Zusammenfassung

Angesichts der wachsenden Menge öffentlich zugänglicher Informationen müssen Unternehmen Wege finden, um Informationen zu gewinnen, die für ihr Geschäft von entscheidender Bedeutung sein können. Leider werden viele dieser Informationen als unstrukturierte oder halbstrukturierte Dokumente veröffentlicht. Softwaretools können solche Daten nicht analysieren, und Menschen würden so lange brauchen, um eine Aufgabe zu erledigen. Als Lösung für dieses Problem wurde die Informationsextraktion (IE) entwickelt. Ziel der Informationsextraktion aus Text ist es, strukturierte, maschinell verarbeitbare Fakten aus unstrukturierten Dokumenten zu gewinnen.

Dari ist eine ressourcenarme Sprache, die in Afghanistan gesprochen wird. Es gibt keine Korpora oder zuvor etablierte Anwendungen für die Verarbeitung natürlicher Sprache in Dari. Einen Korpus zu erstellen ist schwierig, denn das Erkennen eines Wortes mit mehreren Bedeutungen erfordert Expertenwissen. In dieser Studie haben wir sechs Datensätze erstellt, die online verfügbar sind.

Informationsextraktionssysteme benötigen ein Modell, das beschreibt, wie relevante Zielinformationen in Texten identifiziert werden. Diese Modelle müssen an die Zielinformationen und die Texteingabe angepasst werden, was typischerweise mit Techniken des maschinellen Lernens erreicht wird, die solche Modelle basierend auf Beispielen generieren. Wir haben eine IE-Pipeline aufgebaut, die vier Hauptaufgaben umfasst: (1) Vorverarbeitung, (2) Named Entity Recognition, (3) Coreference Resolution und (4) Relation Extraction. Diese Pipeline skizziert die Aufgabe des IE, und jede Stufe empfängt Eingaben von der vorherigen Stufe und sendet ihre Ausgabe an die nächste Stufe. Bei der Vorverarbeitung werden bereitgestellte Texte mit einem Dari-Satz- / Wortsegmentierungstool in Sätze und Wörter aufgeteilt.

Named Entity Recognition (NER) identifiziert Erwähnungen aus Texten, die zu vordefinierten semantischen Typen wie Person, Standort, Organisation usw. gehören. Frühe NER-Systeme erzielten große Erfolge in Bezug auf die Leistung und verursachten gleichzeitig die Kosten für Human Engineering bei der Entwicklung domänenspezifischer Funktionen und Regeln. In den letzten Jahren hat

Deep Learning eine State-of-the-Art-Performance ermöglicht. In der Sprache Dari unterscheiden sich Named Entities von anderen Sprachen. Es gibt keine Großbuchstaben, um Named Entities zu unterscheiden, und viele Named Entities können in mehr als ein Wort zerlegt werden. Wir haben das Hidden-Markov-Modell und das Bidirectional Long-Short term memory verwendet und festgestellt, dass diese beiden deterministischen und stochastischen Modelle vergleichbare Ergebnisse liefern und die Domänenunabhängigkeit unterstützen.

Die nächste Aufgabe ist die Coreference Resolution, die sich darauf konzentriert, festzustellen, ob sich zwei Ausdrücke im Text auf dieselbe Entität beziehen. Die Erstellung eines Datensatzes zur Coreference Resolution ist zeitaufwendig und erfordert linguistische Kenntnisse. Der DariCoref-Datensatz wurde mit Hilfe der Fakultät für Linguistik der Universität Kabul erstellt. Die Anmerkungswerkzeuge haben auch Einschränkungen bezüglich der Dari-Sprache. Wir haben festgestellt, dass der MMAX2 für den Dari-Text geeignet ist, nachdem wir zahlreiche Tools evaluiert haben.

