*

Die Ermittlung von mehrwortigen Komposita unterliegt einer Reihe von Einschränkungen:

- Bestimmte Adjektive wie z. B. Adjektive mit rein verstärkender oder deiktischer Bedeutung sind ausgenommen:
  - *an **excellent** fit*
  - *most **important** asset*
  - *the **abovementioned** loans*
- Kombinationen bestimmter Adjektive und Substantive sind ausgenommen:
  - *a **large amount***

Auch werden Substantive ausgenommen, die Bestandteile von Funktionsverbgefügen sind, z. B.:

  - *when it makes **sense***
  - *also take into **account***

Grundformenberechnung

Für gebeugte Wortformen wird auf Basis der morphologischen Analyse die lexikografische Grundform berechnet. So können in den Ergebnisdaten unterschiedliche Wortformen in einem Eintrag zusammengefasst und deren Vorkommen zusammengezählt werden, z. B.:

Grundform	Originalwörter
abbreviation	*abbreviation, abbreviations*
accessory	*Accessories, accessories, accessory*

Lassen sich unterschiedliche Wortformen unter Vernachlässigung der Bindestrichsetzung oder der Groß- und Kleinschreibung auf eine gemeinsame Normalform abbilden, werden diese Wortformen wiederum zusammengefasst, wobei Bindestrichvarianten als unterschiedliche Grundformen aufgeführt werden. Dabei werden immer, wenn mehrere Grundformen für einen Termkandidaten ausgewiesen werden, diese nach Vorkommenshäufigkeit der jeweiligen Schreibung sortiert, z. B.:

Grundform	Originalwörter
auto-focus licence *autofocus licence*	*Autofocus licence, auto-focus licence, autofocus licence*
co-observation *coobservation*	*Co-observation, co-observation, coobservation*

Mögliche Fehlerquellen

Auch wenn die sprachverarbeitende Software der IAI Linguistic Content AG sehr umfangreiche Wörterbücher verwendet, können einzelne Wörter oder Wortformen, die in den verarbeiteten Texten vorkommen, außerhalb der Abdeckung dieser Wörterbücher liegen. Dies gilt insbesondere für Vokabular, das speziellen Sachgebieten zuzuordnen ist oder zu einer firmenspezifischen Nomenklatur oder Terminologie gehört. Auch können bei der morphologischen Analyse einzelne Wortbildungen bedingt durch unvollständige Wörterbucheinträge unerkannt bleiben. Die IAI Linguistic Content AG erweitert ihre Wörterbücher kontinuierlich. Dennoch können lexikalische Abdeckungslücken nicht generell ausgeschlossen werden.

Für die grammatische Analyse der verarbeiteten Texte kann außerdem nicht ausgeschlossen werden, dass bestimmte Wortgefüge unzureichend oder falsch analysiert werden. Dies kann u. a. auf das Vorhandensein von orthografischen oder grammatischen Fehlern zurückzuführen sein oder auf das Vorliegen von Mehrdeutigkeiten oder die strukturelle Komplexität der verarbeiteten Texteinheit. Auch können spezielle grammatische Strukturen von den bestehenden Analysekomponenten unberücksichtigt sein. In der Folge kann es vorkommen, dass einzelne Wörter gar nicht oder falsch vereindeutigt werden und somit die Zuordnung der Wortart sowie weiterer grammatischer Merkmale ungenau oder falsch ist.

Generell gilt es festzuhalten, dass die eingesetzten Termextraktionsverfahren auf formalen linguistischen Kriterien sowie statistischen Methoden beruhen. Die Ergebnisse können somit von den Ergebnissen abweichen, die bei einer Termextraktion auf intellektueller Basis erzielt würden.

4. Ergebnisaufbereitung

Im Folgenden erläutern wir Struktur und Inhalte der Ergebnisdateien, die standardmäßig bei einer Termextraktion mit **extraTerm** bereitgestellt werden. Dies sind:

Ergebnisdatei	Inhalt
extraTerm.xlsx	tabellarische Aufbereitung der extrahierten Termkandidaten
extraTerm_ctx.txt	alle ermittelten Vorkommenskontexte pro Termkandidaten
extraTerm_ref.txt	alle Belegdokumente pro Termkandidaten
extraTerm_info.txt	Kennzahlen zur durchgeführten Termextraktion

Termkandidaten

Die bei der Termextraktion ermittelten Termkandidaten und zugehörige Informationen werden in der Datei **extraTerm.xlsx** als Excel-Tabelle erfasst, wie hier ausschnittsweise illustriert:

No.	Lemmatised Term	Word Forms	Parts of Speech	Criterion	Frequency of Occurrence	Number of Documents	Documents
241	configurable key	configurable keys	adj noun	compound.mwu	3	1	LC_09_2015.htm
242	configurable threshold	configurable threshold	adj noun	compound.mwu	1	1	LC_09_2015.htm
243	configurable video output	Configurable video output	adj noun noun	compound.mwu	1	1	LC_09_2015.htm
244	configuration	Configuration configuration configurations	noun	simplex	36	1	LC_09_2015.htm
245	configuration data	configuration data	noun noun	compound.mwu	1	1	LC_09_2015.htm
246	configuration menu	configuration menu configuration menus	noun noun	compound.mwu	5	1	LC_09_2015.htm
247	configuration option	Configuration options	noun noun	compound.mwu	4	1	LC_09_2015.htm
248	configured button	configured button	adj noun	compound.mwu	1	1	LC_09_2015.htm

