Wiederherstellung elidierter Morpheme zur Verbesserung der Informationsextraktion

Institut für Digital Humanities – Sprachliche Informationsverarbeitung | Johanna Binnewitt | 25.05.2019

Das Kooperationsprojekt

„Qualifikationsentwicklungsforschung: Aufbereitung, Annotation und Auswertung eines großen Korpus von Stellenanzeigen“

- Anwendung von Text Mining auf Stellenanzeigen
- 3. Projektphase (11/2018-04/2020)

Kooperationspartner

- Bundesinstitut für Berufsbildung (BIBB)
- Universität zu Köln:
  Sprachliche Informationsverarbeitung (Spinfo)
  (Institut für Digital Humanities)

Das Kooperationsprojekt

- Entwicklung eines Frameworks (*quenfo*) …
  - … zur Extraktion und Klassifikation relevanter Inhalte aus Stellenausschreibungen
  - … mithilfe eines anonymisierten Test-Korpus (ca. 300 Anzeigen) aus einem größeren Korpus der Agentur für Arbeit (> 4 Mio. Anzeigen)
  - … geschrieben in Java
  - … als Open Source-Projekt

Informationsextraktion – Relevante Informationen

Tools	Welche Arbeitsmittel werden in der Jobbeschreibung genannt?
Tasks	Welche Tätigkeiten werden in der Jobbeschreibung aufgeführt?
Compe-tences	Über welche Kompetenzen sollten / müssen Bewerber_innen verfügen?

Beispiel Stellenausschreibung

Wir suchen zum nächstmöglichen Zeitpunkt einen

Logopäden / Klinischen Linguisten (m/w/d)

Ihre Aufgaben:

- Erstellung individueller Behandlungs- und Betreuungspläne durch den Einsatz gängiger Testverfahren
- Behandlung verschiedener Störungsbilder (Schluck-, Sprach-, Sprech- und Stimmstörungen)

Ihr Profil:

- Abgeschlossene Ausbildung zum Logopäden / Klinischen Linguisten (m/w/d)
- Eine qualifizierte eigenständige und strukturierte Arbeitsweise sowie einen empathischen und teamorientierten Umgang mit Patienten und Kollegen

Für Rückfragen stehen wir ihnen unter xxx gerne zur Verfügung.
Senden Sie ihre Bewerbungsunterlagen an xxx@xxx.de

Text Mining

Wir suchen zum nächstmöglichen Zeitpunkt einen

Logopäden / Klinischen Linguisten (m/w/d)

Ihre Aufgaben:

- Erstellung individueller Behandlungs- und Betreuungspläne durch den Einsatz gängiger Testverfahren
- Behandlung verschiedener Störungsbilder (Schluck-, Sprach-, Sprech- und Stimmstörungen)

Ihr Profil:

- Abgeschlossene Ausbildung zum Logopäden / Klinischen Linguisten (m/w/d)
- Eine qualifizierte eigenständige und strukturierte Arbeitsweise sowie einen empathischen und teamorientierten Umgang mit Patienten und Kollegen

Für Rückfragen stehen wir ihnen unter xxx gerne zur Verfügung.
Senden Sie ihre Bewerbungsunterlagen an xxx@xxx.de

Musterbasierte Informationsextraktion

- standardisierte Formulierungen ermöglichen eine Extraktion mithilfe von Mustern
- Muster: Kombination zwischen Lemmata und POS-Tags (Stuttgart-Tübingen-Tagset)

kenntnis|grundkenntnis|erfahrung

+ ART|APPR|APPRART

+ TRUNC

+ KON|$,

+ NN|NE

→ Erfahrung in Verkaufs- und Beratungsgesprächen

Musterbasierte Informationsextraktion

- extrahierte Terme können anschließend strukturiert aufbereitet werden
  - Einordnung in Taxonomie (Welche Art von Kompetenz ist das?)
  - Häufigkeitsanalysen (Wie häufig kommt eine bestimmte Kompetenz in einem Zeitraum oder einer Branche vor?)

