Jan Oevermann

Künstliche Intelligenz und intelligente Textservices

1 Einführung

Mit Anbruch des Zweiten Maschinenzeitalters erlebt auch die Künstliche Intelligenz (KI) eine Renaissance, die durch immer leistungsfähigere Endgeräte und eine Fülle von digital verfügbaren Daten vorangetrieben wird (Brynjolfsson/McAfee 2016). Obwohl die zugrundeliegenden Konzepte nicht neu sind, werden viele Anwendungsszenarien dadurch erst seit Kurzem realistisch umsetzbar. Neben den beeindruckenden Fortschritten im Bereich des maschinellen Sehens, der Mustererkennung und der Verarbeitung gesprochener Sprache sind es im Besonderen die textbasierten Anwendungen, die vom enormen Effizienzgewinn und von höherer Zuverlässigkeit profitieren.

Die Technische Dokumentation als traditioneller Produzent und Konsument von geschriebenem Text kann mit Unterstützung von KI das volle Potenzial der eigenen Inhalte ausnutzen, um Nutzern neuartige Interaktions- und Zugriffsmöglichkeiten zu geben. Durch intelligente Textservices können Redakteure effizient die viel beschworenen intelligenten Informationen erstellen, die heute schon Anwendungen ermöglichen, die vielleicht erst morgen erfunden werden.

2 Ziele des Beitrags

Der folgende Text gibt einen Überblick über den Einsatz intelligenter textbasierter Anwendungen, die Mehrwerte im Bereich der Technischen Dokumentation generieren können. Der Beitrag konzentriert sich dabei auf den aktuellen Stand der Forschung und stellt exemplarisch drei Services vor, die konkret anwendbar sind.

Nach einer kurzen Einführung und Definition des Intelligenzbegriffs im Bereich der Informationsverarbeitung werden Besonderheiten der Textsorte Technische Dokumentation in Bezug auf die maschinelle Verarbeitung gelistet und deren Konsequenzen erörtert. Anschließend werden verschiedene Anwendungen aus drei Beispielgebieten der intelligenten Textservices vorgestellt und deren Funktionsweise erklärt.

3 Was ist intelligent?

3.1 Künstliche Intelligenz

Im Allgemeinen wird mit Künstlicher Intelligenz ein Teilgebiet der Informatik bezeichnet, das sich mit der Automatisierung und Imitation von (menschlicher) Intelligenz befasst. Der Oberbegriff umfasst ein weites Gebiet unterschiedlicher Disziplinen, das von simplen regelbasierten Systemen bis hin zu leistungsfähigen neuronalen Netzen reicht.

Für die Industrie von besonderer Bedeutung sind dabei spezialisierte Anwendungen, die einen Teilbereich menschlichen Denkens entweder übernehmen oder unterstützen können. Dies ist besonders dann sinnvoll, wenn es darum geht, große Mengen von Daten zu verarbeiten oder sich häufig wiederholende Aufgaben zu automatisieren. Auch der Bereich der Technischen Dokumentation kann mit seinen größer werdenden Mengen Content und immer kürzeren Produktlebenszyklen von „intelligenten Anwendungen" profitieren.

Ein Teilgebiet der KI ist das Machine Learning (dt.: maschinelles Lernen), welches auf Basis von Erfahrungen (sprich: vorhandenen Daten) neues Wissen generiert, das dann auf neue, dem System unbekannte Daten angewandt werden kann (Lerntransfer). Das Wissen wird in sogenannten Modellen gespeichert, die mit Hilfe induktiver Methoden aus den vorhandenen Eingabedaten generiert wurden. Enthalten die Eingabedaten auch die erwarteten Ergebnisse, spricht man von „überwachtem Lernen".

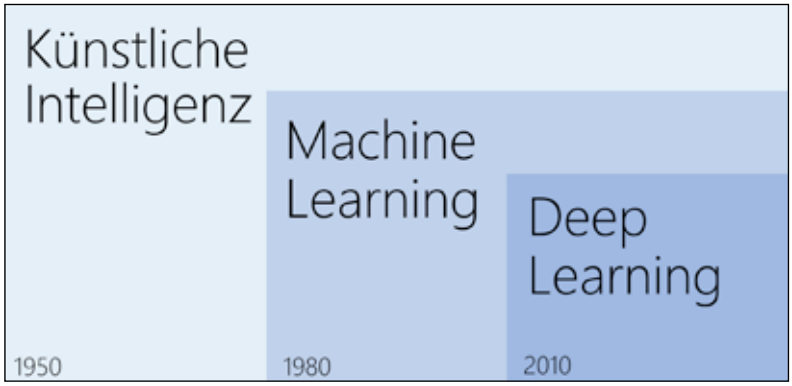

Abb. 1: Machine Learning als Teilgebiet der Künstlichen Intelligenz[1]

