Bei diesem Beitrag handelt es sich um einen wissenschaftlich begutachteten und freigegebenen Fachaufsatz („reviewed paper“).

doi.org/10.37544/1436-4980-2025-06-23

Wissensbewahrung in KMU

Aufbau eines Retrieval-Augmented-Generation-Systems

M. Schmauder, G. Ott, E. Schönwälder, M. Hahmann

ZUSAMMENFASSUNG Wissensbewahrung ist eine Herausforderung für Unternehmen. Da Wissen zumeist als Text kodiert ist, bieten große Sprachmodelle eine vielversprechende Lösung für Erhalt und Nutzung durch ein System, das auf dem Konzept der Retrieval-Augmented Generation (RAG) basiert. Praxisgeeignete Methoden zur Wissensexplikation werden mit lokal ausgeführten Sprachmodellen und traditionellen Retrieval-Algorithmen kombiniert. Der Beitrag zeigt das Lösungskonzept sowie Herausforderungen und erfolgsversprechende Umsetzungsstrategien.

STICHWÖRTER

Mensch und Technik, Wissensmanagement, Arbeitsorganisation

Design of a Retrieval-Augmented Generation (RAG) system for knowledge retention in SMEs

ABSTRACT Knowledge retention is challenging for companies. Since most knowledge is encoded as text, large language models offer a promising solution for preservation and utilization. A system based on the approach of retrieval augmented generation has been developed. Practical methods for knowledge explication are combined with small, locally executed language models and traditional retrieval algorithms. The article shows the solution approach as well as challenges and promising strategies for implementation.

1 Problematik

Erfahrungswissen wird auch in der zunehmend digitalisierten Arbeitswelt weiter eine zentrale Rolle spielen. Darüber herrscht sowohl in der wissenschaftlichen Gemeinschaft als auch in der Wirtschaft weitgehend Konsens. Diese Einschätzung wird durch empirische Befunde eindeutig gestützt. Eine umfassende Erhebung zeigt, dass fast 85 % der Führungskräfte die fundamentale Bedeutung von Erfahrungswissen für den Unternehmenserfolg anerkennen, wobei 43 % diese als „sehr wichtig“ und weitere 35 % als „ziemlich wichtig“ einstufen [1].

Außerdem lassen sich spezifische Aufgabenbereiche identifizieren, in denen Erfahrungswissen besonders zum Tragen kommt. Zu diesen gehören vor allem:

- effektive Bewältigung operativer Problemstellungen
- fundierte Entscheidungsfindung bei unvollständiger Informationslage
- Erkennen komplexer Zusammenhänge
- erfolgreiche Bewältigung von Krisensituationen [1]
- menschliche Eingriffserfordernisse aufgrund noch existenter Fehleranfälligkeit intelligenter Systeme [1, 2]

Zudem ermöglicht die Kombination von digitalem Know-how und Erfahrungswissen die effiziente Umsetzung digitaler Transformationsprozesse.

Angesichts dieser Erkenntnisse bleibt der konsequente und systematische Umgang mit Erfahrungswissen in Unternehmen als Herausforderung bestehen. Ein Risiko besteht insbesondere dann, wenn Mitarbeitende aus dem Unternehmen ausscheiden oder wenn durch innovative technische Lösungen, beispielsweise Automatisierung, fundamentale Veränderungen der Arbeitstätigkeiten erfolgen. In solchen Situationen droht der unwiederbringliche Verlust essenziellen Erfahrungswissens aus der organisationalen Wissensbasis. Parallel dazu erfordert die Sicherung der Wettbewerbs- und Innovationsfähigkeit die Implementierung effektiver Strategien zur kontinuierlichen Identifikation und Integration neuer Wissensbestände und Informationen in die Unternehmenswissensbasis.

Diese Herausforderungen waren der Ausgangspunkt für die kooperative Entwicklung eines Lösungskonzeptes zur Wissensbewahrung und -bereitstellung mit zwei produzierenden Unternehmen im Rahmen des BMBF-geförderten regionalen Kompetenzzentrum der Arbeitsforschung „PerspektiveArbeit Lausitz“ (PAL).

Im Kontext der Forschungsarbeiten wird Erfahrungswissen in Anlehnung an *Plath* [3] als explizites praktisches und theoretisches Wissen (zum Beispiel technisch-technologisches Wissen, Prozesswissen) sowie implizites Wissen (zum Beispiel über Wirkzusammenhänge, funktionale Abhängigkeiten) in Bezug auf Sachverhalte und Vorgehensweisen definiert.

Im Zuge der Digitalisierung in der Arbeitswelt ist der Umfang digital verfügbarer Information in allen Unternehmen immens gestiegen. In vielen Unternehmen, so zeigen Arbeiten im Projekt PAL, verteilen sich die Informationsbestände jedoch auf unterschiedliche Softwarelösungen. Die täglichen Herausforderungen verschieben sich nun dahin, die gewünschten Dateien zu finden und ohne Overload zu verarbeiten [2]. Damit bleibt aber auch die Thematik undokumentierter Erfahrungen nach wie vor virulent.

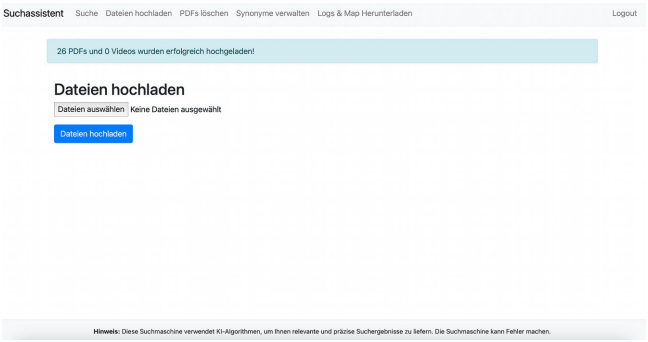

Bild 1. Einpflegen von Dateien in den RAG (Retrieval-Augmented Generation)-Korpus. *Grafik: TU Dresden*

In der Wissensmanagementdebatte spielt zudem das Verhältnis von personenorientierten Methoden und technisch unterstütztem Wissensaustausch eine Rolle. Die Vorteile des personenorientierten Wissensaustauschs gegenüber rein technologieorientierten Ansätzen zeigen sich vor allem in der Ausprägung, implizites Wissen zu vermitteln und ein tiefes Kontextverständnis zu fördern. Während technologiegetriebene Systeme effizient explizites Wissen speichern und bereitstellen, erlauben persönliche Interaktionen den Transfer von schwer fassbarem Erfahrungswissen, Intuitionen und ungeschriebenen Regeln. Dieser Austausch ermöglicht auch eine unmittelbare Validierung und Korrektur von Informationen, wodurch das Risiko von Fehlinterpretationen und Wissenslücken minimiert wird.

