Finden ohne zu suchen: Mit regelbasierter KI automatisch Benachrichtigungen über relevante wissenschaftliche Publikationen erhalten

Hermann Bense
textomatic.AG, Schwarze-Brüder-Str. 1, 44137 Dortmund, GERMANY
hermann.bense@textomatic.ag

Keywords: Finding without Searching (FwS), Language Dictionary, Machine Learning based Translations, Key Value Store (KVS), Natural Language Processing (NLP), Stemming, Tagging, Semantic Search.

1 Einleitung

Jeden Tag erscheint eine Flut von hunderten und tausenden neuer wissenschaftlichen Publikationen. Für Forscher ist es eine mühsame und zeitraubende Aufgabe, dabei den Überblick zu behalten. Ein zentrales Problem von Suchmaschinen wie Google, scholar.google.com und wissenschaftlichen Suchportalen wie z.B. TIB [1] und SpringerProfessional.de ist, dass die Suchergebnisse oft nicht den Erwartungen der Forscher in Bezug auf Aktualität und Relevanz entsprechen. Dieser Artikel beschreibt, wie Forscher aus angewandter und experimenteller Sicht durch eine Methode, die als Finden ohne Suchen (**FwS = finding without searching**) bezeichnet wird, effizient unterstützt werden können. Diese Methode nutzt künstliche Intelligenz in Kombination mit ausdrucksstarken benutzerdefinierten Regeln für Benachrichtungen zu Publikationen.

2 Das News Alert System NAS und die rOb.by-App

Das News Alert System NAS [2] und seine Benutzeroberfläche, die rOb.by-App [3], nutzen einen Korpus von mehr als drei Millionen Publikationen in nahezu allen Teildisziplinen. Mehr als 290 Millionen Tripel von Daten und Metadaten sind im Repository gespeichert. Crawler scannen permanent das Web nach neuen Publikationen und fügen neue Informationen in den Korpus ein. Die Identifizierung der Autoren erfolgt effektiv über die Autorennamen, mit Ausnahme bestimmter asiatischer Autorennamen, wie unten erläutert. Die vorherrschenden Dokumentensprachen, die hier diskutiert werden, sind Englisch und Deutsch. Die Titel der Dokumente aus mehr als 25 Sprachen werden mit Hilfe der deepl.com-API [4] automatisch ins Englische und Deutsche übersetzt. Neben den Schlagwörtern, die den Dokumenten von den Autoren zugewiesen wurden, werden zusätzliche Schlagwörter aus den Titeln und deren Übersetzungen mit Stanza [5] für Englisch und TreeTager [6] für Deutsch ermittelt.

2.1 Wörterbücher und Verschlagwortung

Zur Optimierung der Antwortzeiten fügt das Indizierungstool des NAS alle erfassten Schlagwörter, Autorennamen und zusätzliche Lemmata in den Key-Value-Store KVS des Systems ein. Derzeit enthält der KVS ca. 90 Millionen Tripel. Rund 2.300 englische und deutsche Schlagwörter werden als Stoppwörter verwendet und von der Indizierung ausgeschlossen. Stoppwörter und sprachübergreifende Synonymbeziehungen (Synsets, Synonymringe) [7] sind Teil des 3dna.news-Wörterbuchs. Das 3dna-Wörterbuch wurde im Rahmen des von Google geförderten DNI-Projekts 3dna.news [8] für den Zweck der groß angelegten Nachrichtengenerierung entwickelt [9].

