Horticulture Semantic (HortiSem) – Natural Language Processing bei Entwicklung und Interaktion mit einem semantischen Netzwerk für die Landwirtschaft

Jascha Daniló Jung[1], Xia He[2], Daniel Martini[1] und Burkhard Golla[2]

Abstract: Im Projekt HortiSem wird ein semantisches Netzwerk speziell für den Bereich der Landwirtschaft entwickelt. Ein semantisches Netzwerk ist eine Knowledge Base, in der Begriffe und ihre Bedeutung zueinander in Beziehung gesetzt werden. Dies geschieht üblicherweise über eine Triple-Beziehung, konkret über die Relation „Subjekt → Prädikat → Objekt“. Auf diese Weise können große Datenmengen miteinander verknüpft und maschinenlesbar gemacht werden. Neben vorhandenen, strukturierten Daten aus verschiedenen Datenbeständen (BVL Pflanzenschutzmittel, AGROVOC, PS Info) sollen auch neue, semistrukturierte Datensätze und Informationen aus Textkorpora eingepflegt werden. Dazu werden relevante Texte mit Hilfe von Methoden des Natural Language Processing nach bestimmten Klassen (Kulturen, Schädlinge, u.a.) durchsucht und Annotationen in den Knowledge Graphen integriert. Begriffe sollen dabei möglichst automatisch zu anderen Begriffen und bereits vorhandenen Daten in Relation gesetzt werden.

Keywords: Datenmanagement, Smart und Big Data, Künstliche Intelligenz, Machine Learning, Natural Language Processing, Linked Open Data, semantisches Netzwerk

1 Einleitung

Im Projekt Horticulture Semantic (HortiSem)[3] wird ein semantisches Netzwerk speziell für den Bereich der Landwirtschaft entwickelt. Dazu werden Daten aus unterschiedlichen Quellen, meist Datenbanken, ausgelesen und in RDF-Notation [CWL14] transkribiert. Eine weitere Quelle, die nicht auf Datenbanken basiert, sind Texte. Fachtexte enthalten viele Informationen in sprachlicher Form, die nicht mit einfachen Methoden in passende Datensätze umgewandelt werden können. Viele dieser Texte beruhen auf der Abstraktion eines fachkundigen Autors, die zu Grunde liegenden Daten wurden also interpretiert und zueinander in Bezug gesetzt. Diese Daten und Beziehungen sollen möglichst automatisch in das semantische Netzwerk integriert werden.

Zu diesem Zweck bedienen wir uns Methoden des Natural Language Processing (NLP). Mit Hilfe von NLP-Verfahren können große Textmengen automatisch verarbeitet werden

1 Kuratorium für Technik und Bauwesen in der Landwirtschaft, Datenbanken und Wissenstechnologien, Bartningstraße 49, 64289 Darmstadt, j.jung@ktbl.de; d.martini@ktbl.de
2 Julius Kühn-Institut, Institut für Strategien und Folgenabschätzung, Stahnsdorfer Damm 81, 14532 Kleinmachnow, xia.he@julius-kuehn.de; burkhard.golla@julius-kuehn.de
3 HortiSem, https://hortisem.de

und für das semantische Netzwerk nutzbar gemacht werden. Die Daten und Informationen der Texte werden kontextualisiert, eingeordnet und somit maschinenlesbar gemacht. Neben der Gewinnung von neuen Daten für das semantische Netzwerk sollen auch die Texte selbst automatisch verschlagwortet und innerhalb des Netzes suchbar gemacht werden.

2 Wissensstand

2.1 Semantische Netzwerke

Semantische Netzwerke erlauben es, Daten zueinander in Beziehung zu setzen und maschinenlesbar zu machen. Damit können die zugrunde liegenden Informationen automatisiert und schnell abgerufen werden. Daten, die bereits in verschiedenen Datenbanken vorliegen, werden so zu einem großen Netzwerk zusammengefasst. Dieses Netzwerk kann dann mit zusätzlichen neuen Daten erweitert und Relationen zwischen den verschiedenen Daten können beschrieben werden.