Wurde anfangs die Notwendigkeit einer strategischen Entscheidung zum Umgang mit Erfahrungswissen entweder für eine Personifizierungsstrategie (Orientierung auf persönlichen Wissensaustausch, personenorientierte Methoden) oder einer Kodifizierungsstrategie (technische Verarbeitungsmöglichkeiten für explizites Erfahrungswissen) [4] betont, eröffnen heute neuartige technische Ansätze, wie der Einsatz großer Sprachmodelle, neue Möglichkeiten die Vorteile beider Ansätze zu kombinieren.

2 Potenzial von LLMs und RAG

In den letzten Jahren haben Large Language Models (LLMs) nicht nur in der Forschung, sondern auch in der Industrie erhebliche Aufmerksamkeit gewonnen [5]. Durch ihre Fähigkeit, natürliche Sprache zu verstehen und zu generieren, bieten LLMs ein erhebliches Potenzial für unterschiedlichste Aufgaben wie Zusammenfassungen, Analysen und den kontextuellen Abruf von Informationen.

Hauptursache für diese Fähigkeiten ist die zugrunde liegende Transformer-Architektur, auf der LLMs – genauer gesagt die neuronalen Netze – aufgebaut sind [6]. Durch einen Trainingsprozess mit großen Textkorpora lernen LLMs die Zusammenhänge der jeweiligen Sprache.

Trotz ihres gewaltigen Potenzials mindern sogenannte Halluzinationen erheblich die Zuverlässigkeit von LLMs [7]. Hierbei handelt es sich um generierte Antworten von LLMs, die faktisch falsch sind, beispielsweise aufgrund von Lücken in den Trainingskorpora oder durch Training mit veralteten Daten.

Retrieval-Augmented Generation (RAG) reduziert Halluzinationen, indem es LLMs mit Information-Retrieval-Mechanismen (sprich Suchalgorithmen) kombiniert. Statt sich ausschließlich auf das im Modell gespeicherte Wissen zu stützen, durchsucht ein RAG-System zunächst eine externe Wissensbasis nach relevanten Dokumenten oder Textpassagen. Die relevantesten Dokumente werden anschließend als zusätzlicher Kontext zusammen mit der Nutzereingabe an das LLM übergeben. Somit werden die Qualität und Genauigkeit der generierten Antworten verbessert, und das Wissen der LLMs wird über die Trainingskorpora hinaus ergänzt [7].

3 Systemvoraussetzungen für den Betrieb von LLM- und RAG-Lösungen

Um das in Texten digital gespeicherte Wissen eines Unternehmens zugänglich zu machen, müssen mehrere Randbedingungen, vor allem beim Einsatz in kleinen und mittleren Unternehmen (KMU), eingehalten werden. Im Ergebnis eigener Analysen in Unternehmen werden die folgenden Anforderungen und deren jeweilige Konsequenzen für die Gestaltung eines RAG-Systems formuliert:

1. In Anbetracht der sensiblen Geschäftsinformationen, mit denen ein Unternehmen umgeht, muss das RAG-System den vollständigen Datenschutz gewährleisten und jedes Risiko von Datenlecks vermeiden. Daraus ergibt sich die Anforderung nach einem vollständig lokal lauffähigen System, das nicht auf externe Dienste angewiesen ist, die den Datenschutz gefährden könnten.
2. Dazu muss das System so gestaltet werden, dass die von einem KMU bereitgestellten typischen Hardwareressourcen ausreichen und keine immensen weiteren Kosten für die Systemnutzung anfallen.
3. Im Allgemeinen kann nicht davon ausgegangen werden, dass in den KMU IT-Experten zur Verfügung stehen, sodass sowohl die Einführung als auch die Wartung des Systems so einfach wie möglich sein müssen.
4. Außerdem müssen Benutzeroberfläche und Systemverhalten den Erwartungen der Nutzer und deren Rahmenbedingungen entsprechen, um zur Akzeptanz beizutragen.
5. Das System muss in Eigenregie erweiter- und modifizierbar sein. Daher muss etwa der Code des Systems so geschrieben sein, dass ein externer Entwickler leicht zusätzliche Dateiformate hinzufügen kann. Der zugrunde liegende Wissensbestand muss unkompliziert erweitert und fehlerbehaftete Quellen entfernt werden können.

4 Beschreibung des entwickelten Systems

Um die im Kapitel 3 beschriebenen Systemvoraussetzungen zu erfüllen, die ein Unternehmen zur Inbetriebnahme von LLMs und RAG-Systemen beachten muss, ist das entwickelte System in zwei Komponenten gegliedert.

Die erste Vorverarbeitungskomponente inkludiert den Start des Systems als auch die Aufarbeitung der unternehmensinternen Dateien. Nach einer intuitiven Installation steht das RAG-System zur Verfügung und kann abhängig von der Netzwerkkonfiguration entweder über einen PC oder auf jedem Endgerät im Firmennetz über einen beliebigen Browser abgerufen werden. Zum initialen Start müssen über die Weboberfläche (**Bild 1**, **Bild 2**) zunächst die zu durchsuchenden PDF-Dateien hochgeladen werden.

Nach diesem Schritt werden die Texte der PDF-Dateien geparst, das heißt automatisch analysiert, zerlegt und somit durchsuchbar gemacht. Bislang werden ausschließlich PDF-Datei-

en unterstützt, da dieses Dateiformat in KMU am häufigsten zur Speicherung von Textdaten verwendet wird. Darüber hinaus können Konvertierungen von Formaten wie docx oder pptx leicht ohne manuelle Eingriffe automatisiert werden.

Sobald alle Dateien geparst sind, werden die Texte vorverarbeitet, indem Stoppwörter entfernt und Stemming angewendet wird. Das Entfernen von Stoppwörtern, also von häufigen Wörtern wie „der“ oder „und“, verbessert die Genauigkeit eines Retrievalsystems und erhöht gleichzeitig seine Ausführungsgeschwindigkeit [8]. Das Stemming ist eine Technik, die Wörter auf ihren Wortstamm reduziert, was die Anzahl der unterschiedlichen Begriffe verringert und so die Effizienz und Genauigkeit des Retrievalsystems erhöht [9].

