Named Entity Recognition (NER) von Warndienstmeldungen im Gartenbau: Eine empirische Studie zu Design, Entwicklung und Bewertung der statistischen und Deep-Learning benutzerdefinierten NER-Modelle

Named Entity Recognition (NER) of warning messages in gardening: An empirical study of the design, development, and evaluation of statistic and deep-learning user-defined NER models

Xia He-Bleinagel[1*], Jascha Daniló Jung[2], Burkhard Golla[1]

[1]Julius Kühn-Institut (JKI), Bundesforschungsinstitut für Kulturpflanzen, Institut für Strategien und Folgenabschätzung, Stahnsdorfer Damm 81, 14532 Kleinmachnow

[2]Kuratorium für Technik und Bauwesen in der Landwirtschaft e.V. (KTBL), Team Datenbanken und Wissenstechnologie, Bartningstraße 49, 64289 Darmstadt

*xia.he@julius-kuehn.de

DOI: 10.5073/20220124-072443

Zusammenfassung

Anhand gesammelter Pflanzenschutzhinweise und Warnmeldungen wurden Named Entity Recognition (NER) Modelle zur automatischen Erkennung und Klassifizierung von relevanten Begriffen des Gartenbaus (Kulturen, Schaderreger, Pflanzenschutzmittel, BBCH Stadium) erstellt. NER ist eine Teilaufgabe der Informationsextraktion, die darauf abzielt, benannte Entitäten, die in unstrukturiertem Text erwähnt werden, zu finden und in vordefinierte Kategorien einzuordnen. 114 Dateien mit 105737 Wörtern, davon 12295 verschiedene Wörter, wurden verwendet. Mit dem Annotationswerkzeug Prodigy wurden insgesamt 9019 Entitäten annotiert. Es wurden drei verschiedene Modelle trainiert, basierend auf spaCy, Flair und CRF. Alle drei Modelle erzielten ähnlich gute Genauigkeiten (gemittelte F-Werte), spaCy erreichte mit F1=0.8997 die höchste Genauigkeit über alle vier Klassen. Mit dem hier vorgestellten Projekt werden die Voraussetzungen geschaffen, die Inhalte der Vielzahl von Warndienstmeldungen automatisiert zu erschließen und über gezielte Abfragen und Suchvorgänge zugänglich zu machen. Der Beitrag stellt das methodische Vorgehen und einige Analyseergebnisse beispielhaft dar.

Stichwörter: Maschinelles Lernen, NER, NLP, Prodigy, spaCy, tiefes Lernen

Abstract

We collected a relative small dataset in German language from the warning service of different federal states for Named Entity Recognition (NER) in the Horticulture domain. NER is a subtask of information extraction aimed at finding named entities mentioned in unstructured text and classifying them into predefined categories. It consists 114 pdf files which include 105737 tokens in total and 12295 unique tokens. We used the annotation tool Prodigy to label a total of 9019 entities associated with the 4 semantic classes: Horticulture, Pests, Pesticides, BBCH stage. We built the custom NER models with conditional random field (CRF) statistical model, spaCy model and Flair model which are deep learning neural networks frameworks. All three models have similar performances regarding micro-averaged F1 score, among which spaCy model stands out with F1=0.8997 across all four classes. The project presented here creates the prerequisites for automatically extracting the information of the large number of warning service messages and making them accessible via targeted queries and searches. The article presents the methodical procedure and some analysis results as examples.

Keywords: deep learning, machine learning, NER, NLP, Prodigy, spaCy

Einleitung

Verfahren des Maschinellen Lernens erfreuen sich zunehmender Beliebtheit in vielen wirtschaftlichen Bereichen. Auch in der Landwirtschaft können solche Verfahren produktiv eingesetzt werden. Beispielsweise werden Modelle zur automatischen Erkennung von Schädlingen auf Bildern entwickelt (HOYE et al., 2021). In dieser Arbeit wenden wir Verfahren des Maschinellen Lernens auf Textdaten an, um nützliche Informationen zu gewinnen.

Pflanzenschutzhinweise (oder Warndienstmeldungen) stellen die aktuelle phytosanitäre Situation dar und unterstützen die landwirtschaftliche Praxis bei der Entscheidungsfindung von Maßnahmen, etwa der Wahl eines optimalen Bekämpfungstermins sowie beim sach- und umweltgerechten Einsatz von Pflanzenschutzmitteln. Auch für retrospektive Auswertungen enthalten sie wertvolle Informationen, die im Kontext weitere historischer Umweltdaten (z.B. Klima) zu neuen Erkenntnissen beitragen können.