- Problem: Koordinierte Ausdrücke
  - Erfahrung in Beratungs- und Verkaufsgesprächen

  → Erfahrung in Beratungsgesprächen

  → Erfahrung in Verkaufsgesprächen

Expansion elidierter Morpheme – die Software

CoordinateExpander

- expandiert elidierte Morpheme
- Als GitHub-Repository verfügbar
- Nicht domänenspezifisch
- Anwendungsbereich: Information Retrieval

Morphemkoordinationen

- Koordination: Struktur, in der wiederkehrendes Material entfernt wird (Booji 1985)

  „Paul **isst** einen Apfel und Pia eine Birne.“

- Morphemkoordination: elidierter Teil umfasst lediglich einen Wortteil

  Beratung**s**- und Verkaufsgespräche

- Interfix wird beibehalten → Bindestrich ersetzt die entfernten Morpheme

Morphemkoordinationen identifizieren

- In *quenfo* werden Koordinationen berücksichtigt, die in Informationseinheiten auftauchen

Beratungs-	und	Verkaufsgespräch
TRUNC	KON	NN

- KON[1]*:* Auslöser für Koordinationsexpansion
- TRUNC[2]*:* Anzeichen für Rechtsellipse

1: *nebenordnende Konjunktion*

2: Kompositions-Erstglied

Morphemkoordinationen identifizieren

Fehleranalyse	und	-verfolgung
NN	KON	NN

- KON: leitet Koordinationsexpansion ein
- Linksellipsen: kein eindeutiges POS-Tag

→ Bindestrich zu Beginn des Tokens nach Konjunkt

Morphemkoordinationen expandieren

koordinierte Wörter identifizieren

Deutsch-	und	gute	Englischkenntnisse
TRUNC	KON	ADJA	NN

- Koordiniert Wörter der gleichen Wortart

→ Problem: TRUNC „verschleiert“ Wortart

- Heuristik: großgeschrieben → NN[1] vs. kleingeschrieben → ADJA[2]

1: *normales Nomen*

2: attributives Adjektiv

Kompositazerlegung

elidiertes Morphem im Wort identifizieren

- Kompositazerlegung mithilfe von JWordSplitter[1]
  Beratungs- und Verkaufs | **gespräche**

  Kranken- oder Alten | **pflege | fach | kraft**
- unbekanntes Wort? → letztes Morphem wird als elidierter Teil angenommen

→ Kranken- oder Altenpflegefach**kraft** → Krankenkraft

1: https://github.com/danielnaber/jwordsplitter

Kompositazerlegung

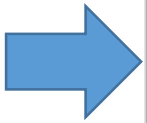
Manuelle Korrektur

splittedCompounds.txt | possibleCompounds.txt

```
1  Krankenpflegefach|kraft
2
```