Sind die Dateien einmal hochgeladen und vorverarbeitet worden, so können diese über das System durchsucht werden. Sobald der Benutzer eine Suchanfrage über die Weboberfläche (**Bild 3**) einreicht, werden über die Suchkomponente des Systems die zur Suchanfrage relevantesten Dokumente aus dem hochgeladenen Datenbestand extrahiert.

In der Suchkomponente des Systems werden zunächst konventionelle Retrieval-Ansätze (BM25-Algorithmus [10, 11]) genutzt, um die für die Abfrage relevantesten Dokumente auszuwählen. Ein Dokument wird als umso relevanter eingestuft, je mehr Vorkommen der gesuchten Begriffe es enthält, insbesondere solcher, die im gesamten Korpus selten sind [12]. BM25 ist zwar sehr effizient und effektiv für die begriffsbasierte Suche, hat aber Probleme, die Feinheiten der natürlichen Sprache zu verstehen, wie Synonyme, Umschreibungen und Kontext, der über einfache Schlüsselwortübereinstimmungen hinausgeht.

Cross-Encoder sind leichtgewichtige Sprachmodelle (im Speziellen: svalabs/cross-electra-ms-marco-german-uncased) und ermöglichen, im Gegensatz zu BM25, einen detaillierten Abgleich zwischen der Suchanfrage und den Dokumenten, wodurch sie komplexe semantische Beziehungen erfassen können. Aufgrund der Komplexität und des Rechenaufwands des Cross-Encoders wird er jedoch im vorgestellten System nur auf eine Teilmenge relevanter Dokumente und nicht auf den gesamten Dokumentenkorpus angewendet. Daher wird BM25 verwendet, um die Dokumente auf ihre Relevanz für die Suchanfrage vorzufiltern, wobei lediglich die zehn relevantesten an den Cross-Encoder übergeben werden. Ausgehend von der resultierenden Ergebnismenge von BM25 werden die relevantesten Dokumente in ihre einzelnen Seiten aufgeteilt und zusammen mit der Suchanfrage vom Cross-Encoder verarbeitet. Während dieser Verarbeitung bestimmt der Cross-Encoder die Relevanz jeder einzelnen Seite zur gestellten Suchanfrage und sortiert sie anschließend nach ihrer Relevanz. Durch die Aufteilung der Dokumente in ihre einzelnen Seiten erhält der Benutzer eine detaillierte Ergebnisliste, die genau jene Seiten präsentiert, die am passendsten zur gestellten Anfrage sind.

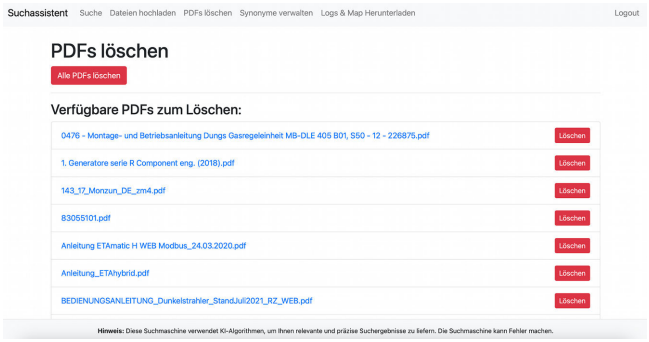

Bild 2. Modifikation und Aktualisierung des RAG-Korpus. *Grafik: TU Dresden*

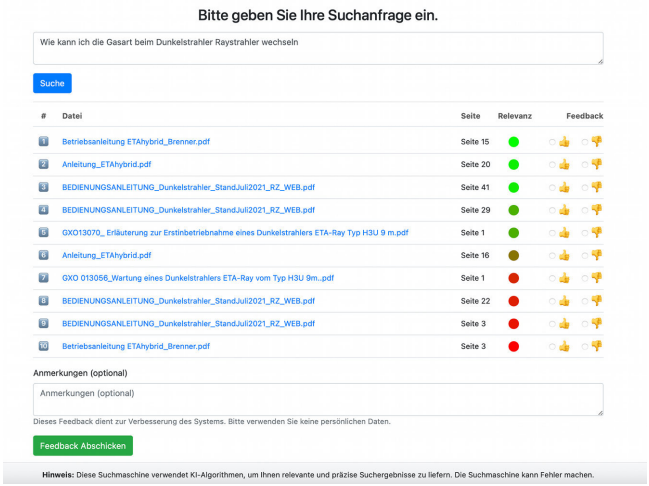

Bild 3. Ergebnisdarstellung mit Feedbackmöglichkeit. *Grafik: TU Dresden*

Auf diese Weise müssen nicht ganze Dokumente, sondern nur einzelne Seiten vom Benutzer gesichtet werden.

In Testläufen bei der Systementwicklung wurde festgestellt, dass die relevantesten zehn Dokumente mit hoher Wahrscheinlichkeit die von der Benutzeranfrage geforderten Informationen enthalten. Der aktuell laufende Test in Unternehmen A (**Tabelle**) bestätigt dies. Da selbst der gewählte ressourcenschonende Cross-Encoder einen beträchtlichen Rechenaufwand erfordert, wird argumentiert, dass die Verarbeitung von zehn Dokumenten ein guter Kompromiss zwischen den Rechenkosten und der Sicherstellung ausreichender Informationen für die korrekte Beantwortung einer Benutzeranfrage ist.

Gemäß dem RAG-Konzept werden anschließend die vom Cross-Encoder als am relevantesten eingestuften Seiten sowie die Suchanfrage an ein großes LLM übergeben, das daraufhin eine vollständig formulierte Antwort generiert. Darüber hinaus zitiert

Tabelle. Charakteristik der Pilotunternehmen.

