FACHBEITRÄGE IM INTERNET

Jürgen Hermes | Manuel Schandock

Stellenanzeigenanalyse in der Qualifikationsentwicklungs-forschung

Die Nutzung maschineller Lernverfahren zur Klassifikation von Textabschnitten

Bundesinstitut für Berufsbildung BIBB

Forschen
Beraten
Zukunft gestalten

FACHBEITRÄGE IM INTERNET

Jürgen Hermes | Manuel Schandock

Stellenanzeigenanalyse in der Qualifikationsentwicklungs-forschung

Die Nutzung maschineller Lernverfahren zur Klassifikation von Textabschnitten

Bundesinstitut für Berufsbildung BIBB

Forschen
Beraten
Zukunft gestalten

Bibliografische Information der Deutschen Nationalbibliothek

Die deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Herausgeber: Bundesinstitut für Berufsbildung, Bonn

Herstellung: Bundesinstitut für Berufsbildung, Bonn

Bundesinstitut für Berufsbildung
Arbeitsbereich 1.4 – Publikationsmanagement/Bibliothek
Robert-Schuman-Platz 3
53175 Bonn
Internet: www.bibb.de
E-Mail: zentrale@bibb.de

ISBN 978-3-945981-50-4

Diese Netzpublikation wurde bei der Deutschen Nationalbibliothek angemeldet und archiviert: urn:nbn:de: 0035-0620-5

Internet: www.bibb.de/veroeffentlichungen

Inhaltsverzeichnis

Abbildungen

Tabelle

1 Kontext

Stellenanzeigen enthalten aussagekräftige und umfangreiche Informationen zu betrieblichen Beschreibungen von Arbeits- und Ausbildungsstellen sowie den entsprechenden Anforderungsprofilen an die künftigen Stelleninhaber/-innen. Allerdings liegen diese Informationen in Stellenanzeigen unstrukturiert und unsystematisch vor. Auf Grundlage eines Kooperationsvertrags zwischen dem Bundesinstitut für Berufsbildung (BIBB) und der Universität zu Köln, vertreten durch das Institut für Linguistik, Sprachliche Informationsverarbeitung (Spinfo), wurde ein Software-Framework entwickelt, mit dessen Hilfe unterschiedliche Verfahren zur Klassifikation von kurzen Textabschnitten evaluiert werden können. Mittels dieser Software wurden unter mehr als 5.000 verschiedenen Verfahren geeignete Kandidaten ermittelt, die Abschnitte aus Stellenanzeigen präzise auf vordefinierte, inhaltlich motivierte Klassen aufteilen können. Diese Verfahren werden auf einer BIBB-internen Datenbank mit mehreren Millionen Stellenanzeigen als erste Stufe einer Methodik mit dem Ziel einer Informationsextraktion aus Stellenanzeigen eingesetzt. Durch die Auswertung von Stellenanzeigen in großem Umfang sollen aktuelle Tendenzen beim Qualifikationsbedarf von Unternehmen sichtbar gemacht werden. Die Ergebnisse dieses Ansatzes können durch ihre Aktualität und Breite in Bezug auf Branchen und Berufe unmittelbar die Qualifikationsentwicklungsforschung im BIBB stärken.

1.1 Stellenanzeigenanalyse als Instrument der Qualifikationsentwicklungsforschung

Die meisten Statistiken, die das Arbeitsmarktgeschehen dokumentieren, lassen in erster Linie Rückschlüsse auf den Status quo der Qualifikationsstrukturen in den Unternehmen zu. Hierzu zählen insbesondere Erhebungen wie der Mikrozensus oder Daten über die sozialversicherungspflichtig Beschäftigten bei der Bundesagentur für Arbeit (BA). Sie dokumentieren das Resultat des Arbeitsmarktgeschehens sowie daran anschließender innerbetrieblicher Entwicklungs- und Qualifizierungspfade. Aber diese Momentaufnahmen bestehender (formaler) Qualifikationsstrukturen geben nicht unbedingt hinreichenden Aufschluss über den eigentlichen Bedarf, da sich die (nicht beobachtbaren) Kompetenzen und Tätigkeiten und die (beobachtbaren) formalen Qualifikationen im Laufe eines Berufslebens voneinander entfernen können. So bringen in Regionen mit Fachkräftemangel nicht immer alle Bewerber/-innen die aus Sicht der Unternehmen notwendigen Qualifikationen in vollem Umfang mit. In der Folge können die Unternehmen gezwungen sein, auf weniger qualifizierte Bewerber auszuweichen und zur Deckung ihres ursprünglichen Bedarfs Weiterbildungsangebote zu unterbreiten. Weder Weiterbildungen noch Lernprozesse am Arbeitsplatz ändern etwas an den formalen Qualifikationen, wie sie bspw. im Mikrozensus erfasst werden. Insofern bilden die genannten Statistiken das Arbeitsmarktgeschehen nicht zufriedenstellend ab.

Demgegenüber formulieren Unternehmen in Stellenanzeigen in der Regel eher idealtypische Qualifikationsanforderungen für die Besetzung von Arbeitsplätzen. Stellenanzeigentexte bieten daher ein besonders geeignetes Analysepotenzial für die Untersuchung des zusätzlichen Arbeitskräftebedarfs und dessen aktuelle Veränderungen. Dies soll anhand eines Beispiels verdeutlicht werden:

Werden Fachkräfte für frei werdende Stellen gesucht, orientieren sich Unternehmen mitunter an der Fachqualifikation der ausgeschiedenen Personen. Dabei wird ein Unternehmen in der Regel darauf bedacht sein, die Stellenbesetzung zukunftsorientiert und in diesem Sinne nachhaltig

zu planen. Beispielsweise würde ein Kfz-Fachbetrieb für die Stelle eines Kfz-Mechanikers, der in den Ruhestand eingetreten ist, möglicherweise eine Neubesetzung durch einen Kfz-Mechatroniker anstreben. Durch die Stellenanzeigen werden solche unternehmensstrategischen Entscheidungen deutlich besser abgebildet als durch Bestandsdaten, die, um bei dem Beispiel zu bleiben, lediglich den Kfz-Mechaniker erfassen würden. Die Folgen des technologischen Wandels für den Arbeitsmarkt lassen sich daher mit der Analyse von Stellenanzeigen besser sichtbar machen. Das Gleiche gilt in stärkerem Maße für Unternehmen, die durch Expansion in neue Märkte vordringen. Aus den Bestandsinformationen ließe sich keine Aussage über die Qualifikationsbedarfe dieser Unternehmen ableiten, in Stellenanzeigen werden diese jedoch unmittelbar artikuliert. Daher ist die Untersuchung von Stellenanzeigen bspw. für Transformationsprozesse hin zu einer Green Economy genauso vielversprechend wie auch zu Themen der Digitalisierung und der Verwendung neuer Arbeitsmittel.

1.2 Information in Stellenanzeigen

Durch einen langfristigen Liefervertrag mit der Bundesagentur für Arbeit (BA) stehen dem Bundesinstitut für Berufsbildung (BIBB) bereits jetzt etwa 2.000.000 Stellenanzeigen für Forschungszwecke zur Verfügung. In den kommenden Jahren werden es noch deutlich mehr werden, da jedes Jahr etwa 350.000 bis 500.000 neue Stellenanzeigen hinzukommen. Die Daten repräsentieren nahezu die Hälfte aller offenen Stellen auf dem deutschen Arbeitsmarkt.[1] Neben der Kerninformation der Stellenanzeigen – den Stellenanzeigentexten – sind zusätzliche Informationen zu jeder Anzeige verfügbar. Von den inserierenden Firmen werden Angaben über den gesuchten Beruf, die Firma (bspw. Betriebsgröße), die offene Stelle (bspw. Stellenumfang) und mehr erhoben. Diese Kombination aus den Stellenanzeigentexten und bereits codierten Informationen machen diesen Datenbestand einzigartig.

Das Interesse an den Daten im BIBB ist hoch. Im Rahmen kleinerer manueller Auswertungen wurden Anforderungsprofile für verschiedene Berufe erstellt, um bspw. Neuordnungsverfahren[2] mit aktuellen empirischen Daten zu unterstützen. Dafür wurden jeweils kleine Stichproben aus den Stellenanzeigen gezogen (200 bis 600 Stellenanzeigen) und mit qualitativen Methoden analysiert. Auf diesem Wege konnten berufs- und branchenspezifische Anforderungs-, Tätigkeits- und Kompetenzprofile abgeleitet werden, die den jeweils aktuellen Bedarf der Unternehmen widerspiegeln. Diese Projekte haben gezeigt, welchen Nutzen die Analyse von Stellenanzeigen für die Arbeit im BIBB mit sich bringt.

An dieser Stelle wird nun angeknüpft. Im Bereich der Computerlinguistik wurden Verfahren entwickelt, die durch den Einsatz maschinellen Lernens in der Lage sind, die bisher erfolgte manuelle Inhaltsanalyse von wenigen Stellenanzeigentexten auf die gesamte Menge der vorhandenen Texte auszuweiten. Da die Texte der Stellenanzeigen dem Urheberrecht und dem Datenschutz unterliegen, müssen die Informationen aus den Anzeigetexten extrahiert und in einer Form gespeichert werden, die den Datenschutz- und den Urheberrechtsvorgaben entspricht.[3] Ein nicht unerheblicher Nebeneffekt der Informationsextraktion ist die Möglichkeit, damit Auswertungen von Trends bezüglich der von Bewerbern geforderten Kompetenzen, der gesuchten Berufe und dergleichen durchzuführen.

[1] http://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Qualitaetsberichte/Generische-Publikationen/Qualitaetsbericht-Statistik-gemeldete-Arbeitsstellen.pdf

[2] Erarbeitung von Ausbildungsordnungen und ihre Abstimmung mit den Rahmenlehrplänen der KMK sowie den relevanten Sozialpartnern.

[3] Die Stellenanzeigen enthalten häufig Firmennamen und -adressen, E-Mailadressen sowie Personennamen und Telefonnummern. Diese und ähnliche Informationen dürfen nicht für Analysen genutzt werden und sind nach einer festgelegten Nutzungsdauer zu löschen.

Die bisher (manuell) durchgeführten Analysen der Stellenanzeigen haben gezeigt, dass die Texte eine Reihe auswertbarer Informationen enthalten. Mittelfristig könnten u. a. folgende Informationen aus den Texten extrahiert werden:

- Kompetenzanforderungen
- geforderte Qualifikationsprofile
- Tätigkeitsbeschreibungen
- Arbeitsmittel
- Anreize der Unternehmen zur Fachkräftebindung (bspw. Kinderbetreuung oder Weiterbildung)
- Eigenschaften der Unternehmen (bspw. inhabergeführt, Start-up, Green Economy etc.)

Durch die geplante automatische Informationsextraktion können diese Informationen sowohl auf den kompletten, bereits vorhandenen als auch auf den zukünftigen Datenbestand angewendet werden. Dies ermöglich insbesondere eine Auswertung der Daten im Zeitverlauf:

- Wie verändern sich Kompetenzanforderungen in bestimmten Berufen oder Branchen?
- Werden Unternehmen in Zeiten des Fachkräftemangels flexibler hinsichtlich der Qualifikationsanforderungen an die Bewerber?
- Wie ändern sich Tätigkeitsprofile innerhalb von Branchen und Berufen?
- Welche Anreize setzen Unternehmen zur Mitarbeitergewinnung und -bindung? Lässt sich ein Zusammenhang zum Fachkräftemangel herstellen?
- Wie verändert sich die Nachfrage nach bestimmten Berufen in neuen Branchen (bspw. Green Economy)? Werden neue Berufsbilder benötigt?

