Natural Language Processing (NLP) Unstrukturierte Beschaffungsdaten in saubere Daten umwandeln

Word, Excel, PDF – in all diesen Formaten kommen Beschaffungsdaten daher. Und das ist bei weitem nicht die einzige Herausforderung, denn neben dem Inhalt sind auch Umfang und Strukturierung in jeder Ausschreibung anders. In diesem Beitrag sprechen wir über diese Herausforderungen und wie man diese am besten angeht.

Gewisse Inhalte und damit auch eine gewisse Struktur sind für Ausschreibungen im öffentlichen Beschaffungswesen vorgegeben. Dennoch bedeutet dies nicht, dass man Informationen direkt strukturiert auslesen kann. Einerseits liegen diese oftmals in verschiedenen Dokumenten und Dateiformaten verteilt vor, andererseits werden sie nicht immer auf dieselbe Weise dargestellt. So können Zuschlagskriterien in Tabellen aufgelistet werden, oder sie werden in Paragrafen beschrieben – oftmals auch in beiden Formen gleichzeitig. Für einen Menschen stellt dies keine Herausforderung dar, aber eine automatische Auswertung von tausenden von Dokumenten erschwert sich dadurch erheblich. Glücklicherweise bietet Natural Language Processing (NLP) Möglichkeiten, die Daten in ein strukturiertes Format zu überführen und diese dann im Anschluss automatisch auswerten zu lassen.

Für jede Aufgabe das richtige Modell

Die Extraktion von strukturierter Information aus Beschaffungsdaten bedarf meistens mehrerer Schritte (vgl. «Intelligence im Beschaffungswesen», S. 26). Das Identifizieren dieser Schritte ist unerlässlich, denn oftmals können diese von unterschiedlichen Methoden unterschiedlich gut gelöst werden. Da die Ausschreibungsunterlagen sehr ausführlich sind und hunderte von Seiten umfassen können, bietet es sich beispielsweise an, zuerst die relevanten Seiten zu identifizieren, auf denen sich die gewünschte Information befindet. Diese Aufgabe ist wesentlich einfacher als die Extraktion selbst, weshalb das Trainieren eines eigenen Modelles naheliegt. Der Vorteil ist, dass hier auch kleinere Modelle gut funktionieren, die sehr kostengünstig und vor allem schnell sind (und oft auch klimafreundlicher, vgl. «Nachhaltige KI», S. 62). Damit dies funktioniert, muss jedoch zuerst ein kleiner Datensatz manuell annotiert werden, damit man die Genauigkeit der einzelnen Komponenten des Systems zu jedem Zeitpunkt überprüfen kann. Nur dann ist sichergestellt, dass man für jeden Schritt das beste Modell oder die beste Methode verwendet und schlussendlich saubere Daten erhält.

Reasoning: Wie LLMs bei der Extraktion von Informationen unterstützen

Bei der Extraktion von Informationen sind grosse Sprachmodelle («Large Language Models», kurz LLMs) die ideale Unterstützung. Mittels natürlichsprachlicher Anweisungen – sogenannten Prompts – lassen sich LLMs instruieren, aus einem gegebenen Textauszug Informationen strukturiert zu extrahieren. Dabei ist die Qualität des Prompts entscheidend: So kann man dem Modell Hintergrundwissen zum Beschaffungswesen oder zu einer spezifischen Branche mitgeben, so dass die Wahrscheinlichkeit einer korrekten Extraktion von Informationen erhöht wird. Natürlich kann man nicht

von einer Erfolgsquote von 100 Prozent ausgehen, weshalb man den strukturierten Output noch einmal validieren sollte. Dadurch kann man sich insbesondere vor sogenannten Halluzinationen besser schützen, bei denen LLMs fiktive Informationen generieren. Wer noch einen Schritt weiter gehen will, kann zwei LLMs unterschiedliche Rollen zuweisen und eine Konversation simulieren: Dann diskutiert der Beschaffungsexperte (LLM 1) mit dem Datenanalyst (LLM 2) mit dem Ziel, möglichst korrekte Daten zu extrahieren. Auch multimodale Modelle wie zum Beispiel Vision Language Models (VLMs) können eingesetzt werden, damit nicht nur der Text in Dokumenten, sondern auch darin enthaltene Bilder analysiert werden können.

Unsere Empfehlungen

1. Testdaten kuratieren
Damit geprüft werden kann, wie gut die Umwandlung von unstrukturierten Beschaffungsdaten in ein strukturiertes Format funktioniert, ist es wichtig, einen Teil der Daten manuell in das gewünschte Format zu überführen. Damit können künftige Extraktionsmethoden evaluiert und miteinander verglichen werden.

2. Verschiedene Modelle evaluieren
Nicht jeder Schritt muss ein LLM involvieren. Idealerweise wird die Lokalisierung der Information von der Extraktion getrennt und für jede Aufgabe wird geprüft, welches Modell oder welche Methode die Aufgabe am besten lösen kann.

3. Expertenwissen einfliessen lassen
Für gute Resultate ist eine intelligente Extraktion unabdingbar. Indem Expertenwissen in Prompts einfliesst, können Sprachmodelle die Daten besser verstehen und diese entsprechend strukturieren.

Mehr Informationen

Kontaktmöglichkeiten und weitere Informationen zu Natural Language Processing:
bfh.ch/ipst/nlp

Kontakt

Luca Rolshoven
Doktorand, wissenschaftlicher Mitarbeiter

luca.rolshoven@bfh.ch
T +41 31 848 62 70

Veton Matoshi
Wissenschaftlicher Mitarbeiter

veton.matoshi@bfh.ch
T +41 31 848 57 89