Um beim überwachten Lernen bestimmte Muster erkennen zu können, müssen Eigenschaften definiert werden, die aus dem Text abgeleitet und numerisch abgebildet werden können (Manning/Schütze 1999). Bei der statistischen Sprachverarbeitung werden dafür häufig

1 Abgeleitet von: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/.

einzelne Wörter oder Wortgruppen verwendet, deren Häufigkeit und Verteilung in den Klassen gezählt wird (Bag-of-Words-Modell).

Deep Learning als relativ junge Disziplin innerhalb des Machine Learnings erweitert künstliche neuronale Netze um weitere Schichten (aus diesem Grunde das „Deep") von Neuronen, die in der Lage sind, komplexe Eingaben (etwa Bilder) durch zahlreiche (Quer-)Verbindungen zu verarbeiten, um am Ende die gewünschte Ausgabe zu erzielen.

3.2 Intelligente Informationen

Als „intelligente Informationen" werden i.d.R. modularisierte und mit (klassifizierenden) Metadaten angereicherte Texteinheiten bezeichnet, die neuartige Zugriffsmöglichkeiten abseits des Dokumentenkontexts ermöglichen. Durch die Verbreitung mobiler Endgeräte mit kleinen Bildschirmen und der steigenden Erwartung von Benutzern, Informationen individuell und kontextsensitiv zur Verfügung gestellt zu bekommen, gewinnt diese Art der Inhaltsaufbereitung immer mehr an Bedeutung. Content-Delivery-Portale und darauf aufbauende Anwendungen können das volle Potential ausnutzen, in dem durch Filter- und Suchmechanismen Informationen zielgerichtet ausgeliefert werden (Ziegler 2015).

Die aus dem Content Management stammenden Methoden der Modularisierung und Klassifikation bekommen durch die Möglichkeiten des Content Delivery eine neue Relevanz und Dringlichkeit. Durch die rasante Digitalisierung in allen Bereichen der Industrie sehen sich viele Unternehmen gezwungen, Informationen systemübergreifend zu integrieren und nach modernen Standards extern bereitzustellen. Treiber sind hierbei vor allem mobile Apps und unternehmensweite Kundenportale. Auch die immer weiter steigende Menge an digitalen Daten in Verbindung mit den in der Technischen Dokumentation bestehenden Aufbewahrungspflichten stellen Unternehmen vor große Herausforderungen des internen Informationsmanagements.

Wichtigste Faktoren um Informationen intelligent zu machen, sind dabei die Modularisierung der Inhalte in abgeschlossene Sinneinheiten (bei DITA z.B. „Topics" genannt) und die Klassifizierung dieser Module mit Informationen zu ihrem Inhalt und ihrer Zugehörigkeit. Eine der verbreitetsten Methoden zur systematischen Klassifikation ist PI-Class® von Ziegler (Drewer/Ziegler 2011). PI-Klassifikationen werden als Taxonomien definiert und können systemunabhängig eingesetzt werden. Intrinsische Klassifikationen kategorisieren eindeutig die Informationsart des Inhalts („Informationsklasse") und verknüpfen ihn mit der beschriebenen Produktkomponente („Produktklasse"). Extrinsische Klassifikationen ergänzen

die Methode um die geplante Verwendung des Moduls für Produktmodelle und Dokumenttypen oder Zielgruppen.

Inhalte können aber auch mit Hilfe semantischer Modelle beschrieben werden. Für die Technische Dokumentation wurde hierzu der Standard iiRDS („intelligent information Request and Delivery Standard“) entwickelt, der neben einem einheitlichen Paketformat zur Verteilung auch ein Schema und Vokabular zur genauen Beschreibung der Eigenschaften des ausgelieferten Contents (tekom e.V. 2017) definiert. Die RDF-basierte Auszeichnung beruht auf standardisierten Konzepten im Bereich von „Linked Data“, welche eine Ein- und Verbindung mit beliebigen anderen Vokabularen und Ontologien zulässt und somit die Integration in komplexere Informationsszenarien ermöglicht (sog. „Linked Open Data Cloud“).