1.3 Kooperation mit der Universität Köln

Aufgrund der großen Zahl auszuwertender Stellenanzeigen sind automatisierte Verfahren der Informationsgewinnung eine grundlegende und unverzichtbare Komponente der Datenanalyse. Sie sind im Sinne des Entdeckens und der Nutzbarmachung von Informationen als erster Analyseschritt (Data Mining) zu sehen, bevor inhaltliche Fragestellungen bearbeitet werden können.

Für eine automatisierte Inhaltsanalyse der Anzeigentexte bedarf es spezifischer Methodenkenntnisse, die nicht zum Repertoire in der Berufs- und Bildungsforschung zählen. Daher wurde bereits 2014 das Institut für Linguistik, Sprachliche Informationsverarbeitung (Spinfo) an der Universität Köln mit einer Vorstudie beauftragt, um die Möglichkeiten und Voraussetzungen der automatisierten Informationsgewinnung für Stellenanzeigentexte zu erörtern. Auf dieser Grundlage wurde 2015 eine Kooperation zwischen dem BIBB und dem Spinfo vereinbart.

Vor der eigentlichen Informationsextraktion[4] wurde in der Vorstudie die Struktur bzw. der Aufbau der Stellenanzeigen ermittelt, womit bereits ein erster Schritt zur automatischen Textanalyse unternommen werden konnte. Ende 2014 wurde im BIBB ein Programm der Spinfo implementiert, welches die Texte vollautomatisch in klassifizierte Textblöcke nach dem folgenden Muster unterteilt:

- Selbstbeschreibung des Unternehmens
- Beschreibung der Stelle bzw. der auszuführenden Tätigkeiten
- Beschreibung der geforderten Bewerberkompetenzen
- Sonstiges (bspw. Kontaktinformationen und Angaben zum Bewerbungsverfahren)

[4] Informationsextraktion nennt sich der Vorgang, in dem unstrukturierte Information (meist in der Form von Texten) in strukturierte Information (bspw. in Form von Datenbankeinträgen) überführt wird (Cowie & Lehnert 1996).

Dieses Verfahren zur Klassifikation der Textabschnitte in Stellenanzeigen wird im Folgenden erläutert und insbesondere die Anwendung auf die Textsorte Stellenanzeige dokumentiert. Durch eine Evaluation unterschiedlicher Klassifizierungsalgorithmen wurde das für diesen Anwendungsfall am besten geeignete Verfahren ermittelt und später auf den gesamten Datensatz angewendet.

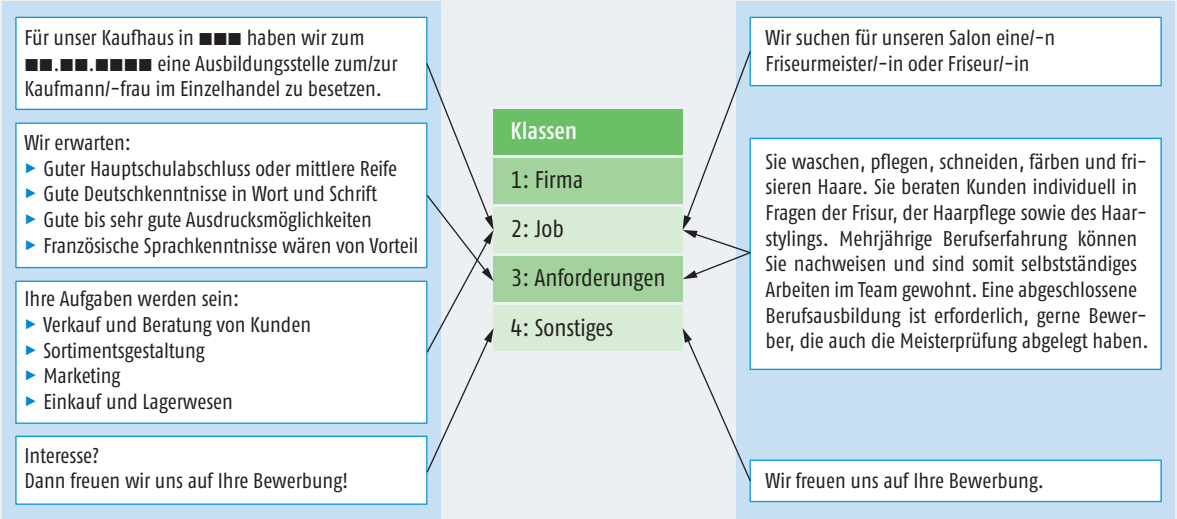
2 Struktur von Stellenanzeigen

Im Vergleich zu anderen Textsorten, wie etwa Zeitungsartikeln oder Erzählungen, handelt es sich bei Stellenanzeigen um eine relativ durchstrukturierte Textsorte – sie bestehen meist aus mehreren Abschnitten, welche spezifischen inhaltlichen Klassen zugeordnet werden können. Die inhaltliche Vorklassifikation der Abschnitte bietet sich als eine Vorstufe zur Extraktion von Informationen an, da auf diesem Wege der Suchraum für die zu extrahierenden Terme eingeschränkt werden kann. So finden sich die Anforderungen an den Bewerber bzw. die Bewerberin in dem Abschnitt, in dem die ausschreibende Stelle ihre entsprechenden Erwartungen formuliert. Die Vorgehensweise der Abschnittsklassifikation (engl. *Zone Analysis*) ist geläufig im Bereich der automatischen Verarbeitung wissenschaftlicher Beiträge in Zeitschriften, Sammelbänden und Proceedings. Derartige Beiträge folgen zum ganz überwiegenden Teil einer Struktur, die mit IMRAD bezeichnet wird (kurz für ***I**ntroduction, **M**ethods, **R**esults, **a**nd **D**iscussion*, vgl. Sollaci/Pereira 2004). Ansätze zur Zuordnung von einzelnen Sätzen oder auch Abschnitten zu den durch die IMRAD-Struktur vorgegebenen Inhaltsklassen wurden bereits mittels manueller (Teufel/Moens 2002) und maschineller Verfahren (McKnight/Srinivasan 2003, Mizuta et al. 2006, Agarwal/Yu 2009) erprobt.

