

# **There is no easy way around disambiguation to construct valid publication-based indicators for science-industry linkage**

Stephan Gauch<sup>\*,\*\*,\*\*\*</sup>

gauch@zhw.eu

\*Humboldt University Berlin, Department of Science Studies, Unter den Linden 6, D-10099 Berlin

\*\*German Centre for Higher Education Research and Science Studies (DZHW), Schützenstrasse 6a, D-10117 Berlin

\*\*\*Technical University Berlin, Chair of Innovation Economics, Marchstrasse 23, D-10623 Berlin

## **Abstract**

The goal of this paper is to assess the potential of different approaches to identify publications from industry organizations in address data provided in the Science Citation Index. Based on a semi-manually processed dataset of types of organizations as a benchmark two broad types of approaches are being utilized. One approach centres on identifying companies based on Legal Form Recognition with and without subsequent manual optimization. The second set of approaches is based on unsupervised machine learning classification approaches. By constructing a testbed of eight different supervised machine-learning algorithms, we can show that maximum entropy-based approaches provide the best solution for this type of problems. Yet, the overall results imply that both approaches are not suitable to arrive at a high quality classification without manual intervention. Reasons for the low level of quality are the highly skewed distribution of the problem with a large number of non-industry mentions as well as the amount of rare events of industry publications not being identified by their legal entity. In the conclusion of this paper different ways are discussed how to use these types of approaches despite their shortcomings. By combining the results of each of the employed methods near acceptable levels of Recall can be reached that can in subsequent manual steps be dealt with.

## **Keywords**

UIC, University-Industry Relationship, Corporate Publishing, Machine Learning, Classification, Industrial R&D

## **Introduction**

The amount of university industry-interaction is a long-standing area of interest multiple contexts both on organizational as well as on comparative macro level. In the Science and Innovation Policy domain indicators aiming at measuring the extent of university-industry relationship both are generally applied in the realm of co-patenting as well as co-publication seek to shed a light on the extent of this phenomenon on a comparative macro level. Organizational evaluations increasingly also capture the notion of this by providing similar measures such as shares of co-publications between universities and industry to account for phenomena. All of these approaches hinge on one substantial methodological caveat. The classification of organizations into different types, most notably into organizations from the realm of scientific research, such as universities or non-university research organizations on the one hand and organizations from the realm of industry on the other. Overall, the degree of coverage and accuracy of such classification or identification issues are central to the quality of the data itself and in consequence for the quality of the analyses and indicators conducted based upon this data (with respect to disambiguation, see e.g. Bourke & Butler, 1996; Van Raan, 2003). In general, the claim that such disambiguation is necessary has also been widely discussed concluding in the overall assessment that such disambiguation is pre-requisite to analysis both on the organizational as well as on the macro-level (e.g. Van Raan, 2003; Bador and Lafouge, 2005, Winterhager et al. 2014, Garcia-Zorita 2006).

Overall, attempts to disambiguate and classify publication address data dates back to the early 1990s, such as the works by De Bruin & Moed (1990) which approached the problem by development and maintenance of master files and rule-based techniques. With the advent of machine learning a number of approaches have been utilized to address this problem. Among the most notable efforts to arrive at a level of disambiguation that satisfies are the works by Galvez & Moya-Aneón (2006, 2007) with their work on transducer-based unification of organization address data via Finite State Graphs. Other approaches, such as the works from Morillo and colleagues (2013a, 2013b) employing semi-automated methods that also include the identification of keywords that reflect sectoral membership. Even though the latter approaches seem to operate fairly well on the whole for universities and public research organizations, the approaches seem to be less suited for classification of companies and “other types of organizations” with as much as around 20% of difference between within hand coding compared to automatic coding. (Morillo et al., 2013a) Other approaches such as developed by Guo et al. (2009) with a special focus on Chinese affiliation data apply a Latent Semantic Association to the data to harmonize datasets. Similar to the above approach this method aims at identifying association keywords and terms that have a high discriminant quality but also takes into account the relative position within strings. While the results show promise regarding the overall precision and recall of the classification the authors give no information in their paper if the performance of this method is dependent on organization type. Cuxac et al. (2013) propose a Naïve Bayes approach arguing that this approach will be superior in differentiating multiple concurrent problems in address data. By employing a two-stage approach the first aim to identify a suitable supervised learning algorithm to classify a completely in hand data source and based on the data and in a second stage aim to identify semi-supervised approaches to eliminate the use of the training dataset by identification of association terms by using specific flavors of multivariate k-means based algorithms which account for overlapping clustering in the corpus. Common to most of these approaches are multi-stage processes involving either automatic or semi-automatic pre-processing of data and application of machine learning algorithms. Also, all research in this area suggest that the need for manual effort can not completely be discarded.

