Hintergrund: Der hunter®CV-EXTRACTOR

Die vom hunter®CV-EXTRACTOR (CVX) verwendeten Technologien des Text-Parsings und der Textanalyse kommen von der Firma DaXtra Technologies aus UK. DaXtra ist einer der Marktführer auf dem Gebiet der semantischen und linguistischen Textanalyse von Dokumenten.

Einen als Word- oder PDF-Datei übermittelten Lebenslauf zu verstehen, stellt für einen Menschen keine besondere Herausforderung dar. Intuitiv greift er für die Analyse und Einordnung der gelesenen Informationen auf eine Wissensbasis zurück, die er sich durch seine Lebenserfahrung erarbeitet hat. Für einen Computer und die darauf ablaufenden Softwareprogramme hingegen bedeutet jeder Text zuerst einmal nur eine lange Folge von Buchstaben, Zahlen und Zeichen. Diese Zeichenfolgen zu analysieren, in Beziehung zu setzen und daraus die Elemente zu extrahieren, die für das Einfügen in eine Bewerbermanagement-Software benötigt werden, ist Aufgabe eines CV-Parsers (CV= Curriculum Vitae; to parse = Analyse der Syntax). Ein solches Software-Programm liefert als Ergebnis etwa in hunter eine komplett ausgefüllte Eingabemaske mit allen extrahierbaren und relevanten Daten zu einer Person.

Herausforderungen des CV-Parsings

Für einen Computer ist es nach wie vor überraschend schwierig, Sprache zu verstehen.
Ein Grund dafür ist ihre Vielfältigkeit: Ein Datum kann auf verschiedene Weisen geschrieben werden. Ebenso kann ein Datum innerhalb eines Dokumentes unterschiedliche Bedeutungen haben, etwa für das Datum des Anschreibens, den Beginn der Ausbildungszeit oder das Ende einer Beschäftigung stehen. Um einen komplexen beruflichen Werdegang zu beschreiben, gibt es unendlich viele Möglichkeiten.

Diese vielen Wege, ein und dasselbe schriftlich auszudrücken, muss der CV-Parser erfassen können. Dies zu codieren bedeutet erheblichen Zeitaufwand. Die größere Herausforderung besteht jedoch in der Ambiguität. Ein Wort oder ein Satz kann je nach Kontext verschiedene Bedeutungen haben. Eine vierstellige Ziffer kann ein Jahr, eine Software-Version oder anderes meinen – je nachdem was die umliegenden Wörter der Zahl für eine Bedeutung geben. Die Parsing-Software muss diese Mehrdeutigkeiten in ihrem Kontext erkennen können.

Typen von Parsern und wie sie funktionieren

Es gibt verschiedene Vorgehensweisen, die Daten aus einem Lebenslauf zu extrahieren. Generell kann zwischen drei Typen von Parsern unterschieden werden:

- der Keyword-Parser,
- der grammatische und
- der statistische Parser.

Die simpelste Methode des CV-Parsings ist die Keyword-basierende. Der ***Keyword-Parser*** identifiziert Begriffe und einfache Muster im Text eines Lebenslaufes mithilfe von heuristischen Algorithmen. Er betrachtet dabei einen Begriff und bestimmt diesen unter Einbezug der angrenzenden Wörter. Sucht das Programm beispielsweise nach der Postleitzahl, versucht es die Wörter, die unmittelbar bei der Zahl stehen, als Adresse zu interpretieren.

Diese Form von CV-Parsing ist allerdings ziemlich ungenau, da hier keine Informationen extrahiert werden können, die nicht unmittelbar bei den entsprechenden Keywords zu finden sind. Ist ein Wort mehrdeutig, kann es leicht falsch interpretiert werden. Die Genauigkeitsrate dieses CV-Parsers liegt daher nur bei etwa 70 Prozent.