In Excel stehen verschiedene Möglichkeiten zur Sortierung und Filterung der Tabelle zur Verfügung. So bietet sich beispielsweise eine Sortierung nach der Anzahl der Belegstellen und eine nachrangige Sortierung nach der Anzahl der Belegdokumente an, um eine Ergebnissichtung gemäß der Vorkommenshäufigkeit und -verteilung der Termkandidaten vorzunehmen. Jede Sortierung kann außerdem mit Filterkriterien kombiniert werden. So können beispielsweise Termkandidaten, die spezielle Extraktionskriterien erfüllen, ausgeblendet oder in den Bearbeitungsfokus genommen werden.

Im Folgenden werden die einzelnen Spalten der Excel-Tabelle erläutert.

lfd Nr

Die Spalte *lfd Nr* enthält eine durchgehende Nummerierung der ermittelten Termkandidaten.

Grundform

Die Spalte *Grundform* enthält die extrahierten Termkandidaten in der berechneten lexikografischen Grundform (siehe **Grundformenberechnung**, S. 5). In allen Fällen, in denen mehrere Grundformen für einen Termkandidaten ausgewiesen werden, sind diese nach Häufigkeit sortiert. Wir empfehlen, diese Spalte als Arbeitsgrundlage für die Weiterverarbeitung der Termkandidaten zu benutzen.

Originalwort

Die Spalte *Originalwort* enthält die Termkandidaten in allen Formen, in denen sie in den Eingabedaten vorkommen. Dies können unterschiedliche gebeugte Formen sein oder auch unterschiedliche Schreibungen.

Wortart

Die Spalte *Wortart* gibt die Wortart der ermittelten Termkandidaten bzw. ihrer Bestandteile an: *noun* (Substantiv), *adj* (Adjektiv), *det* (Artikelwort), *w* (Funktionswort wie z.B. Konjunktion, Präposition oder auch Satzeichen wie z.B. Anführungszeichen), z (Zahl). Vereinzelt kann auch *verb* auftreten, was oftmals darauf hindeutet, dass eine grammatische Struktur bei der linguistischen Analyse nicht hinreichend vereindeutigt wurde (vgl**. Mögliche Fehlerquellen**, S. 6).

Kriterium

Die Spalte *Kriterium* kann zur weiteren Bearbeitung der Ergebnisse sehr nützlich sein, da sie anzeigt, aufgrund welchen Kriteriums ein Termkandidat extrahiert wurde.

Die Extraktionskriterien betreffen unterschiedliche Facetten eines Termkandidaten. Einerseits betreffen sie das Bildungsmuster: mehrwortiges Kompositum, zusammengesetztes Substantiv, einfaches Substantiv. Sie betreffen aber auch spezielle semantische oder worttypologische Klassen, denen einzelne Termkandidaten zugeordnet werden können: Maßeinheit, Akronym. Schließlich betreffen sie den Status der linguistischen Analyse, indem unbekannte (= nicht analysierbare) Wörter speziell kategorisiert werden.

Da die Extraktionskriterien unterschiedliche Facetten eines Termkandidaten betreffen und einzelne Wörter zusätzlich auch noch mehrdeutig sein können, kommt es vor, dass für einen Termkandidaten eine Kombination mehrerer Extraktionskriterien angegeben wird. So würde z. B. *GB* zugleich als Akronym und als Maßeinheit ausgewiesen.

Die verschiedenen Extraktionskriterien werden in nachfolgender Übersichtstabelle erläutert und exemplifiziert.

Kriterium	Erläuterung	Beispiele
acronym[1]	Akronym	*AC, CE, EU*
compound.*	compound.mwu kennzeichnet mehrwortige Komposita bestehend aus einem Substantiv und Adjektiven oder weiteren Substantiven, während compound.word einwortige zusammengesetzte Substantive kennzeichnet; dagegen kennzeichnet compound.hyph den Spezialfall eines Substantivs mit Bindestrich, wobei der Bindestrich im Wörterbuch lexikalisiert ist; es kann nicht eindeutig festgelegt werden, ob eine Zusammensetzung oder ein einfaches Wort vorliegt.	*acoustic absorber,* *active navigation system,* *acrylester, handgrip, microsurgery,* *all-purpose, cut-out, e-mail*
measure	Maßeinheit	*Ah, cm, GHz*
simplex	einfaches Substantiv	*absorption, rubber*
unknown[2]	unbekanntes Wort, Wort mit Rechtschreibfehler; insbesondere bei einer Termextraktion aus PDF-Dateien (vgl. Kapitel 2) können Wortformen an Silbentrennpositionen auseinandergerissen sein, sodass einzelne Teile einer Wortform als unbekannte Wörter ermittelt werden.	*Absorbtion, Autom, AUX* *apsorp, absol,* *tion, ution*

Anzahl Belegstellen

In der Spalte *Anzahl Belegstellen* wird angezeigt, wie oft ein Termkandidat insgesamt im analysierten Dokumentbestand belegt ist.