Das Übertragungsformat der Warndienstmeldungen ist deutschlandweit sehr unterschiedlich. In der Regel werden die Hinweise über Webportale (z.B. https://www.ISIP.de), per Email, SMS oder Fax dem Landwirt übermittelt. Auch die inhaltliche und strukturelle Ausgestaltung der Warndienstmeldungen kann von Land zu Land unterschiedlich sein. Gemeinsam ist den Warnhinweisen der Bezug zu einem geographischen Raum (z.B. Beratungsbezirk), einer Kultur, einem Schaderreger und einer Bekämpfungsmaßnahme zu einem bestimmten Zeitpunkt bzw. Stadium der Vegetationsperiode.

Im Rahmen des Projektes HortiSem arbeiten wir daran bestimmte relevante Entitäten - etwa „Kultur“, „Schaderreger“, „Pflanzenschutzmittel“, „BBCH-Stadium“ - mittels Named Entity Recognition in Texten ausfindig zu machen, zu klassifizieren und in einen „Knowledge Graphen“ zu integrieren. Knowledge Graphen bilden dabei die Datengrundlage eines sog. semantischen Netzes (Semantic Web Technologie). Im semantischen Netz werden Entitäten zueinander in Beziehung gesetzt, sodass sie maschinenlesbar sind und die (automatische) rechnergestützte Datenverarbeitung ermöglicht wird. Das semantische Netz soll im Pflanzenschutz Informationssystem (https://www.pflanzenschutz-information.de/) für gezielte Abfragen von Benutzern integriert werden.

Material und Methoden

Workflow des Versuchs

Abbildung 1 zeigt unseren Workflow bei der Anwendung des maschinellen Lernen-Verfahren. Die Textdaten aus den pdf-Dateien werden in einem ersten Schritt in eine für Prodigy (MONTANI et al., 2018) geeignete Textform konvertiert. Danach können fachkundige Annotatoren die Texte satzweise annotieren (Vergabe von NER Tags für bestimmte Wendungen und Wörter), um Trainings- und Testdatensätze für das NER-Modell zu erstellen. Zum einen wurde ein klassisches Conditional Random Field (LAFFERTY et al., 2001) Modell erstellt, zum anderen zwei Modelle mit spaCy (HONNIBAL et al., 2017) und Flair (AKBIK et al., 2019), die auf Deep Learning-Verfahren basieren.

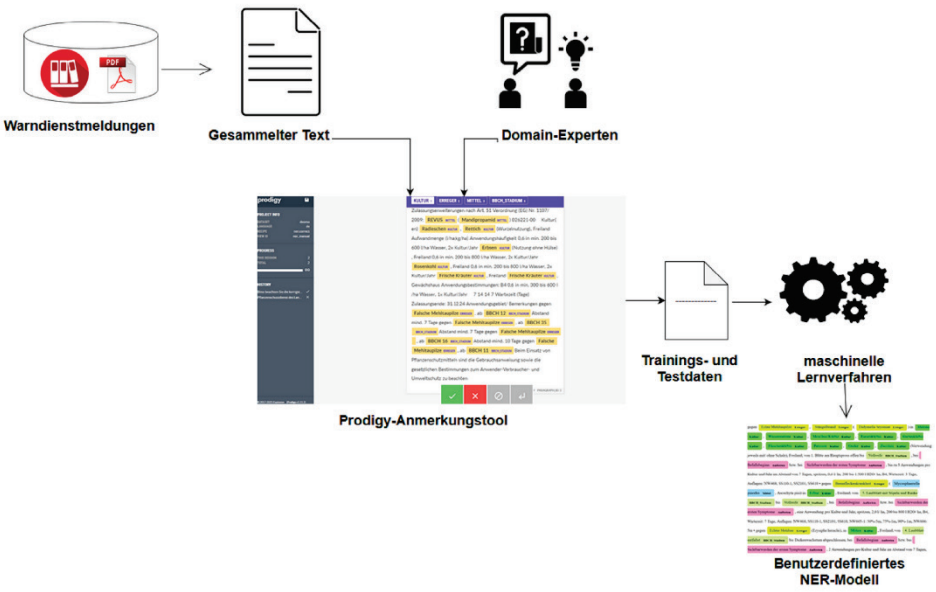

Abbildung 1 Flowchart der Entwicklung eines NER-Modells.

Figure 1 *Flowchart of the NER model development.*

Datenquelle

Dem Julius-Kühn-Institut liegen mehr als 8000 Warnmeldungen vor, die von 1994 bis heute von den Beratungsdiensten der Länder herausgegeben wurden. Diese beinhaltenen Informationen über Flächenkulturen, aber auch Gemüse, Früchte- und Zierpflanzen aus 11 Bundesländern. Für die Erstellung der NER-Modelle wurden die Veröffentlichungen aus den Ländern Brandenburg, Baden-Württemberg, Sachsen und Thüringen verwendet. Zwar sind möglichst viele Daten für maschinelles Lernen immer vorteilhaft, jedoch ist die manuelle Annotation sehr zeitaufwändig und damit auch teuer. Für die hier durchgeführte Analyse wurden daher nur die Warnmeldungen des letzten vollständigen Jahres als Trainigsdaten zur Modellerstellung annotiert sowie einige neuere Texte als Testdaten zur Validierung des Modells.