Auch Stellenanzeigen lassen sich mehrheitlich auf ein spezifisches Schema zurückführen. Dieses kann man unter dem Kürzel AKJF fassen (für ***A**rbeitgeberinformationen, **K**ompetenzanforderungen an den Bewerber, **J**obinformationen und – als Default-Klasse – **F**ormalia/Sonstiges*). Im Regelfall sind diesen inhaltlich motivierten Klassen ganze Abschnitte von Stellenanzeigen zugeordnet, oft sogar in genau der angegebenen Reihenfolge. Zum Teil allerdings variiert diese Reihenfolge, in manchen Fällen müssen mehrere Abschnitte derselben Klasse zugeordnet werden, mitunter werden unterschiedliche Inhaltsklassen auch im selben Abschnitt behandelt. Alles in allem erschien es ratsam, tatsächlich vollständige Abschnitte bzw. Paragrafen (und nicht etwa einzelne Sätze) als zu klassifizierende Einheiten festzulegen und Klassifikationsverfahren zu mo-

Abbildung 1

Klassifikation von Abschnitten zweier anonymisierter Stellenanzeigen in inhaltlich vordefinierte Klassen



dellieren, welche Mehrfachklassifikationen ermöglichen. Dabei soll gewährleistet werden, dass sowohl mehrere Abschnitte einer Klasse (Abbildung 1, linke Seite) als auch einzelne Abschnitte mehreren Klassen zugeordnet werden können (Abbildung 1, rechte Seite).

3 Regelbasierte Klassifikation und maschinelle Lernverfahren

Für die Klassifikation von Texten existieren bereits eine ganze Reihe von Ansätzen (vgl. Sebastiani 2002). Wie in den meisten Bereichen der Verarbeitung natürlicher Sprache lassen sich dabei zwei Hauptrichtungen unterscheiden: *regelbasierte Verfahren* und *maschinelle Lernverfahren*.

Bei regelbasierten Verfahren werden durch den Anwender bzw. die Anwenderin feste Regeln codiert (im Beispiel von Stellenanzeigen wäre das z. B. die Regel: *wenn „Wir sind ein weltweit führendes Unternehmen im Bereich" dann Kategorie „Arbeitgeberinformation"*). Sie arbeiten sehr präzise, allerdings ist ihre Erstellung mit einem hohen Aufwand verbunden, da sämtliche anzuwendenden Regeln manuell codiert werden müssen. Um eine hohe Ausbeute zu erreichen, ist vor allem bei hoher Heterogenität der zu klassifizierenden Einheiten die Erstellung eines umfangreichen Regelsets notwendig.

Maschinelle Lernverfahren (ML) lassen sich meist mit weniger Aufwand als regelbasierte Verfahren umsetzen, da sich hierfür etablierte, offen zugängliche Algorithmen verwenden lassen und keine manuelle Codierung von Regeln notwendig ist. ML lassen sich in überwachte und unüberwachte Lernverfahren unterteilen. Für Klassifikationsprobleme werden in den meisten Fällen überwachte Verfahren eingesetzt. Für diese werden Trainingsdaten benötigt, die mit den entsprechenden Klassen vorausgezeichnet sind. Aus diesen Trainingsdaten wird ein Modell gebildet, auf dessen Grundlage ein Klassifikator Zuordnungen neuer, nicht vorausgezeichneter Daten vornehmen kann. Für ein solches Modell ist es notwendig, für die zu klassifizierenden Einheiten Merkmale zu definieren. Als Merkmale können dafür z. B. Kontext- oder Metainformationen herangezogen werden. Alternativ oder ergänzend dazu können Merkmale aus den zu klassifizierenden Einheiten selbst extrahiert werden. Im Bereich der Textklassifikation kommen als Merkmale z. B. die enthaltenen Wörter, n-Gramme (über Buchstaben oder über Wörter) sowie numerische Daten wie Wort- oder Satzlängen infrage. Die Modellbildung kann dabei auf sehr unterschiedliche Arten erfolgen. Verwendet man etwa Wörter, so ist es möglich, bestimmte Wortklassen auszuschließen, etwa hochfrequente Artikel, Konjunktionen oder Präpositionen (sogenannte Stoppwörter). Unterschiedliche Wortformen können auf die identische Stammformen (durch Lemmatisierer oder Stemmer) zurückgeführt oder normalisiert werden (etwa durch durchgehende Kleinschreibung oder Zusammenfassung von Zahlausdrücken). Zusätzlich untersuchten wir den Einsatz von Suffixbaum-Repräsentationen der Texte für die Merkmalsextraktion (vgl. Chim/Deng 2007), um die Abfolge von Einheiten berücksichtigen zu können (Geduldig 2015), sowie die Einbeziehung eines Mutual-Information-Filters (Xu et al. 2007), der die aussagekräftigsten Merkmale pro Klasse aufzufinden vermag.

Die meisten ML-Klassifikationsverfahren operieren auf numerischen Daten. Merkmale müssen in diesem Zuge auf numerische Werte abgebildet werden. Diese Überführung kann durch einfache Zählung der Häufigkeit und evtl. daran angeschlossene Gewichtungsverfahren (etwa durch das *Term-Frequency-/Inverse Document-Frequency*-Maß oder die *Log-Likelihood-Ratio*) gewährleistet werden.