Jedes Schlagwort im KVS hat ein Meta-Attribut KWF (KeyWord Frequency), das angibt, wie viele Dokumente von den Schlagworten referenziert werden. Die Liste der Schlüsselworthäufigkeiten zeigt, dass das Top-Schlüsselwort *Engineering* ca. 130k Dokumenten zugeordnet ist. Es folgen die Schlagwörter *Systems, Intelligence, Analysis, Management, based, System, Theory, computational* und *Information*, die jeweils mehr als 60k Dokumente indizieren. Mit den Top 55 Stichwörtern sind 2,7 Millionen Dokumente indexiert, was fast der Größe des gesamten Korpus entspricht. Schlüsselwörter, die mehr als 10k Dokumente indizieren, werden als sehr hochfrequente Schlüsselwörter (VHFK) bezeichnet. Etwa 340 Schlüsselwörter erfüllen dieses Kriterium. Die Schlüsselwörter, die weniger als 100 Dokumente indizieren, werden als very low-frequency keywords (VLFK) bezeichnet, die mit weniger als 1k Dokumenten als low-frequency keywords (LFK). Etwa 57k Keywords indizieren zwischen 50 und 100 Dokumente.

2.2 Benutzerdefinierte Regeln für die automatische Recherche

Eine Schlüsselfrage für das Auffinden relevanter Publikationen ist die Zuordnung von Schlüsselwortkombinationen zu der beabsichtigten lexikalischen Semantik. Die Menge der Dokumente, die in eine Klasse von semantischen Konzepten fallen, könnte durch eine oder mehrere Regeln beschrieben werden, die als Query-Strings wie "`Named, Entity, Recognition`" oder "`Sentiment, Analysis`" oder "`human, intelligence, ~machine,~artificial`" definiert sind. Ein Query-String ist also de facto eine Definition eines lexikalischen Konzepts. Ein Regelname wie "*Menschliche Intelligenz*" könnte dann die beabsichtigte Semantik der Regel/Abfrage alt Regelname zusammenfassen.

In Abbildung 1 wird gezeigt, wie der Benutzer eine Regel in der rObby-App definiert. Jede Zeile

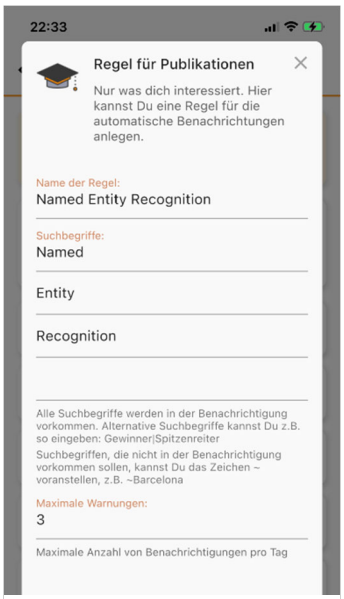

Abbildung 1 Abb. 1: Regel für das Auffinden von Dokumenten zur Keyword-Kombination „Named, Entity, Recognition“

unter Suchbegriffe kann einen einzelnen Suchbegriff oder mehrere ODER-verknüpfte Suchbegriffe enthalten.

2.3 Benachrichtigung über Neuerscheinungen am selben Tag:

Was bedeutet **Finden ohne Suchen**? Das NAS-System arbeitet wie ein Wachhund. Für jede neue Publikation, die in den Korpus aufgenommen wird, werden alle benutzerdefinierten Regeln auf Übereinstimmung mit den Suchbedingungen geprüft. Trifft eine Regel zu, wird der Rechercheur noch am selben Tag per Push-Benachrichtigung oder per E-Mail über die neue Publikation informiert. Die Komplexität der Abfragen spielt in diesen Fällen keine große Rolle, da jeder Publikation im KVS das Datum der Veröffentlichung zugeordnet ist. Damit entlastet das NAS-System den Anwender von der regelmäßigen Suche nach relevanten Dokumenten.

Zentrale Komponente im NAS-System dafür ist der Rule-Processor. Dieser überprüft kontinuierlich alle Regeln aller Benutzer. Es wird ebenfalls protokolliert, welche und wie viele Benachrichtigungen ein Benutzer bereits erhalten hat. Dadurch wird sichergestellt, das er keine doppelten Benachrichtigungen erhält und die vom Benutzer vorgegebenen maximale Anzahl von Benachrichtigungen pro Regel nicht überschritten wird.