3.3 Intelligente Textservices

Als „intelligente Textservices“ werden in diesem Text alle Softwaredienste bezeichnet, die durch Methoden der Künstlichen Intelligenz oder anderweitige Analyseverfahren Texte so verarbeiten, dass dadurch neue Erkenntnisse entstehen. Dazu gehören u.a. Machine-Learning- oder Deep-Learning-Verfahren und Ähnlichkeitsanalysen.

Diese Services können Redakteure bei bestimmten Tätigkeiten unterstützen, indem z.B. große Datenmengen verarbeitet, untersucht und für eine schnellere manuelle Betrachtung vorbereitet werden. In diesem Fall spricht man von Assistenzsystemen oder „Intelligence Amplification“ (Dengel 2017). Durch die strengen rechtlichen Anforderungen an die Technische Dokumentation sind Szenarien, in denen Teilarbeiten eines Technischen Redakteurs (z.B. Metadatenvergabe) komplett übernommen werden, zwar denkbar, derzeit aber in den meisten Fällen nicht umsetzbar. Die „Maschine als Zuarbeiter“ ist jedoch ein Szenario, das mit einer menschlichen Qualitätssicherung bereits heute effektiv funktionieren kann und im Übersetzungsbereich teilweise schon erfolgreich eingesetzt wird.

4 Technische Dokumentation

Um bei der Anwendung intelligenter Textservices auf Technische Dokumentation optimale Ergebnisse zu erzielen, müssen die Besonderheiten und Eigenschaften dieser Textsorte betrachtet werden. Dabei spielen sowohl äußere Faktoren wie die modulare Erfassung und bestehende Klassifikationsmethoden eine Rolle als auch die inhaltlichen Eigenheiten, die vor allem Standardisierung und rechtlichen Vorgaben geschuldet sind (Oevermann/Ziegler 2016). Im Folgenden soll ein kurzer Überblick gegeben werden, welche Charakteristiken von modularer Technischer Dokumentation relevant für die Verar-

beitung durch intelligente Textservices sind. Diese beziehen sich vor allem auf die statistische Sprachverarbeitung, die überwiegend im Machine Learning und bei der inhaltlichen Ähnlichkeitsanalyse angewandt wird.

4.1 Modulare Inhaltserfassung

Das im Content Management verbreitete Vorgehen, Inhalte modular zu erfassen, hat Auswirkungen auf die Verarbeitung der Texte durch Services. Viele der bestehenden Standardmethoden zur Extraktion und Gewichtung charakteristischer Merkmale gehen von Dokumenten als Referenzeinheiten aus und müssen an die erheblich kleinere Größe von Modulen angepasst werden. Dies kann Nachteile mit sich bringen, da ein kürzerer Text prinzipiell weniger auswertbare Merkmale besitzt. Jedoch eignen sich Inhalte aus der Technischen Dokumentation prinzipiell gut, da viele Herausforderungen der maschinellen Sprachverarbeitung vernachlässigt werden können: Synonymie, Ambiguität, uneinheitliche Terminologie, Emotionen (Sentiments) sowie heterogene oder qualitativ schlechte Lerndaten.

4.2 Klassifikationsmethoden

Die in der Technischen Dokumentation häufig eingesetzte Klassifikation nach PI-Class® gibt in der Regel eine taxonomische Klassenhierarchie vor, die aus intrinsischen und extrinsischen Ausprägungen besteht. Auch alternative Klassifikationslogiken (z.B. nach der iiRDS-Modellierung) lassen sich entsprechend einteilen. Steht einem Service zur Verarbeitung nur der einzelne Text eines Objekts zur Verfügung, so können daraus nur intrinsische Klassen direkt abgeleitet werden. Da die meisten Klassifikatoren auf einem flachen Set an Klassen agieren, muss zudem eine feste Klassenebene definiert werden.

4.3 Standardisierung

Ein besonderes Merkmal der Technischen Dokumentation ist die standardisierte Texterstellung zur Erhöhung der Konsistenz, Steigerung der Häufigkeit der Wiederverwendung und Senkung der Übersetzungskosten. In Kombination mit kontrollierter Terminologie entstehen so Wortgruppen und Formulierungsmuster, die über Module hinweg immer gleich sind (z.B. Handlungsanweisungen). Dies hat zur Folge, dass besonders diese Wortfolgen charakteristisch für bestimmte Module oder Klassen sind. Die Gesamtzahl der eindeutigen Merkmale sinkt.