	Beschäftigtenzahl	Leistungscharakteristik Pilotbereich	Zu lösende Problematik	Getestete Funktionalität RAG-System
Fall A	450	Stationäre Montage- und Inbetriebnahmeprozesse	Kontextbezogene Bereitstellung von Informationen und Erfahrungen	ohne externes Sprachmodell (eigene Instanz eines Sprachmodell in Planung)
Fall B	62	Service- und Wartungsarbeiten beim Kunden		mit externem Sprachmodell

das LLM die Seiten, die für die Generierung jedes Satzes verwendet wurden. Im Hinblick auf die Möglichkeit einer Halluzination des LLM stellt dieser Ansatz sicher, dass der Nutzer den Inhalt der generierten Antwort leicht überprüfen kann.

Der letzte Schritt erfordert erhebliche Hardwareressourcen, damit das LLM eingesetzt werden kann und in angemessener Zeit reagiert. Wenn entsprechende Ressourcen nicht im Unternehmen verfügbar sind, muss ein externes LLM, wie etwa GPT-3, über eine API (application programming interface) abgefragt werden. Handelt es sich um sensible Daten, kann dieser Schritt auf Anforderung des Unternehmens auch deaktiviert werden.

Bild 3 zeigt die Ergebnisse einer natürlich-sprachlichen Anfrage ohne die Ausgabe einer natürlich-sprachlichen Antwort mittels externem Sprachmodell (Schritt 4).

Die identifizierten Seiten werden als Linkliste, geordnet nach ihrer Relevanz für die Anfrage, angezeigt. Außerdem enthält diese Darstellung einen Feedback-Mechanismus, über den die Nutzer die Relevanz einer abgerufenen Seite und die vollständig formulierte Antwort bewerten sowie Kommentare zu den Ergebnissen schreiben können.

Die Interaktionen der Beschäftigten mit dem System werden mitgelogged. Im Logfile werden folgende Informationen erfasst:

- Antwortzeit: Wie lange hat das System zur Antwort gebraucht?
- Dokument-IDs: Welche Dokumente wurden von BM25 ausgewählt?
- Seiten-IDs: Welche Seiten wurden vom Cross-Encoder ausgewählt?
- Suchanfrage: Welche Anfrage hat der Benutzer an das System gestellt?

Zusammen mit dieser Protokollierungsfunktion im Backend erlaubt das Feedback eine iterative Verfeinerung des Systems durch die Analyse des Feedbacks sowie der protokollierten Abfragen und Antworten.

5 Pilothafter Einsatz in zwei Unternehmen

Die Lösungserarbeitung und -erprobung erfolgt gemeinsam mit zwei Unternehmen, die neben dem Dauerthema Steigerung der Effizienz der Auftragsbearbeitung, etwa durch Reduktion des Aufwandes bei der Fehleridentifikation und der Fehlerbehebung, speziell die Sicherung und Nutzung von Erfahrungswissen thematisiert hatten.

Die Komplexität der Problematik erfordert einen ganzheitlichen, interdisziplinären Ansatz für den Implementationsprozess, der gleichermaßen technische, organisatorische und menschliche Faktoren berücksichtigt.

Bei der Umsetzung geht es nicht nur um die nötige technische Befähigung der betrieblichen IT-Administratoren. Vielmehr müssen auch die Unternehmensprozesse angepasst und ein adäquates Commitment aller relevanten Akteure zur Bewahrung von Erfahrungswissen hergestellt werden. Daher erfolgt eine schrittweise, adaptive Implementierung mit kontinuierlicher Evaluation und Anpassung im Rahmen eines interdisziplinären Projektteams aus Unternehmensvertretern und Wissenschaftlern, um die Herausforderungen zu bewältigen.

In den Pilotunternehmen werden jeweils mehrere Akteure unterschiedlicher Hierarchieebenen und Arbeitsfelder nicht nur punktuell, sondern kontinuierlich in die Implementation und Aufrechterhaltung des Systems einbezogen. Neben dem jeweiligen unternehmensinternen Projektmanager und dem IT-Administrator werden Beschäftigte des jeweiligen Pilotbereiches vor allem bei folgenden Fragestellungen involviert:

1. Analyse des Wissensbedarfs
2. Bereitstellung ihres Erfahrungswissen für die Anreicherung der Wissensbasis
3. Feedback zu den Antworten des RAG-Systems als Kuratierungsbasis
4. Erweiterung der domänenspezifischen Sprachbasis

Im Mittelpunkt steht die menschengerechte Gestaltung der Arbeitstätigkeiten [13] und Arbeitsbedingungen.

Bei der Lösungsentwicklung wurde daher unter Einbeziehung der Beschäftigten als Erstes der Gestaltungsbedarf der Tätigkeiten bei eventuellen Belastungen und ungelösten Fragestellungen erfasst. Im Fall B beispielsweise wurde die wiederholte Nachfrage zu gleichen Informationen als Belastung im Innendienst identifiziert, die beseitigt werden soll.

Die Konzeptentwicklung begann jeweils mit der Bedarfsermittlung aus Sicht der Beschäftigten und Führungskräfte im Rahmen von Workshops. Fokussiert wurden die Bestimmung der wesentlichen Informationen und Wissensquellen, die im RAG-System abzubilden sind, und der dafür notwendigen Anwendungsvoraussetzungen.

Einen längeren Bearbeitungszeitraum nahm die Identifikation und Sammlung von strukturierten und unstrukturierten Daten in den Unternehmen für die Wissensbasis des RAG-Systems in Anspruch. Die organisationale Wissensbasis in den Projektunternehmen besteht überwiegend aus digitalen Bedienhandbüchern, Arbeitsanweisungen und Protokollen zu ausgeführten Arbeiten mit deren Ergebnissen sowie umfangreichem, nicht dokumentiertem Erfahrungswissen der ausführenden Beschäftigten. Wie die testweise Verarbeitung von Dokumenten der Projektunternehmen während der Entwicklung des RAG-Systems zeigte, erfordert die automatische Verarbeitung digitaler Dokumente die Einhaltung einiger Regeln bei deren Erstellung. Unternehmensintern entstandene Dokumente müssen deshalb teilweise nachbearbeitet beziehungsweise Ersteller neuer Dokumente zu diesen Regeln geschult werden.