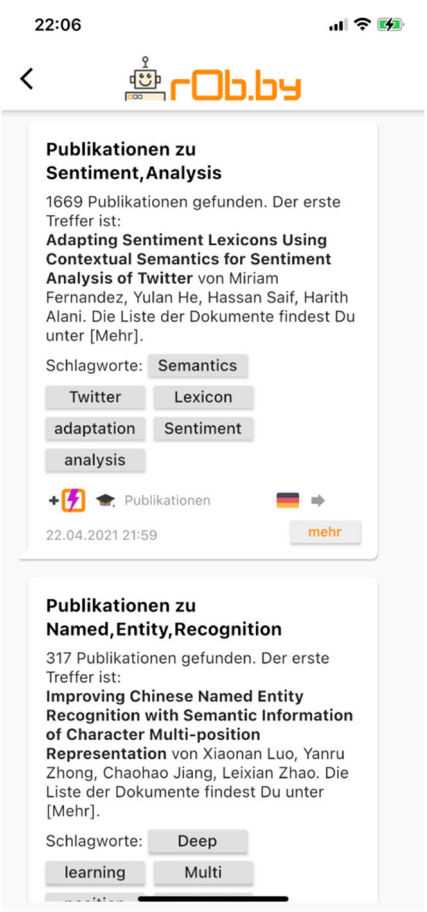

Abbildung 2: Beispiel für eine automatische Benachrichtigung in der rObby-App

Das Beispiel in Abbildung 2 zeigt, dass zu der Suchabfrage wie "`Named, Entity, Recognition`" eine neue Publikation gefunden wurde und das es insgesamt 317 Publikationen zu dieser Keyword-Kombination gibt. Ebenso werden die weiteren Keywords wie „*deep, learning, Multi*“ etc. angezeigt, die der Publikation von den Autoren oder durch die Verschlagwortungsfunktion des NAS-System zugewiesen wurden. Durch Klicken auf den [mehr]-Button gelangt der Benutzer in die web-basierte rObby-Suchmaschine und zur Liste aller gefundenen Dokumente. Dort kann er dann weitergehende Recherchen durchführen.

2.4 Die rObby-Suchmaschine

Der Benutzer kann die rObby-Suchmaschine [3] für die Vor- und Nachrecherche verwenden. Im Vergleich zu anderen Suchmaschinen für wissenschaftliche Publikationen werden zu den Suchbegriffen die Keyword-Häufigkeiten angezeigt (Abbildung 3). Außerdem kann der Benutzer auf intuitive Weise eine neue Suche starten, indem er auf einen orangenen Keyword- oder Autoren-Link klickt. Alternativ kann er die vorhandenen Suche mit einem Klick auf ein + Symbol um weitere Keywords ergänzen.

3 Unterstützung von Forschern bei der Suche nach relevanteren Publikationen:

Es mag trivial erscheinen, aber hochfrequente Schlüsselwörter (HFK) wie z. B. *Engineering* haben eine deutlich geringere Selektivität, während niedrigfrequente Schlüsselwörter (LFK) eine effektive Einschränkung der Ergebnismengen ermöglichen.
Namen von Autoren gehören im Allgemeinen zu den LFKs, da es sehr unwahrscheinlich ist, dass Autoren mehr als 1k Publikationen haben. Es gibt bemerkenswerte Ausnahmen, da asiatische Autorennamen wie *Huang, Kim, Lee, Liu, Yang, Zhao, Zhang* und *Zhou* und indische Autoren wie *Kumar* und *Singh* zu den VHFKs gehören.