4.4 Qualitätssicherung

Durch die hohen Qualitäts- und Sicherheitsanforderungen in der Technischen Dokumentation unterliegen Prozesse, die Inhalte automatisiert verarbeiten, besonderen Ansprüchen an Zuverlässigkeit.

Darum müssen Services, die z.B. Metadaten generieren oder Content analysieren, Mechanismen vorsehen, um die eigene Zuverlässigkeit zu beurteilen und im Zweifel menschliche Entscheidungen herbeizuführen. Ein Beispiel ist die automatisierte Klassifizierung, bei der in Grenzfällen der Mensch über die endgültige Zuordnung entscheidet.

4.5 Vorhandene Daten

Eine Besonderheit in der Technischen Kommunikation sind die vorhandenen Daten. Auf der einen Seite besitzen Unternehmen, die mit einem Content-Management-System arbeiten, qualitativ hochwertige, oft semantisch strukturierte Daten in einem medienneutralen Format. Auf der anderen Seite müssen durch die Aufbewahrungsbestimmungen weiterhin dokumentbasierte Anleitungen zur Verfügung gestellt werden können, die nur in unstrukturierten Formaten wie PDF vorhanden sind. Dies führt zu verschiedenen Ausgangslagen, die z.B. in der Trainingsphase des maschinellen Lernens beachtet werden müssen.

5 Anwendungen

Die folgenden Abschnitte stellen drei Bereiche der intelligenten Textservices im Umfeld der Technischen Dokumentation vor und zeigen deren Einsatzmöglichkeiten. Die darauf basierenden Anwendungen beziehen sich vor allem auf die Gebiete Bestandsdatenaufbereitung, Analyse und Qualitätssicherung sowie Autorenunterstützung.

- Die manuelle (Nach-)Bearbeitung von Bestandsdaten durch Redakteure ist oft aufwendig, repetitiv und langwierig und ist deshalb besonders geeignet für die softwaregestützte Verarbeitung. Darunter fallen die nachträgliche Klassifizierung von Modulen, Bereinigungsarbeiten im Datenbestand oder das Segmentieren von Dokumenten.
- Intelligente Analysen über die Inhalte können helfen, tiefere Einblicke in die Struktur des Datenbestands zu bekommen und damit Qualitätsprobleme frühzeitig zu erkennen. Neben Kennzahlen zur Klassifikationsqualität können auch die Zusammensetzung und die Ähnlichkeit der Module untersucht werden.
- Durch die Einbindung der Services in den Erstellprozess können Autoren effektiv bei ihrer täglichen Arbeit unterstützt werden. So können z.B. händisch vergebene Metadaten automatisch auf Plausibilität geprüft werden und redundante Module durch Ähnlichkeitsvergleiche im Datenbestand verhindert werden.

5.1 Automatisierte Klassifizierung

Textklassifizierung ist eine der am besten erprobten Disziplinen des

Maschinellen Lernens. Unter Beachtung der im Abschnitt 4 aufgeführten Besonderheiten technischer Inhalte können CMS-Module zuverlässig mit Klassifikationen versehen werden (Oevermann/Ziegler 2016). Die Zielsetzung dabei ist immer, ein der Maschine bisher unbekanntes Modul mit Vorhersagen bestimmten (zuvor definierten) Klassen zuzuordnen. Im Bereich der intrinsischen Klassifikationen handelt es sich um sogenannte Multi-Class-Probleme, bei denen einem Modul immer nur eine Klasse aus einem geschlossenen Set (z.B. Informationsklasse) zugeordnet werden kann.

5.1.1 Service-Steckbrief

Automatisierte Klassifizierung	
Eingabe	Nicht klassifizierte Module
Ausgabe	Klassifizierte Module
Voraussetzungen	Klassifizierte Trainingsdaten (ca. 500-1000 Module)/ Modell
Anwendungen	• Klassifizierung von Bestandsdaten (Migration) • Intelligenter Import in Content-Delivery-Portale • Autorenunterstützung durch Klassifikationskontrolle

5.1.2 Funktionsweise

Mit Hilfe von manuell klassifizierten Modulen wird ein Modell trainiert, welches durch bestimmte sprachliche Merkmale die Charakteristiken der verschiedenen Klassen „unterscheiden lernt“ und dieses Wissen auf neue – nicht klassifizierte – Inhalte anwenden kann. Dabei kommt in den meisten Fällen eine statistische Sprachverarbeitung zum Einsatz, die ohne Kenntnis der zugrundeliegenden sprachspezifischen Grammatik die Merkmale eines Textes anhand von Wort- oder Wortgruppenhäufigkeiten erfasst. So kann sprachneutral klassifiziert werden, solange Trainingsdaten und zu klassifizierende Module die gleiche Sprache haben.