Als relevant haben sich folgende Aspekte herausgestellt:

- Für die erfolgreiche Verarbeitung im RAG-System ist es erforderlich, dass alle digitalen Dokumente in einem maschinenlesbaren Format (pdf-Dateien) vorliegen und der Lesezugriff auf die Dateien gegeben ist, falls es sich um Dateien mit eingeschränktem Benutzerzugriffen handelt.
- In Unternehmen A ist bisher üblich, einen erheblichen Teil der Informationen in Bildern darzustellen, auf denen zum Beispiel bestimmte Teile einer Maschine abgebildet sind. Da das System nicht über einen Mechanismus zur Erkennung von Bildern verfügt, müssen die Bilder mit Text, zum Beispiel über aussagekräftige Abbildungsbeschriftung beschrieben werden, um die darin enthaltenen Informationen durchsuchbar zu machen. Bildschirmschnappschüsse mit enthaltener Abbildungsbeschreibungen sind ungeeignet. Es muss zudem ein räumlicher Bezug von Bild, Bildbeschreibung beziehungsweise Text, der auf das Bild Bezug nimmt, in den Dokumenten gegeben sein.
- Tabellen sind unvorteilhaft für die Verarbeitung. Wenn es nicht zwingend notwendig ist, sollte darauf verzichtet werden.
- Fließtext ist grundsätzlich besser geeignet als Stichworte, da der inhaltliche Zusammenhang besser identifizierbar ist. Es gibt jedoch keine Anforderungen hinsichtlich der Schrift- oder Absatzformatierung.

In beiden Projektunternehmen wird parallel zur systematischen Bereitstellung von digitalen Dokumenten daran gearbeitet, bisher nicht dokumentiertes Erfahrungswissen der Beschäftigten verfügbar zu machen und über das System bereitzustellen.

Dazu wird das Tandemkonzept [14] aufgegriffen und weiterentwickelt. Das Konzept beruht darauf, dass im persönlichen Gespräch (vor Ort oder per Videokonferenz) die Erfahrungen sehr erfahrener, oft älterer Mitarbeiter rückwirkend abgefragt und dokumentiert werden. In den Tandems werden alle Fakten und Informationen zusammengetragen, die zukünftig eine schnellere und sichere Auftragserledigung ermöglichen können. Das heißt, es geht nicht allein um Faktenwissen, sondern auch um Begleitumstände und Intuition bei der Lösungsfindung. Das vorhandene Erfahrungswissen soll jedoch nicht allein auf den Tandempartner übergehen, sondern wird gleichzeitig so dokumentiert, dass es auch anderen Beschäftigten zur Verfügung steht. Dem zweiten Tandempartner kommt dabei eine wichtige Funktion zu. Er fungiert als Korrektiv und stellt sicher, dass die erfassten Informationen vollständig und nachvollziehbar sind.

Zur Unterstützung der Dokumentation werden die Gespräche aufgezeichnet und anschließend automatisch mittels Whisper (Spachmodell zur Spracherkennung und -transkription von Open AI, selbstgehostet) transkribiert, manuell gesichtet und grob bereinigt. Die Bearbeitung und Veredelung der Transkripte bei hoher Performance kann relativ einfach gehalten werden. Eine komplette Ausformulierung und Korrektur des Textes ist nicht erforderlich, da dies die Verarbeitung durch das Sprachmodell in Bezug auf eine konkrete Fragestellung übernimmt.

Für die automatische Verarbeitung dieser speziellen Dokumente, aber auch anderer Dateien im RAG-System sind zwei weitere Aspekte zu beachten. Erstens verwenden Unternehmen oft ein spezialisiertes, domänenspezifisches Vokabular. Einige Begriffe werden selten, wenn überhaupt, in der Alltagssprache verwendet und können als Synonyme oder Abkürzungen auftreten. Selbst wenn die Begriffe A und B Synonyme sind, können herkömmliche Retrievalsysteme sowie auf Sprachmodellen basierende Cross-Encoder diese Dokumente nicht als relevant identifizieren, da diese spezialisierten Begriffe in allgemeinen Trainingsdatensätzen nicht vorkommen. Um domänenspezifische Inhalte durchsuchbar zu machen, wird daher im RAG-System eine Synonymliste (**Bild 4**) bereitgestellt, in der jeder Fachbegriff zusammen mit den entsprechenden Synonymen aufgeführt ist. Auf diese Weise kann die Suchanfrage eines Benutzers vom System mit den entsprechenden Synonymen erweitert werden, sodass das Retrievalsystem an den spezifischen Bereich angepasst werden kann.

Der zweite Aspekt betrifft die inhaltlichen Zusammenhänge der transkribierten Aussagen. Sind diese für den Gesprächspartner aus dem Gesprächsverlauf heraus problemlos nachvollziehbar, fehlen diese dem automatischen System als Kontext. Texte sollten daher in Sätze, Absätze, Überschriften und Abschnitte unterteilt und in einem logischen Zusammenhang stehen, beispielsweise: Problem – Lösung, Baugruppe – zu beachtende Sachverhalte oder Schrittfolgen.

Die dann grob bereinigten und an die oben genannten Regeln angepassten Transkripte werden als Datenquelle ebenfalls in das RAG-System eingepflegt und stehen damit allen Systemnutzern zur Verfügung.

Da die Aufbereitung von Erfahrungswissen in Textform an Grenzen stößt, wenn es sich um schwer explizierbares Wissen handelt, wurden im Projektunternehmen B mit Erfolg Videoaufzeichnungen ausgewählter Arbeitsschritte durch die Beschäftigten angefertigt. Dabei kam eine ActionCam zum Einsatz, die auch unter anspruchsvollen Praxisbedingungen eine akzeptable Videoqualität lieferte und sich durch eine niedrige Hemmschwelle bei den Beschäftigten für eigene Aufzeichnungen auszeichnet. Die Videoclips werden grob mit kostenfreier Software geschnitten. Das heißt, Aufzeichnungs- und Bearbeitungsaufwand sind gering und mit „Bordmitteln“ eines Unternehmens realisierbar. Um die Videos jedoch in der aktuellen RAG-Version zu verarbeiten, ist es erforderlich, zusätzlich ein Transkript des gesprochenen Textes mit Zeitstempeln zu hinterlegen. Mithilfe automatischer Transkriptionssoftware verursacht auch dies nur sehr geringe Zusatzaufwände für die Verarbeitung. Die Videos werden ebenfalls als relevante Fundstellen aufgeführt (Bild 3). Die Zeitstempel im Transkript dienen als Sprungmarken für die konkrete Antwort auf die formulierte Fragestellung im Videoclip.