Wir verwendeten ein vergleichendes Erwähnungspaar-Modell mit einem Entscheid- ungsbaum-Klassifikator für zwei Datensätze. Zuerst trainierten wir ein auf Kategorien (z. B. lexikalisch, syntaktisch und semantisch) und Erwähnungstyp (z. B. Pronomen, Eigennamen) spezialisiertes merkmalsbasiertes Modell im DariCoref-Datensatz. Parallele Matching-Algorithmen wurden auch für eine String-Match-Funktion und die Kankor-Datenbank (University National Entry Exam) verwendet,

die Nachnamen, männliche und weibliche Vornamen für die Geschlechtervereinbarung und eine Alias-Funktion enthält. Zweitens wurde bei einem strukturierten Datensatz (1000 Wörter) der Entscheidungsbaum aus der Scikit-Learn-Bibliothek verwendet, um numerische Werte zu klassifizieren. Das merkmalsbasierte Entscheidungsbaummodell übertraf den strukturierten Datensatz in Bezug auf die Genauigkeit.

Schließlich ist die Relationsextraktion der Prozess der Bestimmung semantischer Beziehungen zwischen Entitäten in einem Satz. Es gibt mehrere Methoden, um Beziehungen aus unstrukturiertem Text zu extrahieren. Diese Methoden können nur dann eine höhere Genauigkeit erzielen, wenn der Datensatz manuell mit Anmerkungen versehen wird. Die vorliegende Forschung ruft binäre Beziehungen aus dem unstrukturierten Text durch einen partiellen syntaktischen Parser (Dari NER und Dari Part-of-Speech-Tagger) ab. Die Polynom-, RBF- und Sigmoid-Kernel wurden zur Klassifizierung in die Support Vector Machine übernommen.

Abstract

With the growing amount of information, companies must think of ways for mining information that may be critical to their business. Unfortunately, much of this information is published as unstructured or semi-structured documents. Software tools cannot analyze such data, and humans would take so long to complete a task. Information Extraction (IE) is developed as a solution to this issue. The objective of extracting information from text is to obtain structured, machine-processable facts from unstructured documents.

Dari is a low-resource language spoken in Afghanistan. There are no corpora or previously established applications for natural language processing tasks in Dari. Creating a corpus is difficult because recognizing a word with several meanings necessitates expert knowledge. In this research, we created six datasets, which are available online.

Information Extraction systems require a model that describes how to identify relevant target information in texts. These models need to be adapted to the target information and the textual input, typically accomplished using Machine Learning techniques that generate such models based on examples. We built an IE pipeline that included four major tasks: (1) preprocessing, (2) entity extraction, (3) coreference resolution, and (4) relation extraction. This pipeline outlines IE's task, and each stage receives input from the previous stage and sends its output to the next stage. During preprocessing, provided texts are split into sentences and words using a Dari sentence/word segmentation tool.

Named entity recognition (NER) identifies mentions from text belonging to predefined semantic types such as person, location, organization. Early NER systems achieved great success in terms of performance, but at the cost of lower recall and months of work by computational linguistics in the development of domain-specific features and rules. In recent years, deep learning afforded state-of-the-art performance. In the Dari language, named entities are different from other languages. There are no capital letters to distinguish named entities, and many named entities can be decomposed into more than one word. We used the Hidden Markov Model and Bidirectional Long-Short term memory

and found that these two deterministic and stochastic models can produce comparable results and support domain independence.

The next task is coreference resolution, which focuses on identifying whether two expressions in the text refer to the same entity. The creation of a dataset for coreference resolution is time-consuming and needs linguistic knowledge. The DariCoref dataset was created with the help of Kabul University's Department of Linguistics. The annotation tools also have restrictions with the Dari language. We found that the MMAX2 is appropriate for the Dari text after evaluating numerous tools.

We employed a comparative mention-pair model with a decision tree classifier on two datasets. First, we trained a feature-based model specialized in categories (e.g., lexical, syntactic, and semantic) and mention type (e.g., pronouns, proper nouns) on the DariCoref dataset. Parallel matching algorithms were also utilized for a string match feature and the Kankor (University National Entry Exam) database, including last names, male and female first names for gender agreement, and an alias feature. Second, on a structured dataset (1K words), used the decision tree from the Scikit-learn library to classify numerical values. The feature-based decision tree model outperformed the structured dataset in terms of accuracy.

Finally, relation extraction is the process of determining semantic relations between entities in a sentence. Several methods exist for extracting relations from unstructured text. These methods can achieve higher accuracy only when the dataset is manually annotated. This research retrieves binary relations from the unstructured text through a partial syntactic parser (Dari NER and Dari part-of-speech tagger). The polynomial, RBF, and Sigmoid kernels were adopted into the Support Vector Machine for classification.