Requirements in Job Advertisements: Automated Detection and Classification Into a Hierarchical Taxonomy of Work Equipment (Tools)

Manuel Schandock
Federal Institute for Vocational Education and Training (BIBB)

BigSurv18, Barcelona, 27th Oct. 2018

- Job Postings as Data Source
- Detecting terms of interest
- ML Approach to sum up different terms into canonical terms and to classify this canonical terms into a taxonomy of tools
- Outcome

Job Postings as Data Source

In labour market research we face a serious lack of empirical information about ongoing trends for the employers demand for competences, skills and experience with technical equipment (tools).

- Job advertisements are a rich source of information for employers requirements.
- They are very up to date.
- And they are convenient and quite cheap to collect – in the case of using online sources.

Job Postings as Data Source

But there is also a challenge:

The vast amount of information in job advertisements is completely unstructured. Which means we have to deal with natural language texts and we have to dig for the information before we analyze the data.

Job Postings as Data Source

The data we use:

- job ads hosted by the German Federal Employment Agency (BA) from 2011 up to 2017 (99 Percent German)
- overall amount of more than 2.500.000 single job ads
- including a bunch of information by self-disclosure (such as industry, occupation, alternative occupation, et cetera)

Job Postings as Data Source

Wir suchen fur unsere Projekte im In- und Ausland in den Regionen Stuttgart oder Heilbronn-Franken

eine/n Architekt (m/w)

Sie entwickeln in einem interdisziplinaren Team Konzepte, Entwurfe und Layouts fur die Objektplanung. Sie gestalten ansprechende Produktions-, Buro- sowie Entwicklungs- und Laborbereiche.

Ihre Aufgaben: - Objektplanung und Gestaltung - Termin- und Kostensteuerung - Fuhren und moderieren von Besprechungen - Vertrags- und Baustellenmanagement

Ihr Profil: - Erfolgreich abgeschlossenes technisches Studium - Praktische Erfahrung in der Bearbeitung der LP 1-5 HOAI - Analytisches Denken, Planungs- und Organisationstalent sowie Kommunikationsfahigkeit zeichnen Sie aus - Sie arbeiten gerne im Team, fuhren Ihr Projektteam proaktiv und runden Ihr Profil durch gutes Englisch und den professionellen Umgang mit Software-Tools (CAD, MS Office, ggf. MS Project) ab

Wenn Sie diese Aufgaben interessant finden, senden Sie bitte Ihre aussagefahige Bewerbung mit der Kennziffer ### (vorzugsweise per e-Mail) an bewerbung@###.de.

Ansprechpartner ist Herr ###, der Ihnen fur Fragen unter +49 ### - ### gerne zur Verfugung steht

Detecting terms of interest

Wir suchen fur unsere Projekte im In- und Ausland in den Regionen Stuttgart oder Heilbronn-Franken

eine/n Architekt (m/w)

Sie entwickeln in einem interdisziplinaren Team Konzepte, Entwurfe und Layouts fur die Objektplanung. Sie gestalten ansprechende Produktions-, Buro- sowie Entwicklungs- und Laborbereiche.

Ihre Aufgaben: - Objektplanung und Gestaltung - Termin- und Kostensteuerung - Fuhren und moderieren von Besprechungen - Vertrags- und Baustellenmanagement

Ihr Profil: - Erfolgreich abgeschlossenes technisches Studium - Praktische Erfahrung in der Bearbeitung der LP 1-5 HOAI - Analytisches Denken, Planungs- und Organisationstalent sowie Kommunikationsfahigkeit zeichnen Sie aus - Sie arbeiten gerne im Team, fuhren Ihr Projektteam proaktiv und runden Ihr Profil durch gutes Englisch und den professionellen Umgang mit **Software-Tools** (**CAD**, **MS Office**, ggf. **MS Project**) ab

Wenn Sie diese Aufgaben interessant finden, senden Sie bitte Ihre aussagefahige Bewerbung mit der Kennziffer ### (vorzugsweise per e-Mail) an bewerbung@###.de.