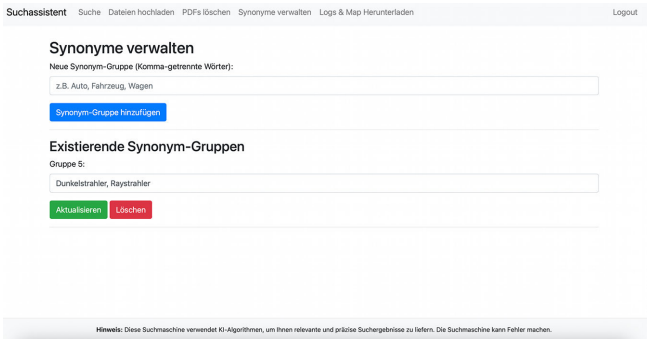

Bild 4. Synonymverwaltung im RAG-System. *Grafik: TU Dresden*

Obwohl aktuell noch Daten in den Projektunternehmen gesammelt werden, um das System und seine Konsequenzen zu bewerten, wurden bereits verschiedene Aufgaben identifiziert, die sich aus seiner Nutzung ergeben (**Bild 5**) und detailliert zu untersuchen sind.

Aufgabe 1: Recherche nach Lösungsvorschlägen und technischen Informationen

Durch das RAG-System steht den Beschäftigten browserbasiert ein Recherchetool zur Verfügung, das auf natürlichsprachliche Anfragen Antworten beziehungsweise Lösungsvorschläge aus qualitativ unterschiedlichen Datenbeständen zur Verfügung stellt. Die Weboberfläche ermöglicht die Eingabe von Suchanfragen und die Anzeige relevanter Ergebnisse sowie ein Feedback zur Qualität der Antworten (Bild 3).

Der Einsatz der vorgeschlagenen Lösung reduziert den Suchaufwand erheblich und erhöht die Treffsicherheit der Antworten gegenüber einer manuellen Suche in Dokumenten oder einer konventionellen Volltextsuche. Das vereinfacht die Nutzung von vorhandenem Wissen und Know-how und erhöht letztendlich die Qualität der ausgeführten Arbeiten beziehungsweise reduziert den Aufwand für eventuell mehrfache Rückfragen erheblich.

Da der RAG-Ansatz auf der Verarbeitung unstrukturierter Textdaten basiert, verlagert sich der Schwerpunkt von der arbeitsintensiven Aufbereitung und Bereinigung auf das kontinuierliche Sammeln von aussagekräftigen Informationen und Erfahrungswissen. Dadurch wird der Prozess der Wissensbewahrung und -kuratierung verbessert und gleichzeitig der in die Arbeit integrierte Lernprozess der Mitarbeiter unterstützt. Insgesamt führt

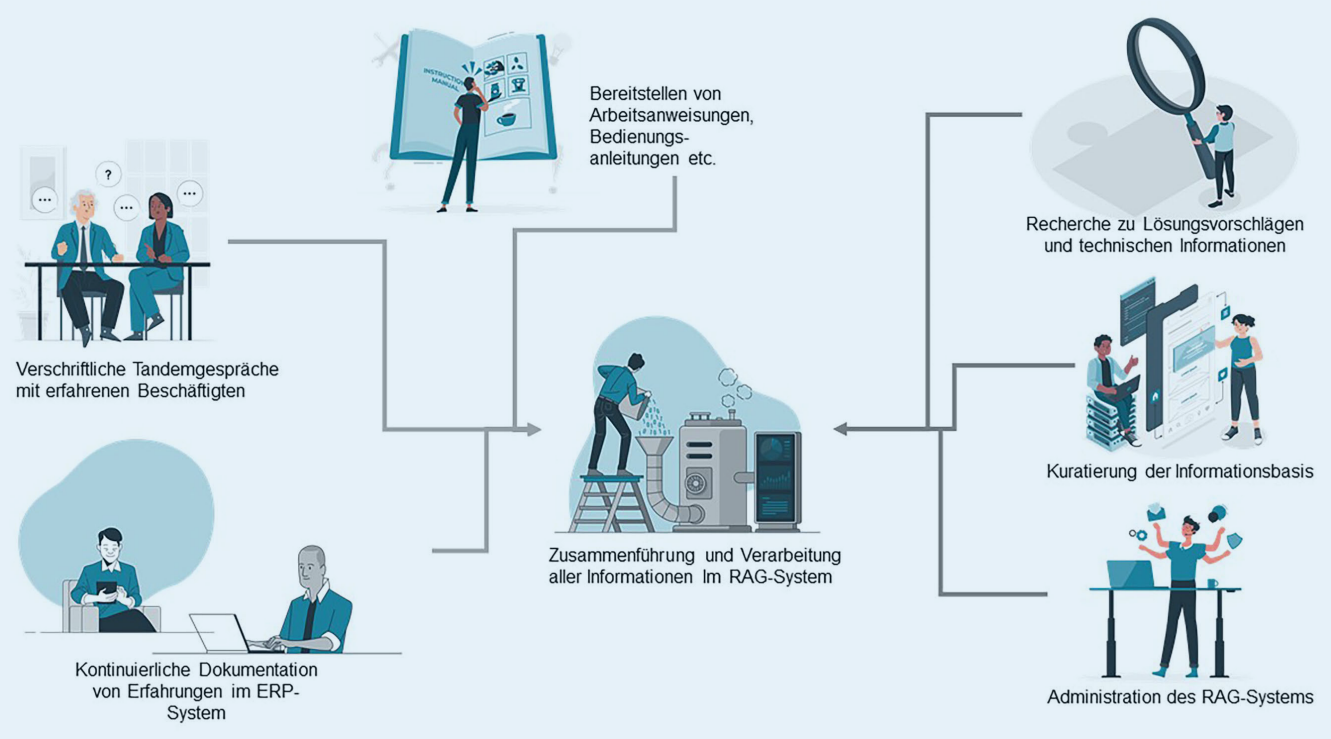

Bild 5. Resultierende Arbeitsaufgaben. *Grafik: TU Dresden, Illustration: https://storyset.com*

dies zu einer Verbesserung der Unternehmensflexibilität, der Leistung und der Attraktivität des Arbeitsplatzes.