Schließlich gilt es, neben Merkmalsauswahl und Merkmalsgewichtung noch das ML-Klassifikationsverfahren selbst auszuwählen und evtl. zu konfigurieren. Für diese Studie wurden Verfahren mit teilweise sehr unterschiedlichen Herangehensweisen gegeneinander evaluiert: erstens das sehr verbreitete und performante *Naive-Bayes-Verfahren*; zweitens der *Rocchio-Algorithmus* als auch performantes, lineares Klassifikationsverfahren; drittens *Support Vector Ma-*

chines (SVM, Geduldig 2014) und schließlich viertens der *k-Nearest-Neighbour-(KNN-)Algorithmus* als präzises, aber langsameres Verfahren.

ML-Klassifikationsverfahren wurden im Bereich der Zone Analysis bisher vor allem auf wissenschaftliche Veröffentlichungen angewendet. Die dort erzielten Ergebnisse lassen sich nicht ohne Weiteres auf die Zone Analysis in Stellenanzeigen übertragen, da diese im Allgemeinen kürzer sind, aber allem Anschein nach auch eine noch stärker ausgeprägte Binnenstruktur aufweisen. Aus diesem Grund wurde eine große Auswahl von Modellen gebildet und mit den vier unterschiedlichen ML-Klassifikationsverfahren kombiniert. Zu diesem Zweck wurde ein Framework entwickelt, das die Durchführung von mehreren Tausend Experimenten mit unterschiedlicher Konfiguration unterstützt und mit dem die erzielten Resultate miteinander verglichen und gerankt werden können.

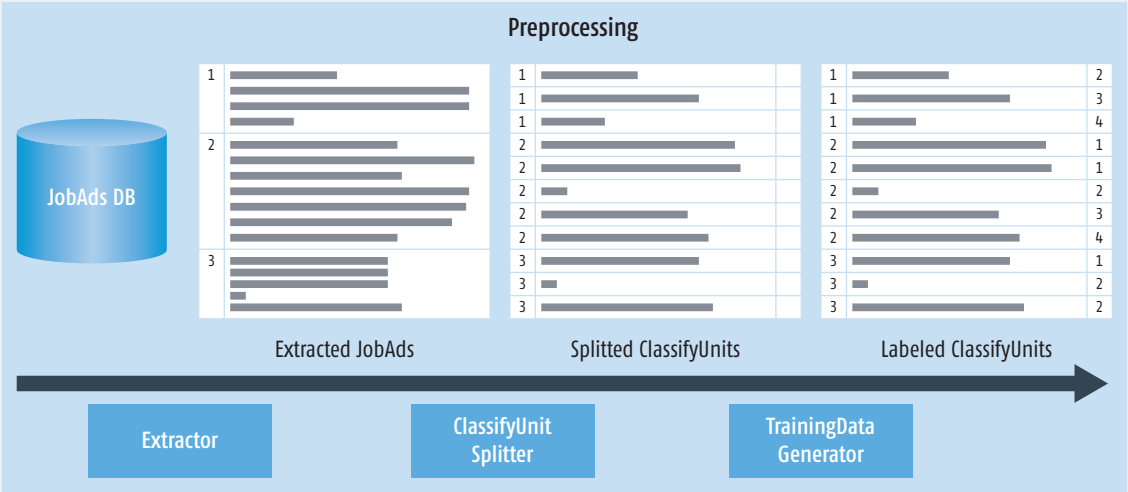
4 Umsetzung: Der Job Ad Section Classifier (JASC)

Für das Training der Klassifizierer und deren Evaluation wurden vom BIBB knapp 280 Stellenanzeigen anonymisiert und der Spinfo zur Verfügung gestellt. Mittels regulärer Ausdrücke wurden diese Stellenanzeigen in einzelne Paragrafen (knapp 1.500) unterteilt. Im Anschluss daran wurden sie manuell mit den entsprechenden Klassenlabels versehen. Der diesen Arbeitsschritten entsprechende Workflow ist in Abbildung 2 dargestellt.

Mehrfachklassifikationen sind mit den genutzten, gängigen Klassifikationsverfahren nicht ohne Weiteres möglich und mussten aus diesem Grund gesondert behandelt werden. Zu diesem Zweck wurden für mehrfach zu klassifizierende Abschnitte zunächst eigene Klassenlabels festgelegt. Wenn etwa geforderte Bewerberkompetenzen (Klasse 3) und die Beschreibung der Arbeitsstelle (Klasse 2) innerhalb eines Abschnitts behandelt werden, wurde dieser mit einem Label für Klasse 5 ausgezeichnet. Vor der Evaluation wurden diese Mehrklassenlabels wieder auf die vier Grundklassen zurückgeführt, sodass die tatsächlichen Precision- und Recall-Werte auch für die mehrfach ausgezeichneten Klassen berechnet werden konnten. Um ein Beispiel zu geben: Würde ein Verfahren einem Abschnitt die Klasse 3 zuordnen, der Abschnitt wäre aber mit Klasse 5 (also 2 und 3) gelabelt, würde das Evaluationsverfahren ohne Rückführung auf die Mehrfachklassen die Zuordnung als falsch positiven in Klasse 3 und falsch negativen in Klasse 5 bewerten. Mit Rückführung kann die Zuordnung als richtig positiver für Klasse 3 und falsch negativer für Klasse 2 gezählt werden. Insofern entspricht die Evaluation der Zuordnung nach der Rückübersetzung in die Mehrfachklassen im stärkeren Maße dem, was wirklich gemessen werden sollte.