The problem of identifying organizations of the type of industry is related to the problem of disambiguation in many ways. First and foremost, similar methods can be applied to address this problem ranging from rule-based approaches to advanced supervised learning techniques. Second, attribution of sectoral membership in some approaches are a by-product of the overall disambiguation attempts. Sometimes even used as a supporting information to disambiguate organizational address information. Yet, the problem of sectoral classifications can be argued as being a less complex problem compared to fully developed disambiguation approaches. Constraints on the quality of the results are considerably lower and may therefore be achieved by less efforts. This is especially relevant in cases where large-scale international comparative studies are among the goals of indicator construction. We propose that embracing a less complex approach to sectoral classification might bridge a gap between the status-quo of largely not fully disambiguated country-specific corpora and a hopefully future state of existence of such high quality data. This motivation led us to use the fully disambiguated corpus of German affiliations within Web of Science data as available within the German Competence Centre for Bibliometrics to conduct a case study aiming at identifying if a less strict approach can reap valuable benefits to arrive at timely indicators of science-industry linkage. In short our question can be formulated as follows:

*“Can lightweight approaches with low manual effort provide high quality classification of industry vs. non-industry organizations?”*

To address this question we devised various test-beds aimed at benchmarking different approaches against a well maintained semi-automated and quality-checked classification of German affiliation data available to organizations of the German Competence Centre for Bibliometrics. We explicitly excluded approaches that rely on matching large scale commercial or government company register data. The reason for this is twofold. First, large scale company registers are not universally available for all countries that might be relevant for this type of analyses. Second, even well kept commercial registers can be incomplete, especially in the light of small businesses that do not feature extensive

fiscal reporting regimes. Such approaches also are impacted by a time-lag between founding a company, their inclusion in such registers and the point in time of availability. Such approaches therefore would underestimate crucial parts of science-industry linkage, e.g. the impact of spin-offs or start-ups. Rather than relying on processing such register data, we want to focus on approaches that upon applicability can be utilized directly presupposing availability of address data within publications.

First, we employed a quasi-naïve regular-expression-based Legal Form Recognition (LFR) approach, i.e. classifying an organization by presence of their legal form such as “GmbH” within an affiliation string. Such approaches can be considered a short-cut to more elaborate and time-intensive approaches such as the ones proposed by Morillo et al. (2013a, 2013b). Rather than identifying suitable keywords from a pre-classified corpus, we want to assess if legal forms, which are more easily to be identified and require no pre-classification effort lead to acceptable results. One main reason to follow this approach is the low level of effort necessary to arrive at such a set of keywords. The extent of such lists of legal forms is considerably limited and can be collected from various sources shifting the effort from manual coding of affiliation data towards less time consuming pattern matching approaches using plain matching or advanced matching approaches such as regular expression matching or fuzzy string matching. Despite the argument for a lower manual effort, the list of legal forms can be easily extended towards the international domain by extending legal forms by national or supranational legal forms. Therefore, if such an approach proves to provide good results an extension towards the international realm is a matter of simple extension of such a list. Moreover, such an approach can build upon previous work from the patenting domain, which we will discuss in the methodological part. Still, such approaches based on legal forms will have its caveats, which mainly depends on the overall mix of university and extra-university research organizations within a country. The higher the share of extra-university research organizations that feature a legal form that implies them to be incorrectly sorted into the industry sector, the lower the precision of such approaches. Yet, we argue that such approaches might still provide a benefit for future classification efforts as the pool of organizations subjected to manual subtractive processing will be considerably smaller than processing the overall pool of organizations in question. Germany, with its considerably high share of extra-university research organizations provides a good starting point to assess issues that are linked to this problem. Another challenge to these types of approaches lies in the practice of authors from industrial organizations to mention the legal form in the address string of a publication. In contrast to address data within patents, in which the legal form has a substantial impact on the ownership of the intellectual property vis a vis the patent applicant, we have to assume such a practice. An assumption, which as we can show in this paper is not universally correct and has more far reaching consequences on other types of approaches. In cases where such information about legal forms is not provided by an author from a company, the approach will identify fewer organizations as accurate, i.e. a lower recall for the approach selected.