Aufgabe 2: Kontinuierliche Weiterentwicklung der Wissensbasis

Die zweite resultierende Arbeitsaufgabe umfasst die kontinuierliche inhaltliche Aktualisierung und Fortschreibung der Wissensdatenbank mit formalen (wie etwa neue Arbeitsanweisungen, Bedienungsanleitungen) und informellen Datenbeständen (wie etwa dokumentierte Tandemgespräche, Notizen aus der Arbeitsausführung). Zahlreiche Beschäftigte eines Unternehmens sind gefordert, immer wieder Ergänzungen für die Wissensdatenbank zu liefern. Dies technisch zu verarbeiten, stellt für das RAG-System nur eine geringe Herausforderung dar. Für die Beschäftigten hingegen bedeutet es einen Mehraufwand, der nur durch die Schaffung einer Win-Win-Situation gerechtfertigt werden kann, beispielsweise durch Zeitersparnis infolge zukünftiger Rückgriffsmöglichkeit auf Erfahrungswissen anderer Beschäftigter.

Aufgabe 3: Kuratierung der Wissensbasis

Die Kuratierung der Wissensbasis, das heißt die Sicherstellung der technischen Richtigkeit, Aktualität, Vollständigkeit, Gültigkeit, Verständlichkeit oder Klarheit der gespeicherten Informationen, kann derzeit nur mit zusätzlichen personellen Ressourcen, unterstützt durch das implementierte Feedbacksystem, geleistet werden. Eine solche Überprüfung kann aber zum Beispiel auch zur Korrektur oder Anpassung bestehender Arbeitsanweisungen führen. Aus Sicht des Unternehmens ergeben sich hieraus also auch positive Effekte, indem Abläufe systematisch überprüft und hinterfragt werden, was sich letztendlich in effektiveren Arbeitsprozessen niederschlagen könnte.

Aufgabe 4: Administration des Systems

Der Aufwand des Administrators, um das System ordnungsgemäß zu betreiben, ist verhältnismäßig gering. Die Installation erfolgt mittels dem Open-Source-Tool „Docker“ Dadurch reduziert sich der Installationsaufwand in einem Unternehmen deutlich und erfordert keine Spezialkenntnisse seitens der IT-Administration. Während die Einrichtung des Systems selbst wenig Herausforderungen mit sich bringt, da die Installation unkompliziert ist und ein Endnutzer über einen einfachen Webbrowser darauf zugreift, kann die Absicherung geeigneter Dokumentenformatierungen und die Pflege der Synonymbasis besonders in der Einführungsphase des Systems relativ aufwendig sein.

Bild 1 und Bild 2 (siehe Kapitel 4) zeigen die dem Nutzer mit Administrationsrechten vorbehaltenen Funktionen zur Pflege des Dokumenten-Korpus.

Ebenfalls für einen Nutzer mit spezifischen Zugriffsrechten ist die Möglichkeit vorgesehen, die unternehmensspezifische Synonymliste zu aktualisieren (Bild 4).

Aufgabe 5: Kontinuierliche Dokumentation von Erfahrungen

Diese Aufgabe hat für eine nachhaltige Wirkung des Konzeptes eine große Bedeutung. Die aktuellen Arbeiten in den Pilotunternehmen konzentrieren sich darauf, die im Einsatz befindlichen ERP- oder Störungserfassungssysteme um Erfassungsmöglichkeiten von auftragsbezogenen Informationen zu erweitern, die dann ebenfalls Eingang in das RAG-System finden.

6 Erkenntnisgewinn und Limitationen

Das Konzept stößt auf großes Interesse in den Unternehmen. Seit April 2025 erfolgt der Testbetrieb im Unternehmen A durch circa 20 Beschäftigte. Die dortige Erfahrungsbasis enthält 81 Dokumente unterschiedlichen Umfangs. **Bild 6** zeigt erste Erfahrun-

gen aus dem Einsatz des Systems. Als leistungsbestimmendes Hardwareelement erweist sich die Grafikkarte.

Eine Sichtung erster Anfragen und Ergebnisse lässt die Vermutung zu, dass der Aufwand zur Beschaffung benötigter Informationen aus Sicht des Beschäftigten deutlich (> 50 %) sinkt. Eine Verifizierung dieser Aussage steht noch aus.

Die Testphase in Unternehmen B beginnt mit Zugriffsmöglichkeit auf ein System mit angebundenem „großen“ Sprachmodell. Dann können auch vom LLM ausformulierte Antworten erprobt und bewertet werden. Auch ohne die Umsetzung dieser letzten Funktionalität ist eine deutliche Vereinfachung der Informationsbeschaffung und ein Beitrag zur Unterstützung systematischer Erfassung von Erfahrungswissen zu erwarten.

Die systematische Erprobung in den Pilotunternehmen umfasst als ersten Schritt den Vergleich von Antworten zu einem gemeinsam formulierten fachspezifischen Fragenpool sowie die Auswertung der Feedbackdatei und der Logdatei. Die Beurteilung der Inhalte wird mit einem Perfomancetest des RAG-Systems gekoppelt. Als weiterer Schritt ist eine Studie zur Nutzung und zur Akzeptanz des Systems durch die Beschäftigten nach einem circa achtwöchigen Nutzungszeitraum geplant. Dazu kommt ein spezifisch entwickelter Fragebogen an die Nutzer zum Einsatz, der unter anderem Usability-Kriterien, die Beurteilung der Antwortqualität, die individuelle Technikaffinität [15] und Veränderungen der Arbeitstätigkeit erfasst.

Vor Abschluss der vollumfänglichen Erprobung des Systems ist noch keine abschließende Beurteilung der Wirksamkeit möglich. Es zeigt sich, dass neben den angesprochenen Voraussetzungen in den Unternehmen wichtige technische Rahmenbedingungen für den Einsatz von KI-Lösungen in den Unternehmen geschaffen werden müssen, wie etwa die Verfügbarkeit geeigneter Hardware oder Datenschutzprotokolle. Dies ist gegenwärtig mit erhöhtem organisatorischen Aufwand bei der Beschaffung verbunden.

7 Ausblick

Dieser Beitrag untersucht die Herausforderungen der Wissensbewahrung in kleinen und mittleren Unternehmen (KMU) und stellt ein Lösungskonzept basierend auf Retrieval-Augmented Generation (RAG) vor, das den Suchaufwand nach kontextbezogenen Informationen im Arbeitsprozess deutlich reduziert und KMU eine effiziente Möglichkeit zur Nutzung ihres vorhandenen Wissens bietet.