Ansprechpartner ist Herr ###, der Ihnen fur Fragen unter +49 ### - ### gerne zur Verfugung steht

Detecting terms of interest

Work tools frequently can be found in certain regions (sections or paragraphs) of the job ads. Namely where the jobs being described and where the requirements are verbalized. Hence we have to search in that regions to find work tools.

1. employer/company: size, markets, famous produkts, locations, history, ..
2. job description: working schedule, tasks, working tools, ..
3. requirements: certificates, skills, working tools, ..
4. miscellaneous: contact information, formal informations, ..

Detecting terms of interest

Wir suchen fur unsere Projekte im In- und Ausland in den Regionen Stuttgart oder Heilbronn-Franken

eine/n Architekt (m/w)

Sie entwickeln in einem interdisziplinaren Team Konzepte, Entwurfe und Layouts fur die Objektplanung. Sie gestalten ansprechende Produktions-, Buro- sowie Entwicklungs- und Laborbereiche.

Ihre Aufgaben: - Objektplanung und Gestaltung - Termin- und Kostensteuerung - Fuhren und moderieren von Besprechungen - Vertrags- und Baustellenmanagement

Ihr Profil: - Erfolgreich abgeschlossenes technisches Studium - Praktische Erfahrung in der Bearbeitung der LP 1-5 HOAI - Analytisches Denken, Planungs- und Organisationstalent sowie Kommunikationsfahigkeit zeichnen Sie aus - Sie arbeiten gerne im Team, fuhren Ihr Projektteam proaktiv und runden Ihr Profil durch gutes Englisch und den professionellen Umgang mit **Software-Tools** (**CAD**, **MS Office**, ggf. **MS Project**) ab

Wenn Sie diese Aufgaben interessant finden, senden Sie bitte Ihre aussagefahige Bewerbung mit der Kennziffer ### (vorzugsweise per e-Mail) an bewerbung@###.de.

Ansprechpartner ist Herr ###, der Ihnen fur Fragen unter +49 ### - ### gerne zur Verfugung steht

Detecting terms of interest

Splitting the texts into paragraphs and classify

- Normalisation (reducing heterogeneity): lower cases, removing stopwords, removing punctuation, lemmatization
- Classification of paragraphs into 4 classes (bag-of-words + machine learning): build up some training data manually, compute feature vectors of words and n-grams, train a knn-classifier model (cross validated), classify all paragraphs
- We get a list of target values that indicates the class of each paragraph. Hence we can filter the paragraphs for information extraction.

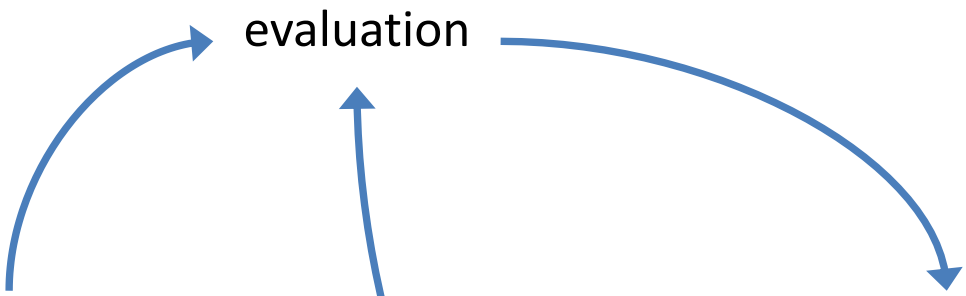
Detecting terms of interest

Initially we take frequently used phrases that indicates the occurrence of work tools (context). For example:

- .. safe handling of [work tool]
- You have to use [work tool] ..
- .. practical experience with [work tool]
- You'll have knowledge of [work tool] ..
- .. etc. (circa 20 phrases)

That is just to give an idea. It´s made more sophisticated by computer linguists.