Second, we used the previously mentioned high quality classification provided by Bielefeld<sup>1</sup> as a starting point for applying supervised machine learning algorithms that are focused strictly at solving the classification of industry organizations within the data. The testbeds designed in light of the latter approach were explicitly aimed at a second objective. Namely, if these supervised approaches offer a sweet spot, i.e. an absolute number of or a relative share of organization address strings that upon manual or semi-manual classification can lead to results that meet a high level of accuracy and coverage and can therefore be applied to produce high quality indicators. In contrast to LFR, these approaches rely on learning algorithms that use a set of preclassified objects to inductively classify unknown objects. The quality of such approaches generally increase with the share of pre-classified objects vis-a-vis the total set of objects to be classified. Such approaches might alleviate some of the shortcomings that eventually might be prevalent in the first approach, i.e. the lack of consistent mentioning of legal form in cases where industry organizations appear frequently with their actual organization name but legal forms are not mentioned in all of the cases. Such cases would not be identified as industry applying the first approach. These methods have the potential of shifting the

---

<sup>1</sup> For the remainder of this paper we will refer to this classification as the „Bielefeld classification“.

manual effort of classification towards only classifying a fraction of organizations and selection of appropriate algorithms to solve the classification problem. In order to reflect this, our test-beds feature two variable components. First, we employed multiple supervised machine learning algorithms to assess which of these algorithms provide the best results. Second, we modified the amount and share of training data, which we derived from the high quality classification, to assess if such an aforementioned sweet spot exists. The latter aspect is especially relevant as it will provide a first assessment of efforts that have to be addressed to replicate the results for data from other countries.

### **Data sources and data pre-processing**

The data utilized to conduct this study was extracted from the in-house data source of the German Competence Centre for Bibliometrics. One of the central features of this database is the diachronically accurate Bielefeld classification. We limited our analysis to publications listed within the Science Citation Index Expanded. From this data source the individual address data was extracted based on the following constraints:

- The address has to correspond to an organization situated in Germany.
- The address has been used in terms of a publication between 2000 and 2014.
- The address has been used on a publication that is among the class of citable items.

For this corpus we extracted the original information address information, the sectoral attribution as being coded within the process of disambiguation, the organization identifier as attributed post-disambiguation. Applying the above constraints we arrived at a dataset comprising a total of 368010 entries. These entries are distributed among the following eight sectors “University” (UNIV), “Organizations attributed to the Fraunhofer Society” (FHG), “Organizations attributed to the Helmholtz Association” (HGF), “Organizations attributed to the Max-Planck Society” (MPG), “Organizations attributed to the Leibniz Association” (WGL), “Government Labs (GOV)” (RES), Industry (IND) and “Other Organizations” (OTH)<sup>2</sup>. For all later purposes we regard and treat these disambiguated data as accurate in these sense as they have been manually post-processed and devoid of any attribution errors.

Table 1 shows the overall distribution of the raw number of addresses and their sectoral attribution as well as the accurate number of organizations after disambiguation<sup>3</sup>. The data reflects that the disambiguation efforts have an enormous impact on the overall number of organizations in question. We also find, that the practice of unified mentions seems more homogenous for companies. Especially the reductive performance of harmonizing university addresses suggest, that the approaches employed in this paper can not directly be related to approaches in which classification by type is executed post disambiguation stages as the number of organizations in the case of universities is drastically reduced in such a step, which will have an impact at accuracy assessments of algorithms. Still, as we aim at assessing light-weight approaches to the problem we will work on the original data, rather than disambiguated data. This also reflects the point of departure for classifying industry type organization in hitherto not standardized and disambiguated national data.

---

<sup>2</sup> The type “Other organizations“ mainly reflect organizations in the health sector such as clinics.

<sup>3</sup> The authors want to highlight that the shares reported in Table 1 do explicitly not reflect the share of absolute numbers of publications attributed to these types of organization but rather the number of individual organizations present in the data set.