2.2 NLP in der Landwirtschaft

Es ist zunächst nicht offensichtlich, dass der Einsatz von NLP im landwirtschaftlichen Bereich Sinn ergeben kann. NLP beschäftigt sich mit der automatischen Verarbeitung von linguistischen, also sprachlichen Daten und hat damit eine große Schnittmenge mit den Sprachwissenschaften, die zu den Geisteswissenschaften gehören. Damit stehen sich die eher theoretischen Überlegungen der Sprachwissenschaften und der sehr praxisorientierte Bereich der Landwirtschaft gegenüber. Dieser Widerspruch lässt sich auflösen, indem man Sprache als Informationsmedium definiert, das Daten und Informationen enthält. Auch im praktischen Bereich der Landwirtschaft erfolgen der Informationsaustausch und die Forschung über sprachliche Mittel. Computerlinguistische Verfahren sind also in allen Bereichen anwendbar, in denen Informationen über Sprache ausgetauscht werden.

2.3 NLP State-of-the-art

Bei der Named Entity Recognition (NER) handelt es sich um ein typisches Klassifizierungsproblem im Machine Learning Bereich. Es wird ein Modell entwickelt, das automatisch bestimmte Begriffe erkennen und klassifizieren, also einer bestimmten Kategorie wie „Erreger“ oder „Kultur“, zuordnen soll. Klassifizierungsprobleme wie NER werden üblicherweise nach Metriken wie Genauigkeit, Recall und Präzision bewertet. Diese Werte beschreiben das Verhältnis von falsch positiven, falsch negativen, richtig positiven und richtig negativen Ergebnissen. Zudem gibt es Mischwerte, die eine gewichtete Bewertung aus mehreren dieser Metriken darstellen. Bei NLP-Anwendungen

wird oft der F_1-Score verwendet, der das harmonische Mittel aus Präzision und Recall beschreibt.

Der State-of-the-art im Bereich NER ist in den letzten Jahren stetig gestiegen und liegt inzwischen bei vielen Verfahren bei einem F_1-Score von über 90 % (siehe etwa [SSH21; Wa21; Ya20])[4]. Das heißt, dass im Schnitt über 90 % der gesuchten Klassen gefunden und korrekt klassifiziert werden.

Neben den Verfahren trägt die Qualität der Trainingsdaten für Machine-Learning-Verfahren maßgeblich zur abschließenden Performance der Anwendung bei. Trainingsdaten müssen daher besonders sorgfältig ausgewählt und erstellt werden.

3 Methoden

Bei der Arbeit am Projekt HortiSem hat sich ein Workflow herauskristallisiert, der im Folgenden dargestellt wird.

3.1 Erstellung des Korpus, Datenquellen, Annotation

Für die Erstellung eines NER-Verfahrens musste zunächst eine Textsammlung relevanter Texte gesammelt werden. Insbesondere die Pflanzenschutz-Informationstexte der Warndienste der Beratungseinrichtungen der Länder sollen später automatisch auf relevante Begriffe analysiert und verschlagwortet werden können, um diese dann in das semantische Netzwerk einzupflegen. Relevante Klassen sind dabei insbesondere Kulturen, Schädlinge, BBCH-Stadien, Pflanzenschutzmittel sowie Maschinen.

3.2 Named Entity Recognition mit spaCy

Für die Erstellung eines NER-Modells zur automatischen Klassifizierung von Begriffen wird zunächst eine möglichst große Menge an Trainingsdaten und Testdaten benötigt. Diese müssen zunächst von Hand annotiert werden. Anschließend werden die annotierten Daten als Input für die Erstellung des NER-Modells mit spaCy[5] [HM17] verwendet.

Für die manuelle Annotation wurde prodigy[6] verwendet, das auch von den spaCy-Entwicklern herausgegeben wird. Dadurch wird die Kompatibilität von Trainingsdaten und NER-Verfahren sichergestellt. spaCy ermöglicht es, einen Teil der annotierten Texte automatisch als Trainingsdaten zu verwenden, um das Modell zu trainieren. Der restliche

4 s.a. die folgende Übersicht über relevante wissenschaftliche Veröffentlichungen auf der Webseite NLP-progress, http://nlpprogress.com/english/named_entity_recognition.html, Stand: 28.10.2021.
5 spaCy Homepage, https://spacy.io
6 prodigy Homepage, https://prodi.gy

Teil kann dann zur automatischen Evaluation des Modells verwendet werden, um die Performance bewerten zu können.