Abbildung 2

Preprocessing – Von den Stellenanzeigen in der Datenbank zu manuell ausgezeichneten Abschnitten

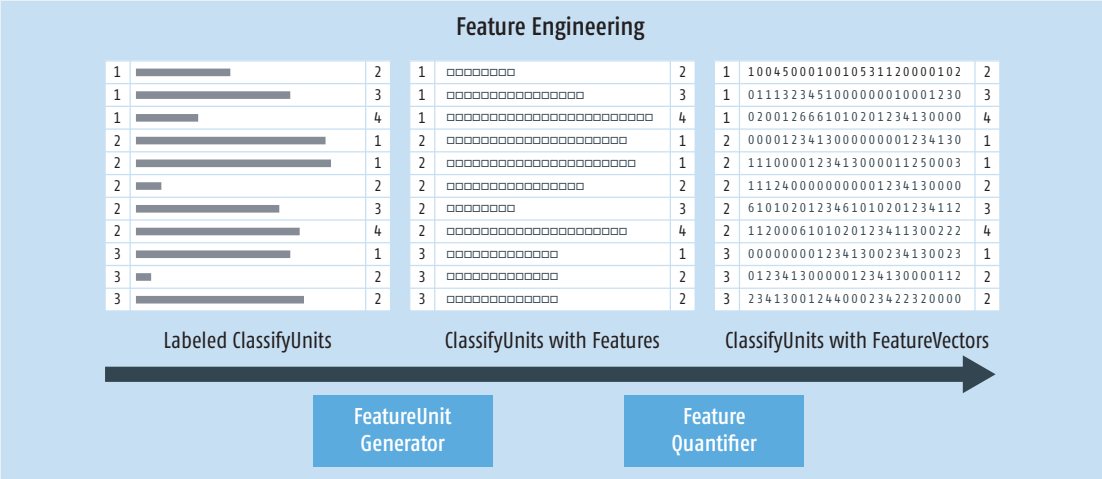


Mit dem Resultat des Preprocessing-Workflows können nun in einem weiteren Schritt die unterschiedlichen Modelle für Trainings- und Testphase der Klassifizierer gebildet werden. Dazu wurden die gelabelten Paragrafen (in der Abbildung 2 *ClassifyUnits* genannt) in zehn Gruppen unterteilt, von denen in einem Kreuzvalidierungsverfahren jeweils neun als Trainingsdaten und die

verbliebenen als Testdaten verwendet wurden. Mit den Trainingsdaten wurde über unterschiedliche Auswahlen und verschiedene Gewichtungen von Merkmalen insgesamt mehr als 1.000 Modelle gebildet – jedes durch einen spezifisch konfigurierten Workflow *Feature Engineering*, schematisch dargestellt in Abbildung 3.

Abbildung 3

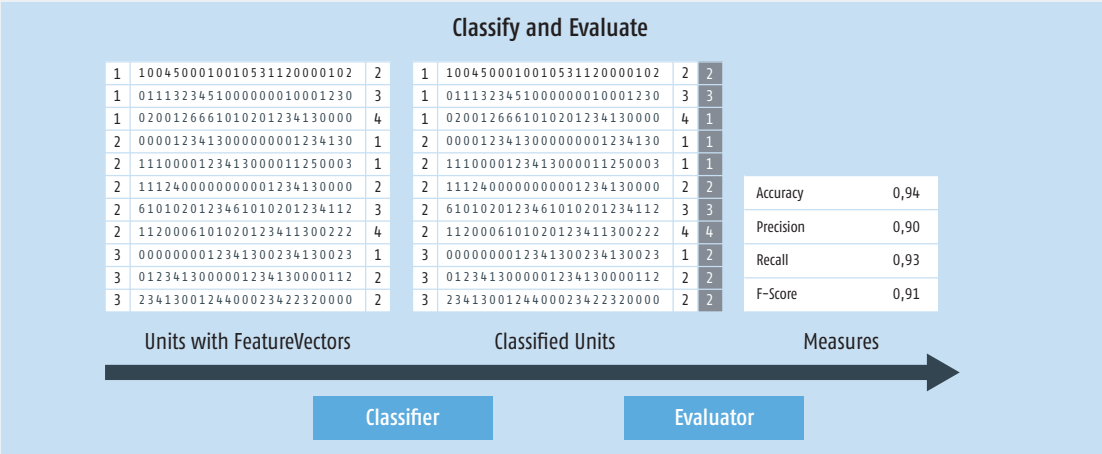
Feature Engineering – Bildung von Modellen für die einzelnen Klassen durch die Auswahl von Merkmalen über einen FeatureUnit Generator und deren Gewichtung über einen Feature Quantifier



Im Anschluss werden über die verschiedenen Modelle die Daten aus den Testsektionen der Kreuzvalidierungsgruppen anhand der drei verschiedenen ML-Algorithmen klassifiziert. Über den Vergleich mit dem Gold-Standard der manuell vorgenommenen Vorauszeichnungen können die gängigen Evaluationswerte Precision (Anteil der richtig klassifizierten Abschnitte an allen identifizierten Abschnitten), Recall (Anteil der richtig identifizierten Abschnitte an allen Abschnitten der jeweiligen Klasse), F-Score als harmonisches Mittel über die beiden vorangegangenen Werte und schließlich Accuracy (Anteil aller richtig identifizierten Abschnitte an allen Abschnitten) durch die Evaluator-Komponente ermittelt werden (Workflow *Classify and Evaluate* in Abbildung 4).

Abbildung 4

Klassifizierung der Testdaten und Evaluation gegen den Gold-Standard



Durch einen vorgeschalteten, regelbasierten Klassifizierer, der das Auftreten von 65 Ausdrücken auf die vorgegebenen Klassen abbildete, konnten sämtliche ML-Klassifizierungsverfahren signifikant verbessert werden. Dazu wurden 65 Ausdrücke ausgewählt, deren Einsatz eine Precision von nahezu 100 Prozent und einen Recall von knapp 50 Prozent ergab. Als Aufgabe für die ML-Klassifizierer verblieb, im Anschluss daran den Recall zu erhöhen, ohne die Precision allzu weit zu senken.