Beleuchtet wurden einerseits die Herausforderungen beim Einsatz großer Sprachmodelle (LLMs) in KMU und betont auch die Notwendigkeit, die individuellen Bedürfnisse und Rahmenbedingungen von KMU bei der Implementierung solcher Systeme zu berücksichtigen. Erste Einsatzerfahrungen zeigen Veränderungen in Arbeitsaufgaben auf, die im Ergebnis umfassender Teststellungen verifiziert und hinsichtlich eventuell notwendiger Konsequenzen aus Sicht der Arbeitsforschung bei Kompetenzen und Arbeitsaufgabengestaltung betrachtet werden müssen.

Auch wurden gemeinsam mit den Beschäftigten der beiden Pilotunternehmen die Nutzungsbedingungen kontinuierlich anhand von Tests in den Unternehmen verfeinert und umfassende Erfahrungen bei den Anforderungen an die Dokumentenstrukturierung, die Textverständlichkeit und den Dokumentationsbedarf gesammelt. Daraus werden sich Empfehlungen für die Gestaltung zukünftiger Dokumentationen von Erfahrungswissen ergeben.

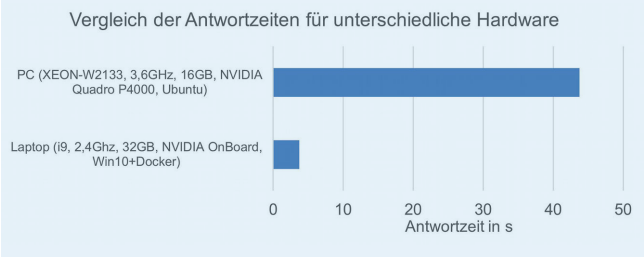

Bild 6. Erfahrungswerte zu Hardwareanforderungen und Leistungsparametern. *Grafik: TU Dresden*

FÖRDERHINWEIS

Das regionale Kompetenzzentrum der Arbeitsforschung „PAL – Perspektive Arbeit Lausitz“ wird vom Bundesministerium für Bildung und Forschung unter dem Förderkennzeichen 02L19C301 gefördert, Projektlaufzeit: 01.11.2021 – 31.10.2026

LITERATUR

[1] Baumhauer, M., Meyer R.: Berufliche Handlungsfähigkeit und Erfahrungswissen: Stellenwert für die Facharbeit in der digitalen Transformation: Eine empirische Analyse am Beispiel der Chemieindustrie. Arbeit 30 (2021) 4, pp. 263–282, doi.org/10.1515/arbeit-2021–0019

[2] Matern, A. (Hrsg.): Wissensmanagement quo vadis? Kuratiertes Dossier in 2 Teilen zum GfWM KnowledgeCamp 2020, Teil 1. Stand: 2020. Internet: www.gfwm.de/wp-content/uploads/2020/10/Kuratiertes-Dossier-gkc20-WM-quo-vadis-Tl1.pdf. Zugriff am 10.06.2025

[3] Plath, H.-E.: Erfahrungswissen und Handlungskompetenz – Konsequenzen für die berufliche Weiterbildung. In: Kleinhenz, G. (Hrsg.): IAB-Kompendium Arbeitsmarkt- und Berufsforschung, 2002. Beiträge zur Arbeitsmarkt-Berufsforschung, BeitrAB 250, S. 517–529

[4] Hinkelmann, K.; Witschel, H. F.: Auswahl der richtigen Wissensmanagement-Methoden. Blickpunkt KMU (2012), doi.org/10.26041/FHNW-2999

[5] Chang, Y. et al.: A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology (TIST), arXiv 2023, doi.org/10.48550/arXiv.2307.03109

[6] Vaswani, A. et al.: Attention Is All You Need 2023. arXiv, doi.org/10.48550/arXiv.1706.03762

[7] Fan, W. et al.: A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. arXiv 2024, doi.org/10.48550/arXiv.2405.06211

[8] Kaur, J., Buttaar, P.: A Systematic Review on Stopword Removal Algorithms. International Journal on Future Revolution in Computer Science & Communication Engineering 4 (2018) 4, pp. 207–210

[9] Flores, F.N.; Moreira, V.P.; Heuser, C.A.: Assessing the Impact of Stemming Accuracy on Information Retrieval. In: Pardo, T.A.S.; Branco, A.; Klautau, A. et al. (eds): Computational Processing of the Portuguese Language. Heidelberg: Springer 2010, p. 11–20, doi.org/10.1007/978–3–642–12320–7_2

[10] Robertson, S.E.; Walker, S.: Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In: Croft, B.W.; van Rijsbergen, C.J. (eds): SIGIR 94. London: Springer 1994, pp. 232–241, doi.org/10.1007/978–1–4471–2099–5_24

[11] Robertson, S.; Walker, S.; Jones, S.; et al.: Okapi at TREC-3. Proceedings TREC of NIST Special Publication 500–225 (1994), pp. 109–126

[12] Robertson, S.; Zaragoza, H.; Taylor, M.: Simple BM25 extension to multiple weighted fields. Proceedings of the thirteenth ACM international conference on Information and knowledge management, NY, USA, 2004, pp. 42–49, doi.org/10.1145/1031171.1031181

[13] DIN EN ISO 6385: Grundsätze der Ergonomie für die Gestaltung von Arbeitssystemen. Deutsche Fassung, Ausgabe 2016

[14] Schulze, D.; Schroth, A.; Lutzmann, H.: Aufgabenbezogener Wissenstransfer durch Tandemarbeit – Leitfaden für kleine und mittlere Unternehmen (SAB, Projektnummer: 080941266). Dresden: 2012

[15] Franke, T.; Attig, C.; Wessel, D.: A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology

Interaction (ATI) Scale. International Journal of Human–Computer Interaction 35 (2019) 6, pp. 456–467, doi. org/10.1080/10447318.2018.1456150

Prof. Martin Schmauder
martin.schmauder@tu-dresden.de
Tel. +49 351 / 463-38510

Dipl.-Ing. Gritt Ott
Technische Universität Dresden
CIMTT Zentrum für Produktionstechnik und Organisation
Helmholtzstr. 10, 01062 Dresden
cimtt.de

Erik Schönwälder, M.Sc.

Dr.-Ing. Martin Hahmann
Technische Universität Dresden
Professur für Datenbanken
Helmholtzstr. 10, 01062 Dresden
tu-dresden.de/ing/informatik/sya/db/