Detecting terms of interest



Pattern matching

.. safe handling of [work tool]

You have to use [work tool] ..

.. practical experience with [work tool]

You'll have knowledge of [work tool] ..

.. etc. (circa 20 phrases)

Bootstrapping

itemizations, lists, conjunctions, ..

- .. [work tool], [work tool], [work tool], ..
- .. [work tool] and [work tool] ..
- .. [work tool] or [work tool] ..

BiBB

Detecting terms of interest

After 20 iterations over more than 2 million paragraphs the algorithm is stopped.

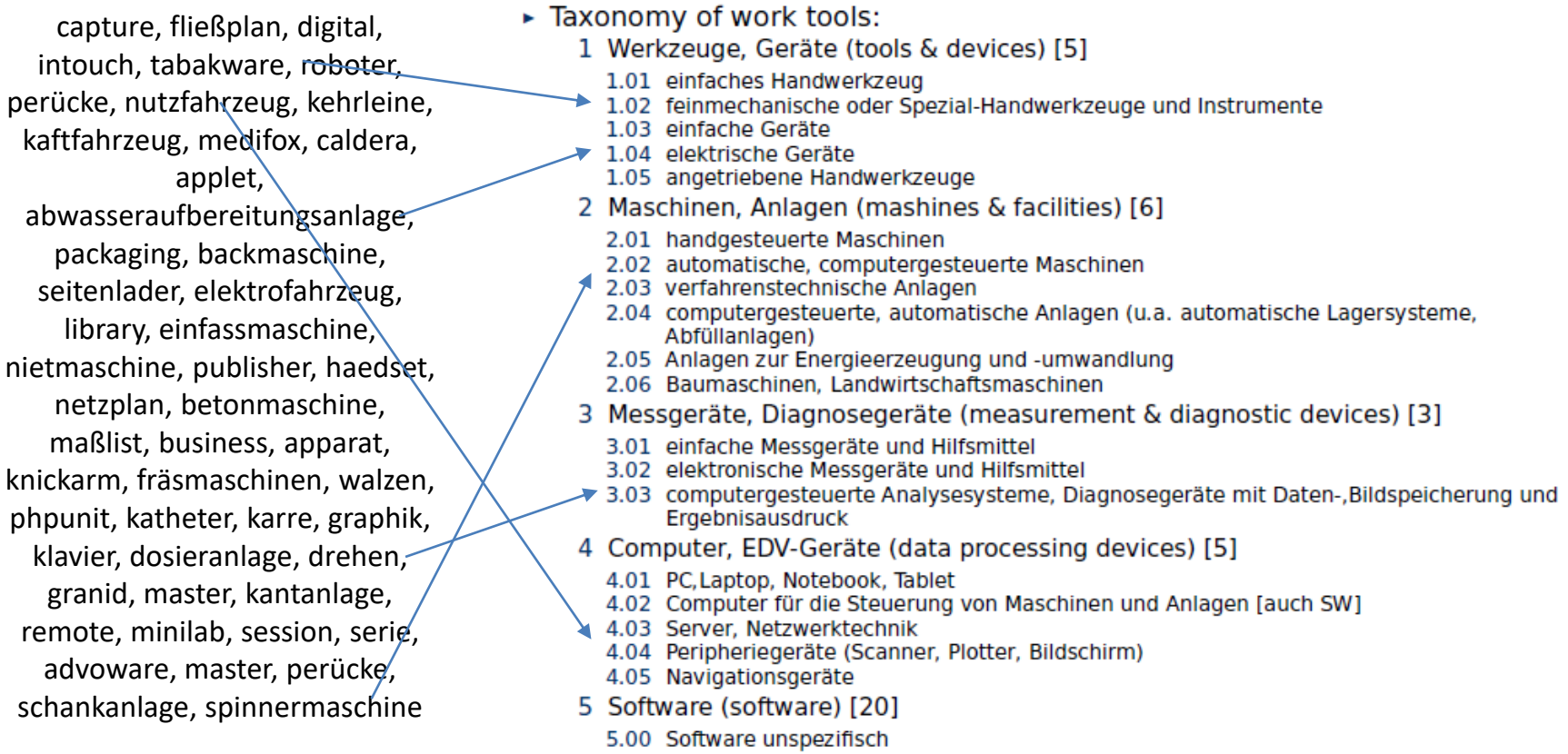
We end up with a list of more then 14.000 true positive extracted tools. (circa 50.000 false positive)