Jedem der extrahierten Textmerkmale wird eine Gewichtung zugeordnet, die sich z.B. über die Verteilung innerhalb einer Klasse kombiniert mit der Vorkommenshäufigkeit über die gesamte Trainingsmenge errechnet. Die sich daraus ergebende Gewichtungsverteilung wird für jede Klasse bestimmt und die zu klassifizierenden Module damit verglichen.

Der Klassifizierer, also die Funktion, die eine bestimmte Vorhersage für eine Klasse trifft, kann das jeweilige Ergebnis mit einer sogenannten Konfidenz bewerten. Dieser Wert beziffert, wie zuverlässig eine Vorhersage getroffen werden kann. Die Konfidenz kann in Prozentwerten ausgedrückt werden und z.B. über einen Grenzwert als Qualitätssicherung verwendet werden (Oevermann/Ziegler 2016)

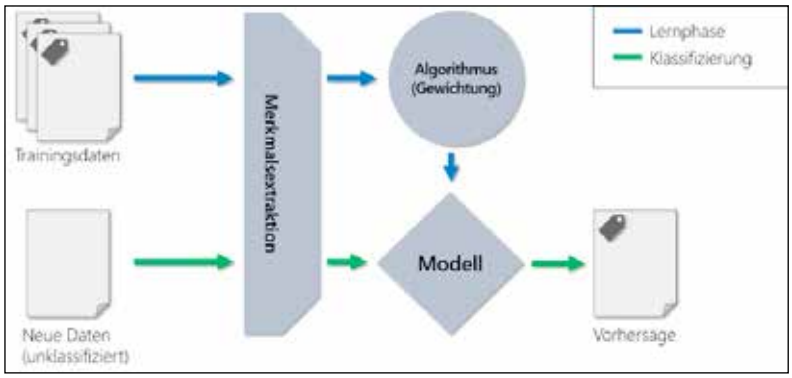

Abb. 2: Automatisierte Klassifizierung; Lern- und Klassifizierungsphase

Der Vorteil einer solchen automatisierten Vorgehensweise liegt in der hohen Performanz der Klassifizierung, welche große Modulmengen in kürzester Zeit verarbeiten kann. Zu beachten ist jedoch die Voraussetzung, dass eine repräsentative Menge an Trainingsdaten vorhanden sein muss, um ein Modell mit ausreichender Qualität anzulernen. Diese werden oft als „Goldstandard" bezeichnet, da auf ihnen die Genauigkeit und Zuverlässigkeit der automatisierten Klassifizierung basiert. Sind Fehler in den Trainingsdaten, so werden diese auch durch das erlernte Modell weitergegeben.

Mit hochwertigen und repräsentativen Trainingsdaten können Ergebnisse zwischen 85-95 % Genauigkeit („Accuracy") bei einer Berechnungszeit von wenigen Millisekunden pro Modul erreicht werden.

5.1.3 Anwendungsbeispiele

Unternehmen, die sich entscheiden intelligente Informationen (und damit klassifizierende Metadaten) einzusetzen, sind oft mit einem massiven initialen Aufwand konfrontiert, da Bestandsdaten nachträglich mit entsprechenden Klassifikationen versehen werden müssen. Ein intelligenter Textservice, der Module automatisiert klassifiziert, kann diese Arbeit nicht komplett übernehmen, jedoch die Arbeit wesentlich erleichtern. Anstelle des kompletten Modulbestands muss nur eine repräsentative und ausreichende Menge (ab ca. 500-1000 Module) händisch klassifiziert werden, um ein Modell zu trainieren, welches dann Klassen für die verbleibenden Module in einem automatisierten Prozess vergibt. Der für jede Vorhersage berechnete Konfidenzwert kann Rückschluss darauf geben, ob eine Nachkontrolle durch einen Redakteur erforderlich ist.