3.1 Erweiterte Suche: Geeignete Schlagwörter für Publikationen finden:

Das logische ODER kann verwendet werden, um Schlüsselwörter mit unterschiedlichen Schreibweisen oder Floskeln oder Homonymen zu verbergen. Das Pipe-Symbol | kann dafür in Suchanfragen wie in `system|systems`, oder `deep|machine,learning` verwendet werden. Das logische NOT kann verwendet werden, um die Ergebnismengen um Wörter zu reduzieren, die nicht bereits in der Menge der Stoppwörter enthalten sind. Z. B. `Mensch,Intelligenz,~Maschine` liefert alle Dokumente, die mit Mensch UND Intelligenz indiziert sind, ohne Dokumente, die mit *Maschine* indiziert sind. Es gibt spezifische Wortformen und Beziehungen zwischen Wörtern und Begriffen, die eine weitere Diskussion verdienen:

3.2 Homonyme

Sehr hochfrequente Homonyme (VHFKs) wie *can* (Verb und Substantiv), *not/Not* (englische Negation und deutsches Substantiv für Notwendigkeit), *may* (Verb und Monatsname), *second* (Zahl und Substantiv) , *set* (Verb und Substantiv), *state/s* (Substantiv für Status und Substantiv für Land) und *use* (Verb und Substantiv) erfordern eine besondere Behandlung. Ein Problem zeigt sich auch bei benannten Entitäten. Beispiele: *Schade* (Nachname des Autors und deutsches Adjektiv für Mitleid) und Siegel (Nachname des Autors und deutsches Substantiv für *Siegel/Signum*).

4 Zusammenfassung und Ausblick

4.1 Synsets und semantische Ringe:

Eine offene Forschungsfrage ist es, herauszufinden, welche Schlüsselwortkombinationen sinnvoll sind und zu relevanten Suchergebnissen und Dokumentenclustern führen. Der Artikel zeigt, wie Forscher bei dieser Aufgabe durch die Methode "Finden ohne Suchen" unterstützt werden. Ein weiterer Fortschritt könnte jedoch dadurch erreicht werden, dass die Schlüsselwörter durch eine semantische Nachbarschaftsbeziehung miteinander in Beziehung gesetzt werden, z. B. durch die Verwendung von Word-Net-Synsets [7] und mehrsprachigen Synonymringen wie *Employment* (EN), *job* (EN/DE), *Anstellung* (DE), *Arbeitsplatz* (DE), *travail* (FR).