We have 2.6 million findings of tools in 1.2 million job postings

circa 3/5 of all job ads contain tools

ML Approach

capture, fließplan, digital, intouch, tabakware, roboter, perücke, nutzfahrzeug, kehrleine, kaftfahrzeug, medifox, caldera, applet, abwasseraufbereitungsanlage, packaging, backmaschine, seitenlader, elektrofahrzeug, library, einfassmaschine, nietmaschine, publisher, haedset, netzplan, betonmaschine, maßlist, business, apparat, knickarm, fräsmaschinen, walzen, phpunit, katheter, karre, graphik, klavier, dosieranlage, drehen, granid, master, kantanlage, remote, minilab, session, serie, advoware, master, perücke, schankanlage, spinnermaschine



- Taxonomy of work tools:
  - 1 Werkzeuge, Geräte (tools & devices) [5]
    - 1.01 einfaches Handwerkzeug
    - 1.02 feinmechanische oder Spezial-Handwerkzeuge und Instrumente
    - 1.03 einfache Geräte
    - 1.04 elektrische Geräte
    - 1.05 angetriebene Handwerkzeuge
  - 2 Maschinen, Anlagen (mashines & facilities) [6]
    - 2.01 handgesteuerte Maschinen
    - 2.02 automatische, computergesteuerte Maschinen
    - 2.03 verfahrenstechnische Anlagen
    - 2.04 computergesteuerte, automatische Anlagen (u.a. automatische Lagersysteme, Abfüllanlagen)
    - 2.05 Anlagen zur Energieerzeugung und -umwandlung
    - 2.06 Baumaschinen, Landwirtschaftsmaschinen
  - 3 Messgeräte, Diagnosegeräte (measurement & diagnostic devices) [3]
    - 3.01 einfache Messgeräte und Hilfsmittel
    - 3.02 elektronische Messgeräte und Hilfsmittel
    - 3.03 computergesteuerte Analysesysteme, Diagnosegeräte mit Daten-,Bildspeicherung und Ergebnisausdruck
  - 4 Computer, EDV-Geräte (data processing devices) [5]
    - 4.01 PC,Laptop, Notebook, Tablet
    - 4.02 Computer für die Steuerung von Maschinen und Anlagen [auch SW]
    - 4.03 Server, Netzwerktechnik
    - 4.04 Peripheriegeräte (Scanner, Plotter, Bildschirm)
    - 4.05 Navigationsgeräte
  - 5 Software (software) [20]
    - 5.00 Software unspezifisch

BIBB

ML Approach

Workflow:

- bag terms with same meaning to reduce amount of terms
  (string similarity, text2vec)
- start to assign bags/terms to taxonomy positions (manually)
- use the information of the assigned proportion to assign further bags/terms automatically
  (string similarity, text2vec)

ML Approach

What´s the information we could use:

- the spelling of terms (string similarity)
  gabelstabler gabelstaper gabelstapler gabelstappler gapelstapler

- the text within the terms occur (text2vec)
  "**gabelstapler** fahren, auch rückwärtiges rangieren mit anhänger"
  „idealerweise können sie hubwagen und **gabelstappler** bedienen?"
  „führen eines **gapelstaplers**"
  „innerbetrieblicher transport mittels **gabelstabler**"
  „transport von waren mittels **gabelstaper** und kran"