Klassifizierung und Annotation

Es erfolgte die Aufteilung in vier Klassen in Absprache mit fachkundigen Annotatoren.

Tabelle 1 Beispiele der vier NER-Klassen

Table 1 *Examples of the four NER classes*

NER Tags	Beschreibung	Beispiele
Kultur	Namen der Kultur	Kartoffeln, Gurken
Erreger	Namen von Schädlingen, Pathogene, Unkräuter	Blattläuse, Mehltau, Kamille
Mittel	Pflanzenschutzmittel	Bandur, Cato
BBCH-Stadium	BBCH-Codes, Beschreibungen von morphologischen Entwicklungsstadien von Pflanzen	BBCH 12, Vollblüte

Erstellung und Annotation des Korpus

Prodigy ist ein webbasiertes Annotationswerkzeug für die Erstellung von Trainings- und Testdaten von maschinellen Lernen-Modellen. Es ermöglicht eine schnelle und effiziente Annotation mittels vorgegebener „Rezepte“. Für die ersten Annotationen wurde das „ma-nual.ner“ Rezept verwendet um die vier Klassen in den vorbereiteten Rohdaten zu annotieren. Zusätzlich wurde die „Pattern“-Funktion verwendet, um mögliche Kandidaten von Prodigy vorschlagen zu lassen. Als Pattern wurden dabei Begriffe aus der BVL

Datenbank für Pflanzenschutzmittel verwendet. Ein Pattern sieht beispielsweise wie folgt aus: {"label": "Mittel", "pattern": "Trimangol Flowable"}. Insgesamt wurden 28 088 Pattern für die vier Klassen verwendet.

Für den Trainingsdatensatz wurden 102 pdf-Dateien annotiert, die aus insgesamt 97 315 Token (d.h. Wörtern und Satzzeichen) bestanden. Davon wurden 80 % als Trainingsdaten und die restlichen 20 % als Testdaten verwendet. Der Testdatensatz besteht daher aus 12 Dokumenten für die Auswertung der Modelle.

Tabelle 2 Die Verteilung der Anzahl der annotierten Token in den vier Klassen

Table 2 *Distribution of the number of annotated token in the four classes*

NER Tags	**Train**	**Test**
Kultur	2835	244
Mittel	1994	162
Erreger	3074	285
BBCH-Stadium	408	17
Anzahl der Dokumente	102	12
Gesamtzahl der Token	97315	8422
Gesamtzahl der eindeutigen Token	9913	2382

Maschinelles Lernen Methoden

1) spaCy Framework

Bei spaCy handelt es sich um eine open-source library für Natural Language Processing. Die spaCy library ist für schnelle Berechnung mit Computerprozessoren optimiert. Da spaCy und Prodigy von den gleichen Entwicklern stammen, lassen sich Prodigy Datensätze sehr leicht mit spaCy verarbeiten.

2) Flair Framework

Flair ist ein mächtiges und einfaches Framework für state-of-the-art NLP Anwendungen, entwickelt von der Humboldt Universität Berlin. Die Stärke des Frameworks liegt vorwiegend in der Verwendung von Worteinbettungen. Bei der Worteinbettung werden Wörtern und Symbolen mehrere Vektoren zugeordnet. Flair verfügt nicht nur über – wie sonst üblich – einen Vektor pro Wort, sondern passt die Anzahl der Vektoren abhängig vom Kontext an. Für die Erstellung des Flair Modells wurden darüber hinaus links- und rechtsläufige Embeddings verwendet.

3) CRF Modell

Beim Conditional Random Field (CRF) handelt es sich um ein ungerichtetes, probabilistisches Modell. Es wird zur Segmentierung von Sequenzen verwendet, um aus einer Eingabesequenz X die gewünschte Sequenz Y zu berechnen.

Formel 1 Definition von CRF

Formula 1 *Definition of CRF*

$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} exp\left(\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t) \right)$$

Z(x) ist die Normalisierungskonstante, θ_k ist eine Gewichtung und f_k (y_t,y_(t-1),x_t) ist eine Funktion. Die Funktion betrachtet dabei nicht nur den aktuellen Input, sondern auch den Output des vorherigen Durchlaufs. Ein Schlüsselschritt ist die Erzeugung der Funktionsmerkmale. In unserem Versuch wurden Wortidentität, Wortsuffixe, Wortform, Wortart und Informationen über den Kontext berücksichtig.