5 Ergebnisse

Für die Evaluation wurden Precison/*prec* und Recall/*rec* der Verfahren für die vier Inhaltsklassen ermittelt. Da die vierte Klasse (Formalia/Sonstiges) als eine Default-Klasse angesehen werden kann, wurde diese nicht in die Mittelwertberechnungen einbezogen, die in Tabelle 1 dargestellt werden. Aufgeführt sind jeweils die besten Ergebnisse der vier getesteten Klassifikationsalgorithmen. Wie vorhergesehen erzielt die relativ langsame KNN-Klassifikation die – mit einigem Abstand – besten Ergebnisse.

Die besten KNN-Verfahren liegen relativ dicht beieinander, solange für k ein Wert zwischen 4 und 6 gewählt wird, als Distanzmaß die Cosinus-Distanz und als Quantifizierer LogLikelihood genutzt werden. Durchgehend werden mit Buchstaben-n-Grammen (mit Kombinationen aus den Längen 2 bis 4), Einbeziehung von Stoppwörtern und SuffixTree-Repräsentationen bessere Ergebnisse erzielt als mit der Verwendung von Wörtern, dem Löschen von Stoppwörtern und ohne den Einsatz von SuffixTrees. Lemmatisierung und Normalisierung haben nur wenig Einfluss auf die Güte der Ergebnisse.

Die Präzision (precision/prec) der besten Verfahren ist mit einem Wert von über 0,98 sehr hoch, dagegen fällt der Recall mit 0,92 etwas ab. Beim Studium der einzelnen Ergebnisse ist das größtenteils dadurch zu erklären, dass Abschnitte, die im Gold-Standard mit mehreren Klassenlabes versehen wurden, nur als zu einer Klasse zugehörig erkannt wurden. Für eine Mehrfachklassifikation sind die Ergebnisse für die besten KNN-Verfahren (F-Score > 0,95, exemplarisch hier Tabelle 1, Zeile 1) dennoch durchaus beachtlich.

Tabelle 1
Ergebnisse der besten Experimente mit unterschiedlichen Klassifikatoren

Nr	fsco	prec	rec	accu	classifier	distance	quantifier	feature units
1	0,95	0,98	0,92	0,97	KNN (k = 4)	Cosinus	LogLike	2 & 3 Grams + SW + Norm + Stem + SuffTree
2	0,93	0,96	0,91	0,93	SVM	–	LogLike	3 & 4 Grams + SW + SuffTree
3	0,92	0,92	0,92	0,95	Rocchio	Cosinus	LogLike	2–3 Grams + SW + Norm + SuffTree
4	0,66	0,54	0,85	0,67	Naive Bayes	–	–	3 Grams + SW + Norm + SuffTree

Hinsichtlich der Merkmalsauswahl lässt sich feststellen, dass die Klassifizierer durchweg mit aus Suffixbäumen generierten Merkmalen und mit kurzen n-Grammen besser performen als mit ganzen Wörtern. Stoppwörter/*SW* zu filtern verschlechtert das Ergebnis, eine Normalisierung/*Norm* (hier: Kleinschreibung und Zurückführen von Numeralausdrücken auf ein Merkmal NUM) verbessert es leicht, eine Lemmatisierung hat dagegen so gut wie keine Auswirkung. Als beste Merkmalsgewichtung (*quantifier*) stellte sich das LogLikelihood-Verfahren heraus, als bestes Distanzmaß (*distance*) die Cosinus-Distanz.

Von den beiden getesteten linearen Verfahren liegt SVM leicht vor dem Rocchio-Algorithmus, was vor allem auf die höhere Präzision zurückzuführen ist. Mit einem F-Score von 0,93 (SVM, Tabelle 1, Zeile 2) bzw. 0,92 (Rocchio, Tabelle 1, Zeile 3) erzielen sie recht brauchbare Ergebnisse, fallen aber im Vergleich zu den KNN-Verfahren relativ stark ab.

6 Diskussion

Die Ergebnisse der Testphase, vor allem die für die KNN-Klassifikatoren, sind von überraschend hoher Qualität. Dem BIBB wurde daher der Einsatz eines KNN-Klassifikators zur Klassifikation der Abschnitte aus den Stellenanzeigen der BIBB-internen Datenbank empfohlen. Zusammen mit dem Code der Klassifikationssoftware wurden beim BIBB Modelle für verschiedene lineare und KNN-Verfahren implementiert.

Die Evaluation des produktiven Einsatzes des Klassifikationstools ist noch nicht endgültig abgeschlossen, denn an Stichproben wurde festgestellt, dass die sehr guten Ergebnisse der Testphase sich nicht ohne Weiteres übertragen ließen, wenngleich die Ergebnisse noch immer sehr brauchbar waren. Die Abweichung liegt vor allem darin begründet, dass die Modelle für die Klassifizierer aus anonymisierten Trainingsdaten gebildet wurden. Die Daten aus der Datenbank liegen dagegen nicht anonymisiert vor, was bedeutet, dass sich in diesen teilweise Merkmale finden, die nicht auf die Trainingsdaten zurückgeführt werden können. Aus diesem Grund wird in der gegenwärtigen Projektphase ein neues Trainingsset mit den Rohdaten erstellt, aus dem dann neue, präzisere Modelle gebildet werden können.

Weiterhin werden zurzeit unterschiedliche Verfahren der Informationsextraktion auf den vorklassifizierten Abschnitten der Stellenanzeigen erprobt. Dies geschieht auf Grundlage der ebenfalls im Umfang der Vorstudie geleisteten Arbeit von Neumann (2014). Vorrangiges Ziel der derzeit bis April 2017 laufenden Projektphase ist die Extraktion der geforderten Bewerberkompetenzen sowie der im ausgeschriebenen Beruf vorgesehenen Arbeitsmittel.