*Table 1: Distribution of raw number of addresses and number of disambiguated organizations*

	Raw number of addresses		Number of Organizations (disambiguated)	
	addresses	Share	(disambiguated)	Share
FHG	3761	1,0	71	4,2
HGF	21905	6,0	21	1,2
UNIV	282169	76,7	239	14,0
MPG	8265	2,2	88	5,2
OTH	24252	6,6	618	36,3
RES	6418	1,7	65	3,8
WGL	10636	2,9	98	5,8
IND	10604	2,9	502	29,5

Prior to the next steps the address data was pre-processed to extract the actual organization name including eventual mentions of legal forms along with potential substructes. To achieve this the address string was split by “,” and total length of the parts per address was assessed. In a next step, stripped the trailing right elements of each of the multi-part strings, which usually within the SCI address data refers to country and city with or without postcode. In cases in which the remaining multi-part strings exceeded the length of 3 the first two elements were extracted. In cases where the length was lower than 3 only the first element was extracted. All strings were transformed to lower-cases.

#### **Legal form recognition (LFR)**

The first set of approaches centers on applying a LFR approach. In order to arrive at a reliable list of legal forms we used a list compiled by Magerman et al. (2006). This list has not been specifically developed in order to harmonize publication address data but rather unify different iterations of legal forms in patent applicant data. The list contains a total of 1061 different iterations of international legal forms. To reduce computational effort the list was transformed from a simple plain text matching towards regular expression based matching reducing the overall list to 234 items that represent legal forms. These include abbreviations of legal forms such as “Gesellschaft mit beschränkter Haftung” (ges.\*[ ]+beschr.\*[ ]+haft.\*\$) to abbreviations such as “GmbH” (gmbh\$). A first set of search terms was set up in a way that legal forms were always identified as being situated towards the end of the multi-part strings. Moreover, a second set of regular expressions were designed that account for intermediary position within a string. In this case the search term was encapsuled in whitespace before and after the string. The matching was then executed on the multi-part strings and Precision and Recall were calculated. The approach yielded a total of 11676 matches for type “Industry” of which a total of 7689 of the total of 10604 classified correctly as companies. 3987 organizations were incorrectly assigned as companies. 2915 companies of the total of 10604 were not captured using this method. The results show that this approach seems to perform only in a very mediocre fashion with a Precision of 65.8% and a Recall of 72.5%.

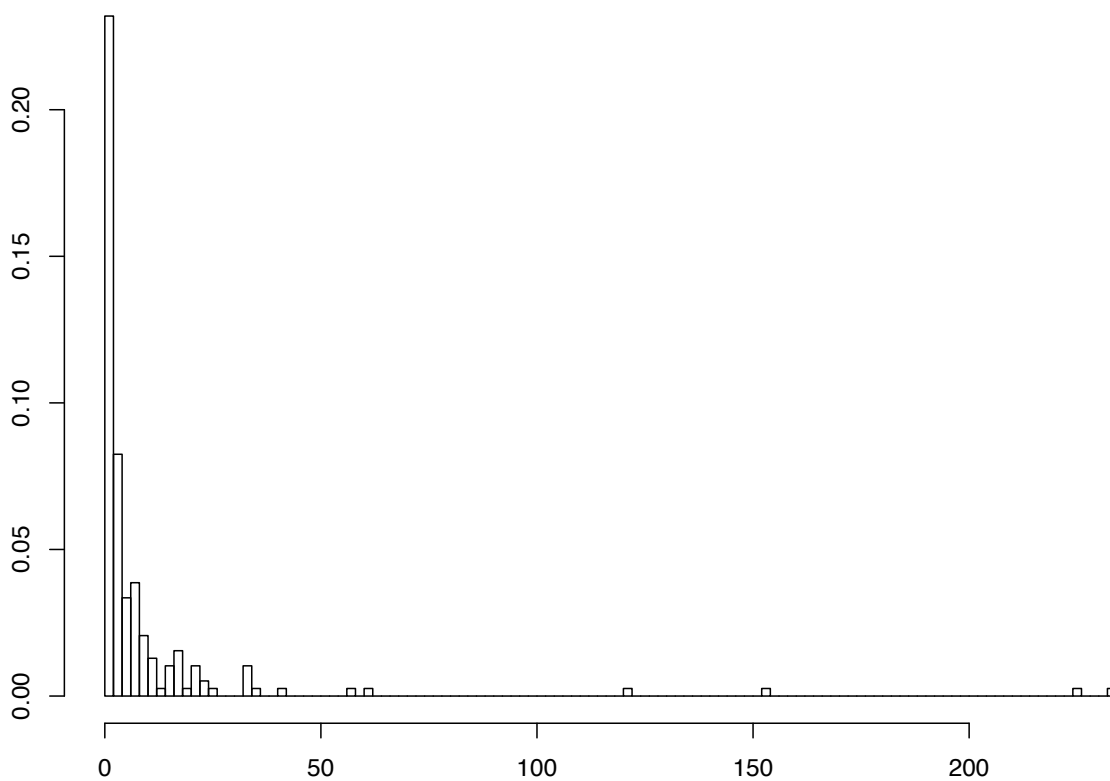
*Table 2: Distribution of False Positives by Type of Organization*