Alternativ ist eine automatisierte (Nach-)Klassifizierung auch ad hoc möglich, z.B. bei einem intelligenten Prozess des Imports in ein Content-Delivery-Portal. Moderne CDP bieten dafür weitreichende Integrations- und Erweiterungsmöglichkeiten, z.B. über WebHooks. Diese registrieren für ein bestimmtes Ereignis (z.B. „Neuer Inhalt hochgeladen") eine auszuführende Aktion (z.B. „Klassifizierungsser-

vice aufrufen“) und führen diese dann beim Eintritt des Ereignisses automatisch aus. So wird der Inhalt beim Import ganz automatisch zur intelligenten Information.

Eine weitere Anwendungsmöglichkeit liegt in der Autorenunterstützung. Zwar legen Technische Redakteure schon beim Anlegen eines Moduls die intrinsischen Klassifikationen fest, jedoch passieren auch an dieser Stelle manchmal Fehler. Diese können durch eine automatische Prüfung, bei der die Klassifikation mit dem Ergebnis eines maschinellen Klassifizierers verglichen wird, abgefangen werden. Beim Auftreten von Unterschieden kann z.B. eine Warnmeldung angezeigt werden, die den Redakteur auffordert, die Klassifikationen des Moduls nochmals zu überprüfen.

5.2 Automatisierte Segmentierung

Das „Portable-Document-Format“ (PDF) ist eine beliebte Art Content zu verteilen und zu archivieren, da es eine exakte visuelle Reproduktion bei Druck und Bildschirmanzeige garantiert. Doch das Format hat auch Nachteile, wenn es darum geht, semantische und strukturelle Informationen zu konservieren.

Um bestehende PDF-Dokumente nachträglich wieder mit Strukturinformationen anzureichern, gibt es Methoden der automatisierten Segmentierung, die z.B. bei Dateien ohne integrierte Lesezeichen die Kapitelstruktur rekonstruieren können.

5.2.1 Service-Steckbrief

Automatisierte Segmentierung	
Eingabe	Monolithisches PDF-Dokument
Ausgabe	Metadaten, die Seitenbereichen bestimmte Klassifikationen zuordnen (im einfachsten Fall die Kapitelstruktur)
Voraussetzungen	Klassifizierte Trainingsdaten (ca. 500-1000 Module)/ Modell
Use Cases	• Aufbereitung von PDF-Dateien für Content Delivery • Analyse von PDF-Inhalten

5.2.2 Funktionsweise

Der Text eines PDFs wird extrahiert und in zahlreiche Blöcke zerteilt, die der durchschnittlichen Modulgröße entsprechen. Diese Textfragmente werden mit einem festgelegten Versatz über das Dokument verteilt und anschließend mit der gleichen Methode wie Module klassifiziert (siehe Abschnitt 5.1). Für die Klassifikationsergebnisse werden Konfidenzwerte ausgegeben und die Textfragmente wieder ihrer ursprünglichen Seitenposition zugeordnet (Oevermann 2016).

Daraus ergibt sich eine charakteristische Verteilung von gleich klassifizierten Blöcken und Einbrüche der Konfidenz an Grenzen zwischen verschiedenen Klassen (Abb. 3). Nun können anhand von

bestimmten Klassen (z.B. Wartungsinformation) bestimmte (Seiten-) Bereiche des PDFs eingegrenzt werden.