Literaturverzeichnis

Agarwal, Shashank; Yu, Hong (2009): Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. Bioinformatics (Oxford, England), 25 (23), 3174–3180. doi:10.1093/bioinformatics/btp548

Chim, Hung; Deng, Xiaotie (2007): A new suffix tree similarity measure for document clustering. In: Proceedings of the 16th international conference on World Wide Web, 121–130.

Cowie, Jim; Lehnert, Wendy (1996): Information extraction. Communications of the ACM, 39 (1), 80–91. doi:10.1145/234173.234209

Geduldig, Alena (2014): Textklassifikation mit Support Vector Machines. MS, Universität zu Köln, Institut für Linguistik, Sprachliche Informationsverarbeitung – URL: www.spinfo.phil-fak.uni-koeln.de/sites/spinfo/geduldia/SVM_Klassifikation.pdf, zuletzt aufgerufen: 09.12.2015.

Geduldig, Alena (2015). Evaluation des Suffixtree-Document-Models als Repräsentationsmodell zur Textklassifikation. MS, Universität zu Köln, Institut für Linguistik, Sprachliche Informationsverarbeitung – URL: www.spinfo.phil-fak.uni-koeln.de/sites/spinfo/arbeiten/SuffixtreeClassification.pdf, zuletzt aufgerufen: 09.12.2015.

Joachims, Thorsten (2001): A statistical learning model of text classification for support vector machines. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '01 (pp. 128–136). New York, USA: ACM Press.

McKnight, Larry; Srinivasan, Padmini (2003): Categorization of sentence types in medical abstracts. AMIA Annual Symposium Proceedings/AMIA Symposium. AMIA Symposium, 440–444.

Mizuta, Yoko; Korhonen, Anna; Mullen, Tony; Collier, Nigel (2006): Zone analysis in biology articles as a basis for information extraction. International Journal of Medical Informatics, 75 (6), 468–487. doi:10.1016/j.ijmedinf.2005.06.013

Neumann, Mandy (2014): Analyse von Anforderungsprofilen. Eine Studie zur Informationsextraktion aus Stellenanzeigen. MA Thesis, Universität zu Köln, Institut für Linguistik, Sprachliche Informationsverarbeitung – URL: www.spinfo.phil-fak.uni-koeln.de/fileadmin/spinfo/projekte/bibb/Masterarbeit_Neumann_final.pdf, zuletzt aufgerufen: 09.12.2015.

Sebastiani, Fabrizio (2002): Machine learning in automated text categorization. ACM Computing Surveys, 34 (1), 1–47. doi:10.1145/505282.505283

Sollaci, Luciana B.; Pereira, Mauricio G. (2004): The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. Journal of the Medical Library Association: JMLA, 92 (3), 364–367.

Teufel, Simone; Moens, Marc (2002): Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. Computational Linguistics, 28 (4), 409–445.

Xu, Yan; Jones, Gareth; Li, Jin Tao; Wang, Bin; Sun, Chun Ming (2007): A study on mutual information-based feature selection for text categorization. Journal of Computational Information Systems. Binary Information Press.

Der Code, der für die Durchführung der oben beschriebenen Experimente genutzt wurde, steht unter der Open-Source-Lizenz EPL 1.0 (Eclipse Public Licence, http://opensource.org/licenses/EPL-1.0) bei GitHub unter https://github.com/spinfo/jasc zur freien Verfügung.

Autoren

Dr. Hermes, Jürgen, Wissenschaftlicher Assistent, Sprachliche Informationsverarbeitung, Institut für Linguistik, Universität zu Köln, Albertus-Magnus-Platz 1, 50923 Köln, 0221 4704430, hermesj@uni-koeln.de

Schandock, Manuel, Wissenschaftlicher Mitarbeiter, AB 2.2 Qualifikation, berufliche Integration und Erwerbstätigkeit, Bundesinstitut für Berufsbildung (BIBB), Robert-Schuman-Platz 3, 53175 Bonn, 0228 1071212, schandock@bibb.de

Abstract

Für die Qualifikationsentwicklungsforschung des BIBB sind Stellenanzeigen eine aussagekräftige Informationsquelle. Darin werden u. a. Anforderungen, Arbeitsmittel und Tätigkeiten der ausgeschriebenen Stellen beschrieben – allerdings unstrukturiert und unsystematisch. Um eine statistische Analyse der Daten zu ermöglichen, entwickelt das BIBB mit der Universität zu Köln eine Methodik zur Extraktion von Informationen aus Stellenanzeigen. Ein Verfahren zur Klassifikation von Textabschnitten wird bereits mit Erfolg auf einer BIBB-internen Datenbank mit mehreren Millionen Stellenanzeigen eingesetzt und schafft damit die Grundlage für weitere Schritte zur Extraktion der Information aus den Volltexten von Stellenanzeigen.

For BIBB's qualifications development research, job advertisements are a significant source of information. The requirements, tools and tasks of the advertised vacancies are described therein – albeit unstructured and unsystematic. In order to facilitate a statistical analysis of the information, BIBB, in cooperation with the University of Cologne, developed a method of extracting information from job advertisements. A classification process of text extracts is already being successfully utilized in a BIBB internal database with several million job advertisements, thus creating the foundation for further steps of extracting information from the full text of job advertisements.

► Schlagwörter/Keywords

Text Mining
Text Classification
Information Extraction
Zone Analysis
Stellenanzeigen
Qualifikationsentwicklungsforschung

Bundesinstitut für Berufsbildung
Robert-Schuman-Platz 3
53175 Bonn

Telefon: (0228) 107-0

Internet: www.bibb.de
E-Mail: zentrale@bibb.de

Bundesinstitut
für Berufsbildung
BIBB

▶ Forschen
▶ Beraten
▶ Zukunft gestalten