Auswertung

Bei NER handelt es sich um ein Klassifizierungsproblem. Die sequentiell vorliegenden Token eines Textes müssen richtig annotiert werden. Um die Ergebnisse der Modelle zu evaluieren und vergleichbar zu machen, werden Präzision, Recall und F1-Wert berechnet. Präzision ist das Verhältnis der korrekt identifizierten positiven Fälle zu allen vorhergesagten positiven Fällen. Recall dagegen ist das Verhältnis der korrekt identifizierten positiven Fälle zu allen tatsächlichen positiven Fällen. Der F1-Wert bezeichnet das harmonische Mittel von Präzision und Recall und beschreibt so die Gesamtgüte des Modells.

Formel 2 Berechnung von Präzision, Recall und F1-Wert (TP = True Positive, FP = False Positive, FN = False Negative)

Formula 2 *Calculation of precision, recall and F1 value (TP = True Positive, FP = False Positive, FN = False Negative)*

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Ergebnisse

Tabelle 3 zeigt die gemittelten F1-Werte. spaCy erzielte hier mit geringem Abstand den besten Wert von etwa 0.9. Für den verhältnismäßig kleinen Datensatz ist dies ein sehr gutes Ergebnis.

Tabelle 3 Zusammenfassung der gemittelten Ergebnisse aller drei Modelle

Table 3 *Summary of the averaged results of all three models*

Model	**Precision**	**Recall**	**F1**
Spacy	0.8947	**0.9048**	**0.8997**
Flair	0.8524	0.8978	0.8745
CRF	**0.905**	0.87	0.887

In Tabelle 4 sind die Ergebnisse nach den vier Klassen aufgeschlüsselt. Für die Klassen „Erreger“ und „Kultur“ erzielte spaCy das beste Ergebnis. Bei der Klasse „Mittel“ erzielte Flair das beste Ergebnis, das CRF erzielte dagegen beim BBCH-Stadium das beste Ergebnis.

Tabelle 4 Zusammenfassung der gemittelten Ergebnisse aller drei Modelle

Table 4 *Summary of the averaged results of all three models*

Model	Tag	Precision	Recall	F1
Spacy3	Erreger	0.8697	0.9368	0.902
	Kultur	0.9194	0.9344	0.9268
	Mittel	0.9189	0.8395	0.8774
	BBCH-Stadium	0.7895	0.6522	0.7143
Flair	Erreger	0.8182	0.9158	0.8642
	Kultur	0.8539	0.9344	0.8924
	Mittel	0.931	0.8333	0.8795
	BBCH-Stadium	0.8095	0.7391	0.7727
CRF	Erreger	0.905	0.8145	0.8488
	Kultur	0.9783	0.7493	0.8418
	Mittel	0.904	0.8405	0.8693
	BBCH-Stadium	0.8345	0.8713	0.8483

Fazit

In dieser Arbeit wurde ein NER Modell für den Gartenbau entwickelt. Es wurden verschiedene Arten entwickelt und evaluiert. Für den verhältnismäßig kleinen Datensatz haben die Modelle sehr gute Ergebnisse erzielt.

Für die Zukunft sind eine genauere Unterscheidung der Klassen sowie ein größerer Trainigsdatensatz geplant. Zudem sollen die Modelle auf Texte aus anderen Bereichen der Landwirtschaft, etwa dem Ackerbau, und auf die Meldungen anderer Bundesländer oder des Bundes angewandt werden.

Mit Hilfe dieses NER Models werden pflanzenschutzrelevante Information in Zukunft automatisch aus textlichen Quellen gewonnen.

Danksagung

Diese Arbeit wurde im Rahmen des Projekts „HortiSem“ durchgeführt, welches vom Bundesministerium für Ernährung und Landwirtschaft (BMEL) gefördert wird. Weitere Informationen finden Sie auf https://hortisem.de/

Literatur

AKBIK, A., T. BERGMANN, D.A. BLYTHE, K. RASUL, S. SCHWETER, R. VOLLGRAF, 2019: FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), p.54-59.

HONNIBAL, M., I. MONTANI, 2017: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, to appear.

HOYE T.T., J. ÄRJE, K. BJERGE, O.L.P. HANSEN, A. IOSIFIDIS, F. LEESE, H.M.R. MANN, K. MEISSNER, C. MELVAD, J. RAITOHARJU, 2021: Deep learning and computer vision will transform entomology. PNAS, 118 (2).

LAFFERTY, J.D.,A. MCCALLUM, F.C.N. PEREIRA, 2001: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289.

MONTANI, I., M. HONNIBAL, 2018: Prodigy: A new annotation tool for radically efficient machine teach-ing, Artificial Intelligence to appear.