Anzahl Belegdokumente

In der Spalte *Anzahl Belegdokumente* wird angezeigt, in wie vielen Dateien aus dem analysierten Dokumentbestand ein Termkandidat vorkommt.

[1] Bei erkannten Akronymen kann es vermehrt vorkommen, dass im analysierten Dokumentbestand ein anderer Gebrauch der betreffenden Formen vorliegt, als im verwendeten Wörterbuch angenommen.
[2] Manche durch OCR-Anwendungen bedingte Fehler sind nicht auf Anhieb ersichtlich. So etwa, wenn statt eines kleinen *L* ein großes *i* erkannt wurde. Solche Fälle werden erst durch Umstellung der voreingestellten Schriftart offensichtlich, z. B. „acrylester“ vs. „`acryIester`“.

Belege

In der Spalte *Belege* werden die Namen der Dokumente aufgelistet, in denen ein Termkandidat vorkommt. Die Dokumentnamen werden zeilenweise gelistet, wobei die Zellenhöhe in der Tabellenansicht begrenzt ist. Für den Fall, dass nicht alle Dokumentnamen in der Zelle Platz finden, wird durch „…“ angezeigt, dass weitere Belegdokumente vorliegen. Die vollständige Liste der Belegdokumente pro Termkandidaten wird in einer separaten Datei bereitgestellt (siehe unten, Abschnitt **Belegdokumente**). Da einige Zeichen, die in Dokumentnamen vorkommen können, in Excel nicht korrekt wiedergegeben werden, werden diese Zeichen sowie das %-Zeichen als Escape-Sequenzen bestehend aus %-Zeichen und zwei hexadezimalen Zeichen dargestellt.

Kontexte

In der Spalte *Kontext* sind bis zu drei Vorkommenskontexte des jeweiligen Termkandidaten angegeben. Die Vorkommenskontexte werden unter Berücksichtigung unterschiedlicher Faktoren sortiert, wie z. B. die Länge des Kontexts, das Vorkommen von Verben oder das Verhältnis zwischen Sonderzeichen und echten Wörtern. Durch die Sortierung werden möglichst einfache Satzstrukturen als Kontextbeispiele priorisiert. Sämtliche Kontexte, in denen ein Termkandidat vorkommt, werden in einer separaten Datei bereitgestellt (siehe unten, Abschnitt **Vorkommenskontexte**).

Belegdokumente

Die Ergebnisdatei **extraTerm_ref.txt** listet pro Termkandidaten sämtliche Belegdokumente. Sie weist ein zweispaltiges Format auf (Tabulator-separiert): In der ersten Spalte wird pro Zeile die laufende Nummer eines Termkandidaten aus der Termextraktionstabelle aufgeführt (Spalte *lfd Nr*), in der zweiten Spalte wird jeweils ein Belegdokument zu der betreffenden Nummer aufgeführt.

Vorkommenskontexte

Die Ergebnisdatei **extraTerm_ctx.txt** listet pro Termkandidaten sämtliche Vorkommenskontexte. Diese UTF-8-kodierte Datei weist ein zweispaltiges Format auf (Tabulator-separiert): In der ersten Spalte wird pro Zeile die laufende Nummer eines Termkandidaten aus der Termextraktionstabelle aufgeführt (Spalte *lfd Nr*), in der zweiten Spalte wird jeweils ein Vorkommenskontext zu der betreffenden Nummer aufgeführt.

Termextraktionskennzahlen

Die Ergebnisdatei **extraTerm_info.txt** enthält grundlegende Kennzahlen zur durchgeführten Termextraktion.

Bezogen auf die Eingabedaten werden angegeben:

- Anzahl der Dokumente, aus denen Textanteile extrahiert werden konnten und die somit bei der Termextraktion berücksichtigt wurden (vgl. Kapitel 2)
- Anzahl der Satzobjekte
- Anzahl der echten Wörter (keine Satzzeichen u. Ä.)

Bezogen auf die Extraktionsergebnisse werden angegeben:

- Anzahl der erkannten Terme (für den Fall, dass bei der Termextraktion eine Bestandsterminologie berücksichtigt wurde)
- Anzahl der ermittelten Termkandidaten

Für die ermittelten Termkandidaten werden angegeben:

- Anzahl der Nominalphrasen
- Anzahl der zusammengesetzten Substantive
- Anzahl der einfachen Substantive
- Anzahl der Akronyme, Maßeinheiten u. Ä.
- Anzahl der unbekannten Wörter

Da im Hinblick auf die unterschiedlichen Extraktionskriterien Mehrfachzuweisungen möglich sind, kann sich die Summe dieser Angaben von der absoluten Anzahl der Termkandidaten unterscheiden.