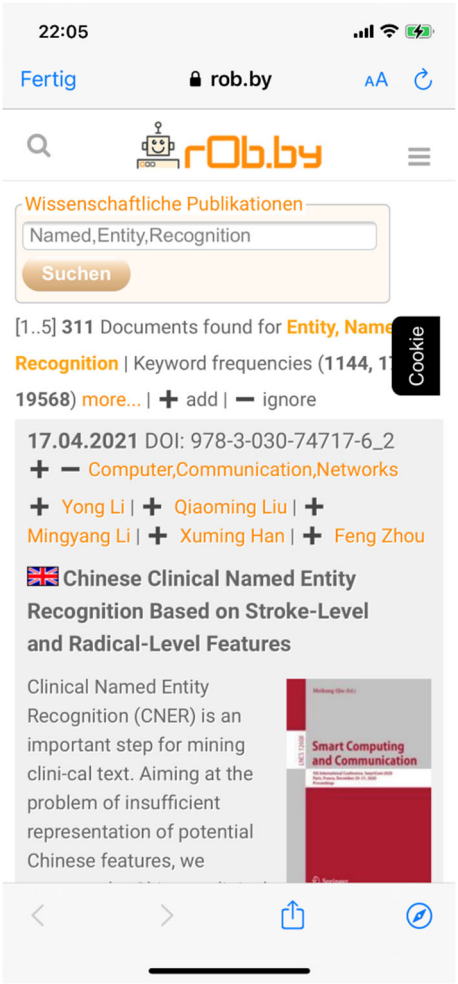

Abbildung 3: rObby-Suchportal

4.2 Domänen-Ontologien:

In der Semantic-Web-Community werden Domänen-Ontologien erstellt, um Konzepte und ihre Semantik zu modellieren. Insbesondere Hypernym-, Hyperonym-, Holonym- und Kausalbeziehungen erlauben es, formalere Taxonomien und Klassenhierarchien zu modellieren. Die EnArgus-Ontologie [10] ist ein Beispiel für eine Ontologie im Bereich der Energieforschungsprojekte, die etwa 12.000 Konzepte wie *Windenergie* oder *Solarkraftwerke* modelliert. Die innerhalb der Domainen-Ontologien gewonnenen Begriffsdefinitionen können durch Anwendung von Inferenzregeln zur Erweiterung der Suchergebnisse mit herangezogen. Die dazu verwendete Suchmethode SbM (Search by Meaning) wurde ursprünglich innerhalb des EnArgus-Projektes [10, 11] entwickelt. Grundlegend hier ist die Idee, Begriffe aus anderen Begriffe zusammenzusetzen, z.B. Kondensator = Speicher (Energie (elektrisch)). Publikationen, die über Kondensatoren berichten, könnten also auch mit der Keyword-Kombination „elektrisch, Energie, Speicher“ oder auch über ontologische Beziehungen und Definition inferiert und gefunden werden.

Nach Kenntnis des Autors verfügt keine der anderen Suchmaschinen und Portale für wissenschaftliche Publikationen über Angaben zu Keyword-Häufigkeiten. Keyword-

Häufigkeiten helfen Autoren bei der Auswahl der richtigen Keywords für die Verschlagwortung ihrer Publikationen. Der Artikel hat gezeigt, wie wertvoll dies auch für die Optimierung von Suchanfragen für wissenschaftliche Publikationen und für die Definition von lexikalischen Begriffen mit Regeln sein kann. Auch die Möglichkeit, die Publikationen mit Suchbegriffen in mehreren Sprachen aufzufinden dürfte einzigartig sein. Als zusätzliche Funktion ermöglicht es die rObby-App, die Benachrichtigungen in Form der Titel und Zusammenfassungen in mehr als 25 Sprachen zu erhalten.

Für die Zukunft ist geplant, weitere Bibliotheksbestände zu erschließen, insbesondere in den Bereichen Bio-Medizin und Patente.

Referenzen

1. TIB, The Leibniz Information Center for Technology and Natural Sciences and University Library, (https://www.tib.eu/de/, last visit: 17.06.2021)
2. NAS, News-Alert-System: https://rob.by/en/NAS/System/, last visit: 17.06.2021)
3. rObby, News-Alert-App (https://rob.by/en/App/, last visit: 17.06.2021)
4. deepl, https://www.deepl.com/docs-api, last visit: 17.06.2021
5. Stanza: Python NLP Library for Many Human Languages - formerly StandfordNLP (https://github.com/stanfordnlp/stanza/, last visit: 17.06.2021)
6. TreeTagger, (https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger/, last visit: 17.06.2021)
7. Christiane Fellbaum (Editors), WordNet: An Electronical Lexical Database , 1998, Mass.:MIT Press, Cambridge (https://books.google.de/books/about/WordNet.html?id=Rehu8OOzMIMC&redir_esc=y, last visit: 17.06.2021)
8. 3dna.news, data-driven-digital news agency (https://3dna.news/en/, last visit: 02.05.2021)
9. Hermann Bense, Using Very Large Scale Ontologies for Natural Language Generation (NLG) , 2017, in Stefano Borgo, Oliver Kutz, Frank Loebe, Fabian Neuhaus (Editors), Jowo 2017 - The Joint Ontology Workshops, Episode 3: The Tyrolean Autumn of Ontology, Bozen-Bolzano, Italy, Sept. 21-23, 2017 (http://ceur-ws.org/Vol-2050/DAO_paper_1.pdf, last visit: 17.06.2021)
10. Hermann Bense, Ulrich Schade, Frederike Ohrem, Lukas Sikorski, Recherche-Unterstützung durch Ontologie Visualisierung im EnArgus2-Projekt , 2014 (https://www.enargus.de/, last visit: 17.06.2021)
11. CNS – Concept Numbering System, https://www.taoke.de/ke/CNS/, last visit: 17.06.2021