BIBB

ML Approach

string similarity (pairwise):

- many different metrics for string similarity are available: **Levenshtein, Jaccard, Soundex, Needleman-Wunsch, ..**
- But how to choose the correct threshold?

word	value
postgresql	0,7
postgressql	0,636364
postresql	0,6
postgres	0,5
postgress	0,454545
postgre	0,4
postgis	0,4
post	0,375
estg	0,375
nosql	0,333333
podest	0,3
postfix	0,272727
endtestgerät	0,266667
transportgestelle	0,263158
sql	0,25
stg	0,25
ges	0,25
prestige	0,25

ML Approach

Let ML do the decision:

- Set up a dataset with pairs of terms
- Compute as many similarity metrics (features) for each of them as you want
- Create a training set
- Train ML on train set
- Classify the entire dataset

id	Word1	Word2	F1	F2	..
1	Office	Offfice	#	#	..
2	Office	Ofice	#	#	..
3	Office	Officer	:	:	:
4	:	:	:	:	:
5	:	:	:	:	:
6	:	:	:	:	:
:	:	:	:	:	:

BiBB

ML Approach

Neural Networks with Feature Extraction

361 samples
66 predictor
2 classes: '0', '1'

Resampling: Bootstrapped (25 reps)

Resampling results:

Accuracy	Kappa
0.8525929	0.6708047

Tuning parameter 'size' was held constant at a value of 10
Tuning parameter 'decay' was held constant at a value of 0.05

ML Approach

führerschein	füchrerschein, führererschein, führereschein, fühhrerschein, führenschein, führererschein, führereschein, führerschei, führerscheihn, führerscheim, fuhrerschein, führerschein, führerscheine, führerscheinkl, führerschen, führerschie, führerschien, führerschin, führerschlein, führersein, führershcein, führershein, führersschein, führeschein, führrschein, fürerschein, fürherschein, fürhershein, fürhrerschein
buchungssoftware	buchungssoftwar, buchungssoftware
abkantpressen	abkanntpresse, abkanntpressen, abkantpress, abkantpresse, abkantpressen, akantpresse
warenwirtschaftssystem	wahrenwirtschaftsystem, warenwirtchaftssystem, warenwirtschafssystem, warenwirtschaftssystem, warenwirtschaftssystema, warenwirtschaftssystemen, warenwirtschaftssytem, warenwirtschaftsystem, warenwirtschaftsystemen
fräsmaschinen	fräsemaschine, fräsmachine, fräsmachinen, fräsmaschiene, fräsmaschin, fräsmaschine, fräsmaschinen
steuerung	steering, steuern, steuerung, steuerungen, steuerungs, steueung, steureung, steurung, steuuerung, streuerung, stuerung
maschinen	amschine, machine, machines, maschiene, maschiener, maschin, maschine, maschinell, maschinen, maschinene, maschiner

BiBB

ML Approach

Some false positives here:

montagebühnen demontageplan, montagebühn, montagebühne, montagebühnen, montagehilf

visualisierungssystem virtualisierungssystem, visualisierungsthema, visualisierungssystem, visualisierungssystema, visualisierunsgssystem, visualisirierungssystem

vorschriften vorschrif, vorschrift, vorschriften, vorschrit, vorspeis vortschrift

steuergeräte steuerbühne steuergerat, steuergerät, steuergeräte, steueringen

BIBB

ML Approach

Adding sense: Use the text within the terms occur (text2vec)

id	Word1	Word2	F1	F2	..
1	Office	Offfice	#	#	..
2	Office	Ofice	#	#	..
3	Office	Officer	:	:	:
4	:	:	:	:	:
5	:	:	:	:	:
6	:	:	:	:	:
:	:	:	:	:	:

+

Text2vec_sim
#
#
:
:
:
:
:

ML Approach

Neural Networks with Feature Extraction

361 samples
67 predictor
2 classes: '0', '1'

Resampling: Bootstrapped (25 reps)

Resampling results:

Accuracy	Kappa
0.8664847	0.6988073

Tuning parameter 'size' was held constant at a value of 10
Tuning parameter 'decay' was held constant at a value of 0.05

ML Approach

Neural Networks with Feature Extraction

361 samples
67 predictor
2 classes: '0', '1'

Resampling: Bootstrapped (25 reps)

Resampling results:

Accuracy	Kappa
0.8664847	0.6988073

Tuning parameter 'size' was held constant at a value of 10
Tuning parameter 'decay' was held constant at a value of 0.05

Outcomes

- A large amount of assignments can be done by the algorithm.
- It´s less expensive to pick out some aliens (false positives).
- With a growing knowledge base the algorithms should get better.
- We end up with a large dataset to deploy the taxonomy inside traditional data bases (open ended questions in survey data).

5.16 Datenbanken (databases):

,sql, acces-db, acces, access-anwendung, access-basiert, access-datenbank, access-tool, access control, access datenbank, access gateway, access point, access sql, access, access/excel, access/oracle, accessdatenbank, accesskenntnis, accsess, acess, acquiring-system, adabas/natural, administrationswerkzeug, adminstudio, agile, appgat, archivsoftware ser, archivsystem, bcdboot, bibliographisch datenbank, eprogesa, pcblut, ca-link, canalyser, canalyzer, cangraph, clearcase, cocreate, crm-datenbank, crm-system oracle, cronjobs, cucm, database-server, daten-bank, datenbank-gestützt programm, datenbank-server, datenbank-softwar, datenbank-tools, datenbank-werkzeug, datenbank sql, datenbank, datenbankapplikation, datenbankenanwendung, datenbankmanagement-system, datenbankmanagementsystem db2, datenbankmanagementsystem, datenbankmodelle, datenbankprogramm, datenbankserver, datenbanksoftware, datenbanksystem l1, datenbanksystem ms-access, datenbanksystem mssql, datenbanksystem mysql, datenbanksystem oracle, datenbanksystem quipsy, datenbanksystem, datenbanksystema, datenmanagementsystem, datentransfersystem, db-server, db-system mysql, db/2, db2-datenbank, db2, db2/informix, ddl, delphi sql, dns, domino-datenbank, dwh, echtzeitdatenbanksystem, edm-systemen, edm-umfeld, edm/pdm-system, edm/pdm/plm, elektromanager, energiedatenmanagementsystem, enterprisedb, esops, essbase, etl-tool, Eventviewer, ezb, fachdatenbank, filemaker, frontrange, git, helpdesk-system, hsql, hsqldb, imds, infomix, informix, it-helpdesk, jdbc, jee/oracle, jsp,awt servlets, kanaldatenbank, lfi-dbs, linq, m5-sql-server, maxdb datenbank, maxdb, microsoft-access, monitoringsoftware, monitoringsystem, monitoringtool, ms-query, ms-server-system, ms-server, ms-sql-server datenbank, ms-sql-server, ms-sql 2000-2012, ms-sql 2008, ms-sql server, ms-sql serversystem, ms-sql, ms¿sql¿server, ms®-sql-datenbank, msql, mssql-datenbank, mssql-server, mssql datenbank, mssql datenbanken, mssql server, mssql, mssql/mysql, mssql/oracle, msssql-server, my-sql datenbanksystema, my-sql server, my-sql, my sql, mysql-datenbank, MySQL 5, mysql, mysql/mssql, mysql/sql/oracle, mysql<U+00BF>flash, mysql¿flash, mysqldatenbank, nagios, netzwerkanalysator, nosql, olap-datenbank, online-datenbank, opencl, oracle-database, oracle-datenbank-kenntnißse, oracle-datenbank, oracle-datenbanken,

Thank you for your attention!