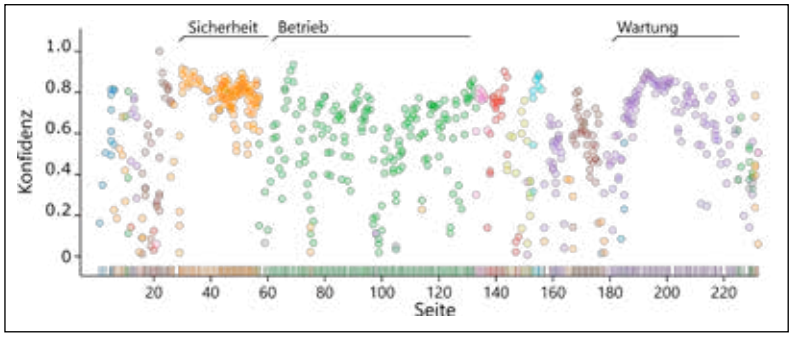

Abb. 3: Ergebnis einer automatisierten PDF-Segmentierung

5.2.3 Anwendungsbeispiele

Bei der Bereitstellung von Inhalten durch Content-Delivery-Portale oder durch mobile Applikationen werden Bestandsdaten, die lediglich in dokumentbasierter Form (z.B. als PDF) vorliegen, oft vernachlässigt, da granulare Zugriffsmethoden wie die Filterung nach Klassifikationen auf Modulebene mit ihnen nicht möglich sind. Durch die zusätzlich generierten Metadaten der automatisierten Segmentierung können PDF-Dokumente auf die gleiche Art und Weise gefiltert werden wie modulare Inhalte. Bei einer Vorfilterung von Inhalten (z.B. für einen spezifischen Service-Auftrag) kann nur der relevante Teil des PDFs übermittelt werden und somit z.B. Bandbreite und Datenvolumen eingespart werden.

Für Analysen, die den Anteil bestimmter Klassen (z.B. von Informationsarten) in langen PDF-Dokumenten aufzeigen sollen, kann ein Textser-vice nach dem gleichen Schema vorgehen und sowohl Position als auch Verhältnisse der Klassifikationen auswerten. Diese Methode kann zum besseren Verständnis der Inhalte großer Mengen an Bestandsdaten verwendet werden (etwa wenn Dokumentarten miteinander verglichen werden).

5.3 Ähnlichkeitsanalyse

Viele Anwendungen im Bereich des „Information Retrieval“ (z.B. Suchmaschinen) und der automatisierten Klassifizierung basieren auf dem Prinzip der Ähnlichkeitsanalyse. Darum liegt es nahe, dieses Verfahren auch in seiner ursprünglichen Form zu nutzen: um zu beziffern, wie ähnlich sich zwei Texte sind. Dabei kommen in den meisten Fällen wieder die Methoden der statistischen Sprachverarbeitung zum Einsatz.

5.3.1 Service-Steckbrief

Ähnlichkeitsanalyse	
Eingabe	Gesamter Modulbestand oder Teile davon
Ausgabe	Ähnlichkeitswerte für Module
Voraussetzungen	Keine
Use Cases	• Bereinigung von Modulbeständen • Ähnlichkeitsbasierte Wiederverwendung

5.3.2 Funktionsweise

Diese Methode hat den Vorteil, die Ähnlichkeit in einem Wert zwischen 0 und 1 auszudrücken, sehr performant zu arbeiten und sprachneutral zu sein. Besonders nützlich ist das Vorgehen beim Identifizieren von Formulierungsvarianten und kleinen Änderungen. Werden die gleichen semantischen Konzepte mit unterschiedlichen Wörtern beschrieben, greift die Analyse in dieser einfachen Variante nicht.

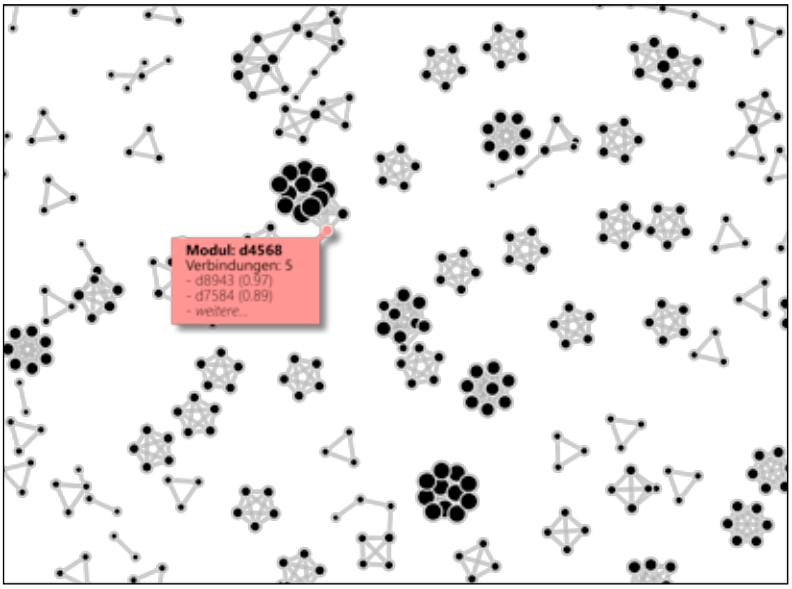

Abb. 4: Analyseergebnis als Cluster von sehr ähnlichen oder identischen Modulen (mit Ähnlichkeit > 0.8). Strichstärke: Ähnlichkeitswert, Durchmesser: Anzahl Verbindungen