splittedCompounds.txt | possibleCompounds.txt

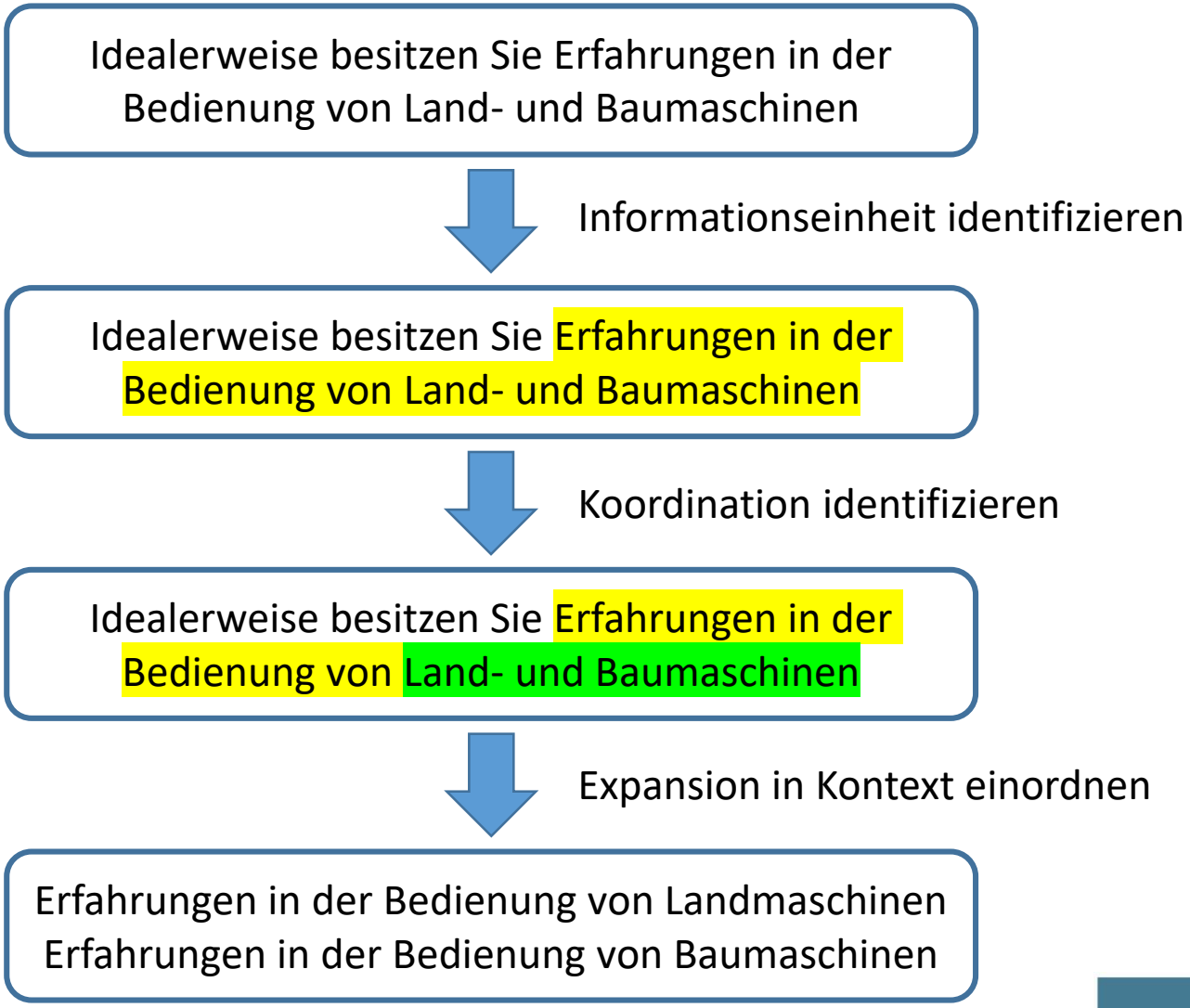
```
1  Kranken|pflegefachkraft
2
```



splittedCompounds.txt | possibleCompounds.txt

```
19  Mess|technik
20  In|landsreisen
21  Kranken|pflegefachkraft
22  Englisch|kenntnisse
23  Landschafts|bau
24  Automatisierungs|technik
25  Fernmelde|technik
26  Weiter|bildung
27  Raum|fahrttechnik
28  Kranken|pflege
29  Aufstiegs|möglichkeiten
30  Restaurant|fach
```

Einordnung in den Kontext



Evaluation

- Entwicklung der Methoden anhand weniger Beispielfälle
- Quantitative Evaluation (noch) nicht möglich
- Methoden wurden nach und nach an neue Fälle (z.B. Linksellipsen) angepasst

Ausblick

- Identifizierung des elidierten Wortteils verbessern
- Andere Koordinationen (z.B. Phrasenkoordinationen) bearbeiten („Maschinen warten und instand halten“)

Literatur

Booij, Geert. (1985) “Coordination Reduction in Complex Words: A Case for Prosodic Phonology”. in Hulst and Smith (eds.) (1985) Advances in Nonlinear Phonology 143 - 160.

Vielen Dank für die Aufmerksamkeit!

Projekt Website: http://dh.uni-koeln.de/37089.html

Code auf GitHub: https://github.com/johannabi/CoordinateExpander