3.3 Relation Extraction

Die automatische Relation Extraction (RE) ist ein sehr aufwändiges Verfahren, um die Beziehung von Worten untereinander zu beschreiben. Diese Verfahren wollen wir in Zukunft verwenden, um Vorschläge für relationale Verhältnisse von Konzepten im semantischen Netzwerk zu erhalten. Denkbar ist etwa eine semantische Nähe zweier Worte über Wortvektoren erkennen zu können. Daneben wurden aber auch linguistische Merkmale verwendet, um solche Beziehungen zu finden. Ist etwa Wort 1 in Wort 2 enthalten (Beispiel: „Häcksler“ und „Feldhäcksler“), so kann es sich um eine Beziehung des Typs „Oberbegriff“ und „Unterbegriff“ handeln.

3.4 Einpflegen in das semantische Netzwerk

Für das semantische Netzwerk HortiSem wird die RDF-Notation Turtle verwendet. Dieser können verschiedene Spezifikationen (Ontologien) hinzugefügt werden, die die genaue Beschreibung der Daten mit den relevanten Informationen und Verhältnissen zu anderen Knoten im Netzwerk auf die jeweilige Fachdomäne zugeschnitten ermöglichen.

Das NER-Modell kann also auf beliebige Texte angewendet werden, um Kandidaten für das semantische Netzwerk zu gewinnen. Es kann dann überprüft werden, ob diese Begriffe bereits im Netzwerk vorliegen. Ist dies nicht der Fall, der Begriff aber für das Netzwerk relevant, so kann eine entsprechende Turtle Notation generiert und in das Netzwerk eingepflegt werden. Bei bereits bestehenden Knoten können gegebenenfalls neue Informationen und Relationen hinzugefügt werden, sofern diese relevant sind. Darüber hinaus kann der Text, auf den das Modell angewandt wurde, selbst verschlagwortet und per URI/URL dem semantischen Netzwerk hinzugefügt werden. Der Benutzer kann damit beispielsweise relevante Texte zu seiner Suche finden. Dabei können auch Links zu Texten, die hinter einer Bezahlschranke liegen und für deren Zugriff der Benutzer Geld bezahlen muss, hinterlegt und diese damit auffindbar gemacht werden.

4 Ergebnisse

4.1 F_1-Score

Während der Entwicklung wurden mehrere Modelle entwickelt und ausprobiert. Auch wurden mehrere Trainingssets erstellt und evaluiert. Für Modelle mit vielen unterschiedlichen Klassen (Schädlinge, Erreger, Kulturen …) wurden zunächst F_1-Werte zwischen 70 % und 80 % erzielt. Der Fokus lag allerdings im bisherigen Verlauf des

Projekts noch nicht auf einer Optimierung der Trainingsdaten und des Modells, da zunächst der gesamte Arbeitsschritt, von den unbearbeiteten Fachtexten bis zur Einpflegung in das semantische Netz, entwickelt wurde. Eine nachträgliche Verbesserung und Optimierung der NER-Modelle ist unproblematisch und kann jederzeit im Nachgang erfolgen.

4.2 Relation Extraction

Die Relation Extraction erfolgt zunächst auf linguistischer Ebene, soll aber auch mit Machine-Learning-Verfahren erfolgen. Denkbar sind beispielsweise Wordvektorverfahren (Word Embeddings), bei denen ähnliche Wörter eine geringe Distanz in einem n-dimensionalen Vektorraum haben. Die Generierung solcher Vektoren erfordert allerdings eine sehr große Menge an Textdaten, um eine zufriedenstellende Genauigkeit zu erreichen [YGD15].

Eine weitere Möglichkeit ist der Rückgriff auf bereits bekannte Knowledge Bases. Hier kann man Beziehungen zwischen Daten teilweise automatisch abrufen.