Der ***grammatische CV-Parser*** hingegen arbeitet mit Regeln, die den Kontext eines Wortes, innerhalb eines Satzes, erschließen sollen. Dieser Parser-Typ erkennt mehr Details und ist imstande, zwischen mehreren Bedeutungen in verschiedenen Kontexten zu unterscheiden. Grammatische CV-Parser können eine Genauigkeitsrate von über 90 Prozent erzielen. Der einzige Nachteil: Solch ein Parser benötigt zum einen viel manuelles Codieren von sprachererfahrenen Ingenieuren und zum anderen eine Menge Tests um sicherzustellen, dass Verbesserungen in einem Bereich nicht die Leistungen in einem anderen verschlechtern.

Der ***statistische Parser*** erkennt Strukturen im CV durch numerische Textmodelle. Auch dieser Parser-Typ kann verschiedene Kontexte von Begriffen erkennen. Um richtig genau arbeiten zu können braucht dieser Parser allerdings eine größere Anzahl von Lebensläufen, anhand derer er die Strukturen lernt, die zur Extraktion der richtigen Begrifflichkeiten führen sollen. Erfährt der CV-Parser zum Beispiel zehn Mal, dass Berlin eine Stadt ist, so wird er dies als Regel festlegen.

Ein statistischer CV-Parser arbeitet grundsätzlich präziser als ein Keyword-basierter, jedoch nicht so genau wie ein grammatischer, wenn der Parser zuvor nicht mit genügend Daten „trainiert“ wurde.

Zentrale Messwerte: Umfang und Genauigkeit

Um zu erfahren, wie gut ein Parsing-Programm letztendlich ist, sollte man bei der Auswahl darauf achten, in welchem Umfang es Daten übertragen kann und wie genau es dabei vorgeht. Grundsätzlich gilt für den Umfang: Je mehr Informationen ein Parser extrahieren kann, desto besser. Nahezu jeder CV-Parser erkennt Daten wie Kontaktinformationen des Bewerbers oder seinen beruflichen Werdegang. Manche, darunter auch der DaXtra-Parser, können jedoch darüber hinaus Referenzen, Hobbys, den gewünschten Arbeitsstandort und viele weitere Bereiche erfassen.

Die Fehlerfreiheit gibt an, wie oft der Parser richtig liegt. Eine Präzisionsrate von 95 Prozent bei der Namensidentifizierung bedeutet, dass der Parser in 95 Prozent der Fälle den Namen des Kandidaten korrekt erkennt. Dieser Wert ist wichtig, denn je geringer die Genauigkeit ist, desto höher sind Aufwand und Kosten, die für eine manuelle Korrektur entstehen. Ein guter Parser hat daher eine Genauigkeitsrate von mindestens 90 Prozent.

Der hunter®CV-EXTRACTOR basiert auf der Technologie des Marktführers

Der DaXtra-Parser ist eine Mischform aus statistischem und grammatischem CV-Parser und kombiniert die Vorzüge beider Typen. Er setzt sprachunabhängige statistische Modelle ein und ergänzt diese durch grammatische Regeln. Durch diese Kombination erreicht er eine Genauigkeitsrate von 95 Prozent. Zudem ist seine mehrsprachige CV-Technologie sehr ausgereift: So ermöglicht der DaXtra-Parser die automatische Datenextraktion für alle Regionen und in den Sprachen Deutsch, Englisch, Holländisch, Französisch, Spanisch, Polnisch, Russisch, Chinesisch und Japanisch. Diese Fähigkeiten machen die Parsing-Lösung von DaXtra zur aktuell besten auf dem Markt – und damit zur idealen Grundlage für den hunter®CV-EXTRACTOR.

Impressum & Kontakt:

Herausgeber: fecher GmbH
Otto-Lilienthal-Str. 12
D-63322 Rödermark

Telefon: +49 (6074) 80577-00
E-Mail: info@fecher.eu
Website: www.fecher.de

Geschäftsführer: Günter Hofmann
V.i.S.d.P.: Günter Hofmann