Type of Organization	False Positives
FHG	89
HGF	1729
HS	812
MPG	110
OTH	1084
GOV	25
WGL	138

The results were then checked qualitatively to assess which problems led to these low levels of Precision and Recall. As expected the results show that legal form can not universally be attributed to

the type “Industry” and public research organizations do not universally exclude their legal form from the affiliation as mentioned in publications leading to high amount of False Positives for organizations by Helmholtz Society. For Fraunhofer Society most of the False Positives could be attributed to the Heinrich-Hertz Institute and Fraunhofer ICT-IMM. For Max Planck Society most False Positives could be identified as due to the Max-Planck Institute für Eisenforschung. For WGL most False Positives were identified as “Leibniz Center for Tropical Marine Ecology”, “German Primate Center” and “German Institute for Economic Research”. For the type other a substantial amount of organizations belong to the health sector, i.e. clinics not directly connected to a university or public research organization. False Negatives were checked qualitatively. The results show that among those organizations a large amount can be accounted by large multinational and national companies. Yet, the distribution of disambiguated organizations in the set of False Negatives show a tendency for organizations appearing only once in the set of affiliations (see Figure 1).

Figure 1: Distribution of divergent mentions of False Negatives



In order to reduce the amount for False Positives a subtractive approach was applied to exclude organizations. A set of search terms was constructed that was inspired by capturing the respective Organizations by mentions of the umbrella group.<sup>4</sup> Moreover, search strings that could identify large public research organizations under this umbrella groups was constructed. We explicitly coded in a way that would be plausible ex-ante to an expert of a national science system. We did not account for errors in spelling or organizations that did not include a reference to their umbrella organizations as such modifications would not be ex-ante inferrable without substantial manual efforts. Applying this approach led to the following results as shown in table 3.

<sup>4</sup> Examples for such search strings are “Fraunhofer”, “FHI’s”, “FHG’s” and respective search terms for other organizations.

*Table 3: Distribution of False Positives by Type of Organization*

Type of Organization	False Positives
FHG	9
HGF	72
HS	789
MPG	0
OTH	1055
GOV	25
WGL	27

For Fraunhofer Society most False Positives could be eradicated. Same for Max-Planck Society for which all False Positives could be accounted for. In terms for WGL and HGF the problem persisted mainly due to generic mentions such as “Forschungszentrum GmbH”. The result of this assessment also show that the main problems with False Positives are due to a high amount of organizations by the type “other” (see Table 2). These results are in line with the implications regarding the quality of disambiguation for these types of organization as shown by Morillo et al. (2013a). For universities most organizations seem to be early level spin-offs which still hold a connection to the universities and mention these accordingly. These attempts led to an improvement of Precision (79.4%). Recall though could not be improved significantly (72.6%). In this case a complementary approach matching a company register could indeed account for some of the problem. Yet, the potential for improvement seem to be limited to multinational and large companies, as the amount of smaller companies which are rarely mentioned is rather high (54% for less than 3 mentions).

#### **Approaches based on supervised machine learning**

A second set of approaches was motivated by the question about the overall performance of machine-learning algorithms in a pre-disambiguation setting. Moreover, we attempted to identify if there is a sweet spot, i.e. a minimum size of randomly selected and coded training data that yield acceptable results. In order to achieve this a test-bed was constructed that feature the 8 most commonly applied machine-learning classification algorithms (see table 4)

*Table 4: ML-Algorithms implemented in classification test-bed*

Classification algorithm	Abbreviation	Source
Maximum Entropy Classification	MAX ENTROPY	Nigam et al. (1999)
Support Vector Machines	SVM	Boser et al. (1992)
Lasso and Elastic-Net Regularized Generalized Linear Models	GLMNET	Friedman et al. (2008)
Random Forrest Classification	FORREST	Breiman (2001)
Supervised Latent Dirichlet Allocation	SLDA	Blei & McAuliffe (2007)
Regression Tree Classification	TREE	Breiman et al. (1984)
Except Stochastic Gradient Boosting	LOGIT-BOOST	Friedman (1999)
Bagging Classification	BAGGING	Breiman (1996)

The test-bed was set up in a way that reflects a random approach to sampling test-data, i.e. a training dataset was randomly selected from the corpus and the Bielefeld Classification used as training information for the algorithm. In a first step the complete corpus was transformed into a Vector Space Document-Term Matrix. To identify if despite the random sampling and the thereby variation of probabilities of training data selected a sweet spot can be identified the test-bed started with a training data size of 2000 organizations. This parameter was then increased in steps of 200. For each step a new random sample of training data was drawn and a classifier was modeled based on this training data. In a subsequent step the classifier was applied to the remainder of the dataset.