4.3 Vorteile für die Landwirtschaft

HortiSem bietet ein durchsuchbares, maschinenlesbares Netzwerk, das direkten Zugriff auf landwirtschaftliche Informationen und Daten enthält und diese in Beziehung zueinander setzt. Es basiert zu Teilen auf Texten, die für herkömmliche Suchmaschinen nicht zugänglich sind, und ist speziell für den deutschsprachigen Bereich entwickelt. Aus Fachtexten kann ein Mehrwert an Informationen gewonnen werden, die für den Benutzer direkt abrufbar sind. Bei der Suche nach relevanten Texten zu einem Thema oder Begriff kann sich für den Benutzer eine große Zeitersparnis ergeben. Für Landwirte ergibt sich damit eine zentrale Anlaufstelle für alle aktuellen und vergangene Warndienstmeldungen und Pflanzenschutzinformationen sowie ein breites Spektrum an kontextualisierten Informationen.

Denkbar sind zudem die Integration weiterer Verfahren, etwa bei der automatischen Bilderkennung. Zukünftig könnten etwa auch Bilder automatisch annotiert und in den Knowledge Graphen eingefügt werden.

4.4 Automatisierung des Ablaufs

Da das Netzwerk nicht unmittelbar aktualisiert wird und die Menge der Input-Texte überschaubar ist, können die Ergebnisse der hier vorgestellten Verfahren durchaus manuell überprüft werden. Es ist nicht nötig, alle Ergebnisse automatisch dem semantischen Netzwerk hinzuzufügen und damit zu riskieren, fehlerhafte Daten einzupflegen. Ebenso können Vorschläge für Relationen von Daten automatisch generiert

werden, die Überprüfung auf Korrektheit, oder zumindest auf Plausibilität, kann dennoch von Fachkundigen übernommen werden.

Die Verschlagwortung von Fachtexten und das Einpflegen von URI/URLs kann dagegen vollautomatisch erfolgen, da es hier bereits bewährte Methoden gibt.

5 Fazit

Das im Projekt HortiSem gewählte Vorgehen hat gezeigt, dass moderne Machine-Learning-Verfahren im Bereich NLP in traditionellen Bereichen wie der Landwirtschaft produktiv eingesetzt werden können. Auch die Landwirtschaft profitiert von der Digitalisierung und Automatisierung von Informationsflüssen. Informationen können gezielter gesucht und in einen fachspezifischen Kontext gesetzt werden. Der Informationsfluss wird also optimiert, insofern die gesuchten Informationen sowie weitere im Kontext stehende Informationen zentral abrufbar werden. So wird die Informationssuche vereinfacht und geht schneller, die Benutzer können Zeit und Geld einsparen. Zudem können Informationen gefunden werden, die für die Fragestellung des Benutzers relevant sind, ohne dass gezielt nach diesen gesucht wurde.

Eine vollständige Automatisierung des vorgestellten Prozesses ist gegenwärtig allerdings nicht ohne Qualitätseinbußen möglich. Aufgrund der bislang hier überschaubaren Datenmenge ist dies allerdings auch derzeit nicht unbedingt nötig.

Literaturverzeichnis

[CWL14] Cyganiak, R.; Wood, D.; Lanthaler, M.: RDF 1.1 Concepts and Abstract Syntax. World Wide Web Consortium (W3C). http://www.w3.org/TR/rdf11-concepts/, Stand: 29.10.2021.

[HM17] Honnibal, M.; Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[SSH21] Straková, J.; Straka, M.; Hajic, J.: Neural Architectures for Nested NER through Linearization. 2021.

[Wa21] Wang, X.; Jiang, Y.; Bach, N.; Wang, T.; Huang, Z.; Huang, F.; Tu, K.: Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning. 2021.

[Ya20] Yamada, I.; Asai, A.; Shindo, H.; Takeda, H.; Matsumoto, Y.: LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. 2020.

[YGD15] Yu, M.; Gormley, M. R.; Dredze, M.: Combining Word Embeddings and Feature Embeddings for Fine-grained Relation Extraction. 2015.