5.3.3 Anwendungsbeispiele

Nach jahrelanger Arbeit innerhalb eines CMS oder einer umfangreichen Migration können sich Module mit identischem oder fast gleichem Inhalt ansammeln, die als unkontrollierte Redundanzen Übersetzungskosten steigern und Inkonsistenzen verursachen. Um diese nicht gewollten Duplikate zu identifizieren, können Ähnlichkeitsanalysen über den Modulbestand durchgeführt werden. Dabei wird zu jeder Modulkombination ein Ähnlichkeitswert berechnet, der auf Basis von Wortgruppenverteilungen ermittelt wird.

Bei großen Datenbeständen oder mehreren Redakteuren kann es vorkommen, dass Module doppelt erstellt werden oder unkontrollierte Kopien eingecheckt werden. Eine Ähnlichkeitsanalyse, die im Hintergrund den aktuellen Text mit den vorhandenen Daten abgleicht, kann helfen, diese Fehler zu vermeiden und die Wiederverwendung zu steigern (Soto u.a. 2015). Im Gegensatz zu Authoring-Memory-Systemen wird hierbei nicht auf Segmentebene verglichen, sondern das gesamte Modul betrachtet. Eine visualisierende Oberfläche kann dem Autor helfen, ähnliche oder gleiche Module und die Unterschiede im Text zu identifizieren.

6 Fazit

Intelligente Textservices sind bereits jetzt Realität und können Technische Redakteure bei ihrer Arbeit sinnvoll unterstützen. Dabei wird der Mensch in seiner Arbeit nicht ersetzt, sondern bekommt einen fleißigen Helfer zur Seite gestellt, der so manche Sysiphusarbeit mit Leichtigkeit übernimmt.

In Verbindung mit den neuen Herausforderungen, die „intelligente Informationen" mit sich bringen, sind die smarten Services der beste Weg, um schnell zu einem Ergebnis zu kommen. So können zum Beispiel große Mengen an heterogenen Daten gezielt aufbereitet werden, um strukturierte Zugriffsweisen zu ermöglichen (Bader/Oevermann 2017).

7 Literatur

Bader, Sebastian/Oevermann, Jan (2017): Semantic Annotation of Heterogeneous Data Sources: Towards an Integrated Information Framework for Service Technicians. Erscheint in: Proceedings of the 13th International Conference on Semantic Systems. SEMANTiCS 2017, Amsterdam, NL. New York City, US: ACM.

Brynjolfsson, Erik/McAfee, Andrew (2016): The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. New York City, US: Norton.

Dengel, Andreas (2017): AI & IA: From Artificial Intelligence to Intelligence Amplification. Präsentation auf der UA Reloaded 17, St. Leon-Rot.

Drewer, Petra/Ziegler, Wolfgang (2011): Technische Dokumentation: Übersetzungsgerechte Texterstellung und Content-Management. Würzburg: Vogel.

Gesellschaft für Technische Kommunikation, tekom (Hrsg.) (2017): iiRDS Specification – intelligent information Request and Delivery Standard. First Public Working Draft. https://iirds.tekom.de.

Manning, Christopher D./Schütze, Hinrich (1999). Foundations of Statistical Natural Language Processing. Cambridge, US : MIT Press.

Oevermann, Jan (2016): Reconstructing Semantic Structures in Technical Documen-

tation with Vector Space Classification. In: Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems. SEMANTiCS 2016, Leipzig, Aachen: CEUR Workshop Proceedings (= CEUR-WS Bd. 1695).

Oevermann, Jan/Ziegler, Wolfgang (2016): Automated Intrinsic Text Classification for Component Content Management Applications in Technical Communication. In: Proceedings of the 2016 ACM Symposium on Document Engineering. DocEng '16, Vienna, New York City: ACM Press, 95-98.

Soto, Axel J. u.a. (2015): Similarity-Based Support for Text Reuse in Technical Writing. In: Proceedings of the 2015 ACM Symposium on Document Engineering. DocEng '15, Lausanne, New York City: ACM Press, 97-106.

Ziegler, Wolfgang (2015): Content Management und Content Delivery. Powered by PI-Class. In: Tagungsband zur tekom Jahrestagung 2015. Stuttgart: tekom.