Figure 2: Precision values for multiple learning algorithms under varying training set sizes

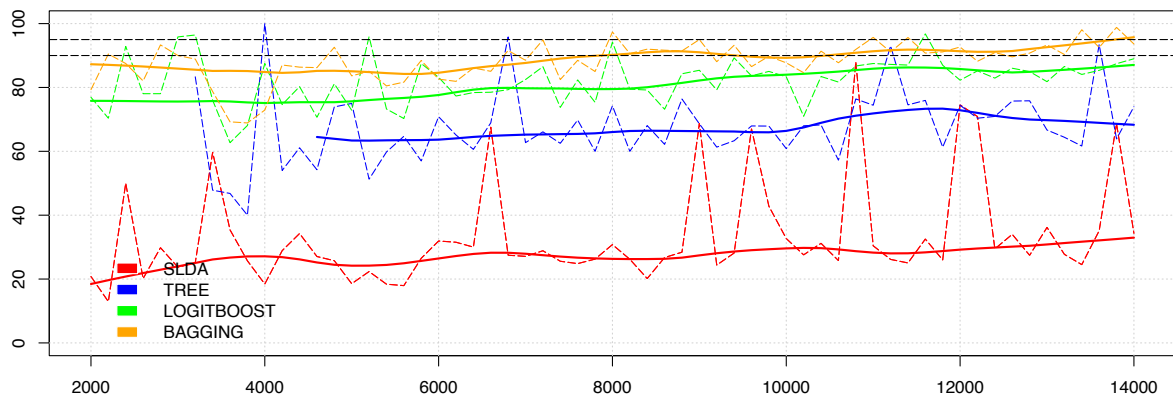
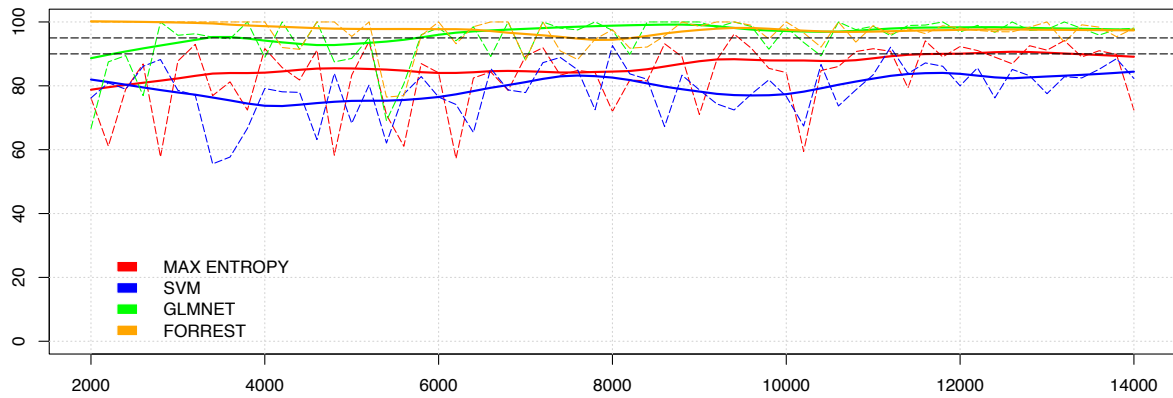
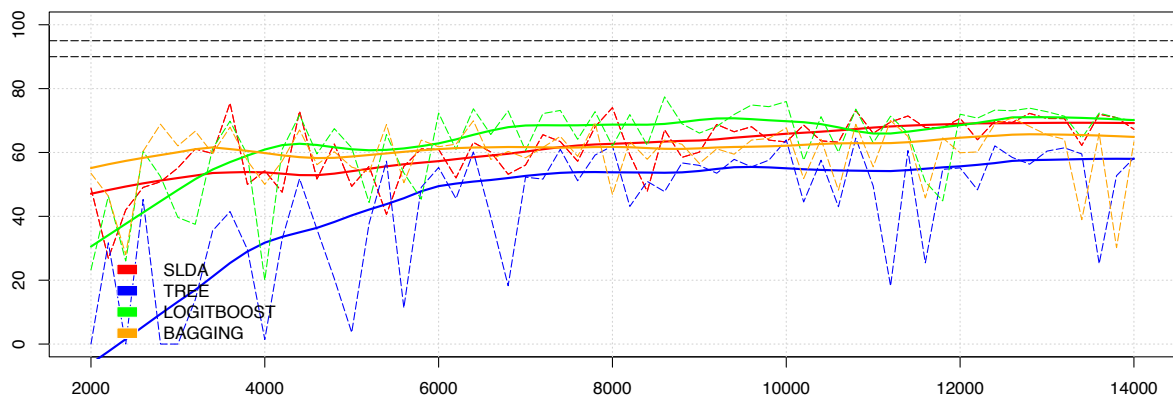
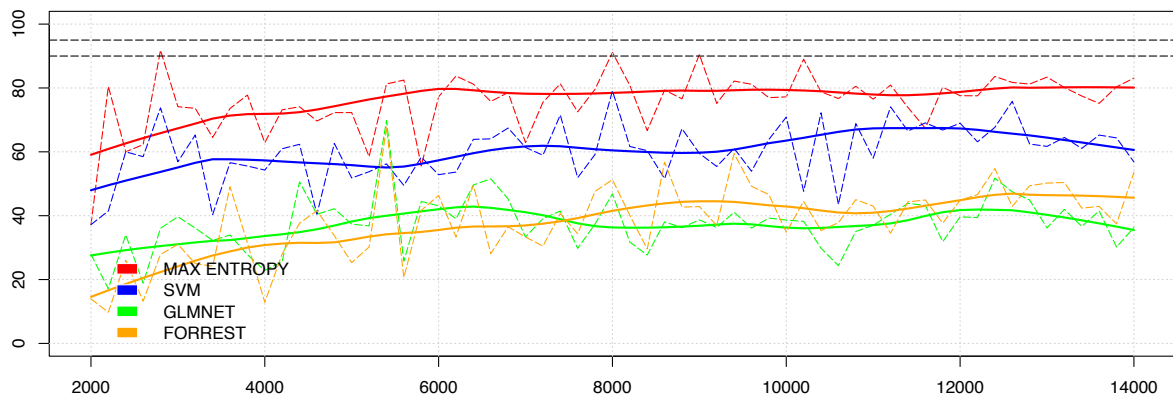


Figure 3: Recall values for multiple learning algorithms under varying training set sizes



For each algorithm we chose a threshold of .90. Subsequently, Precision and Recall values were calculated for each run of the classification attempt by each algorithm.

As the number of industry organization only accounts for 2.9% of the pre-disambiguated address data we expect some level of variation between runs. Yet, we argue that this reflects a realistic approach without ex-ante selection criteria of addresses from the data. To account for this we calculated a lowess-smoother for Precision and Recall over the increasing training data set size. The results of the test-bed are shown in Figure 2 (Precision Values) and Figure 3 (Recall Values).

The results imply that none of the algorithms yield an overall satisfying result and that therefore manual processing of the data can not be avoided. While reaching high levels of Precision seems to be possible for certain algorithms the major drawback lies in low values for Recall. This is in line with our previous attempt to identify organizations via LFR approach. Overall Recall values approach levels of about 80% thereby fairing better than via LFR. This is mainly due to the fact, that this approach accounts for variation in mentioning legal forms for large companies and therefore ameliorate some of the previously observed shortcomings. Yet, the algorithm can not resolve crucial cases that we found in our previous attempt, namely rare events of industry organization mentions without any plausible identifier for the type industry.

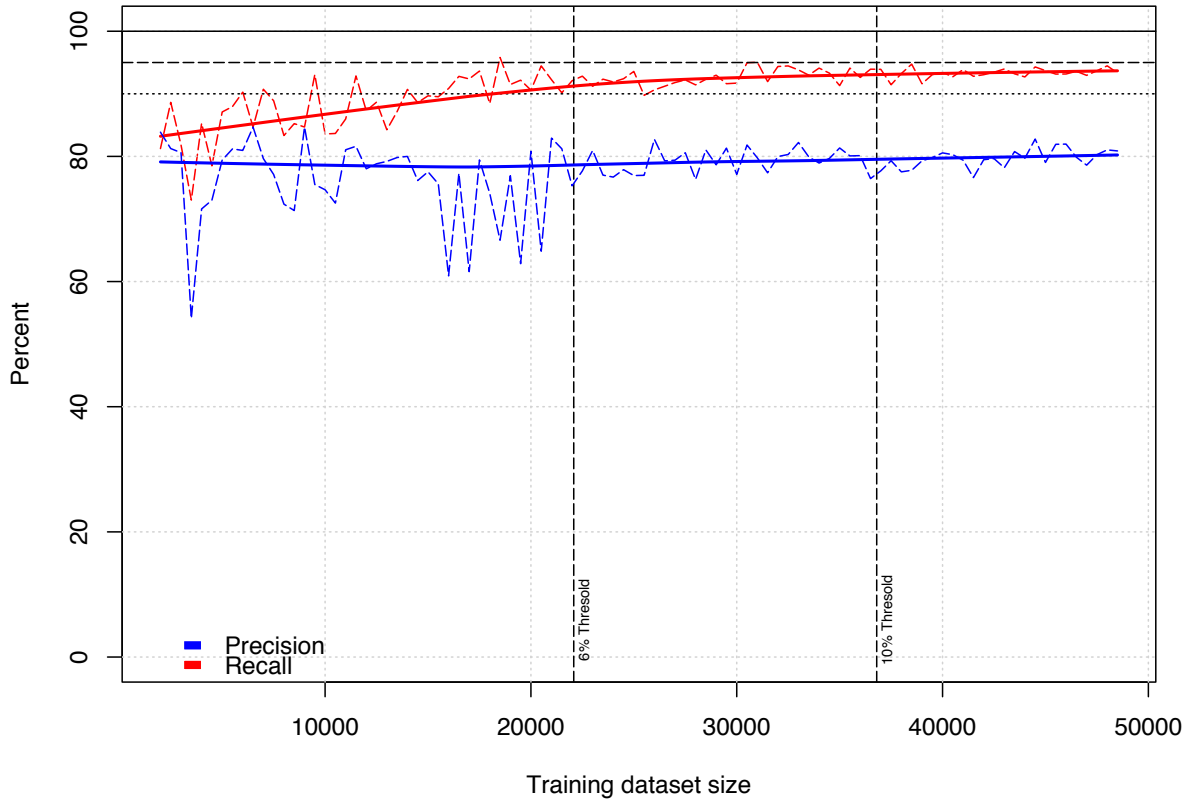
The conglomerate of algorithms features a trade-off between Precision and Recall with most algorithms featuring either high Precision or moderate Recall values. This trade-off is most prominent for GLMNET and FORREST reaching high levels of Precision from even small training datasets approaching almost perfect fit. Yet, these algorithms perform highly sub par in terms of Recall with values approaching around 40%. Similar is true for the BAGGING algorithm which benefits more from an extension of training data size. Yet again, this algorithm did not increase in Recall values over training data size increase and only features a mediocre 60% Recall values even with larger training datasets. The SVM algorithm overall produces Precision values of around .80 and mediocre Recall values limiting at around 60%. The tree-based algorithm did not produce any viable results for low training dataset sizes. The only algorithm that seems to benefit from an increase in training dataset size and reach acceptable results is MAX ENTROPY. Above a training dataset size of 10000 items it features a Precision of around 90% with a Recall of around 80%. Yet, we also find, that beyond this threshold the algorithm does not benefit largely from bigger training datasets. While these results point at a superiority of the MAX ENTROPY approach compared to the manually optimized LFR approach the low Recall values even at higher volumes of training data imply that this approach features an upper limit with these Precision and Recall values. It therefore has to be considered as a supportive algorithm for pre-disambiguation attempts to classify industry organization, while unfortunately its application does not substitute for manual processing.

### **Combining LFR and MAX ENTROPY**

Both the rule-based LFR approach as well as the ML-algorithms used in our testbed feature only moderate levels of Precision and Recall. Taking into account the potential of manual pre- and post-processing we analyzed the correspondence between both approaches, which showed that predictions by both approaches feature a substantial amount of disagreement with Pearson's  $r$  between .70 (1<sup>st</sup> quartile) and .73 (3<sup>rd</sup> quartile). In order to assess, if both approaches can be combined to arrive at better predictions we set up another test-bed which focused on the MAX ENTROPY and the LFR predictions. Similar to the first testbed we randomly selected training datasets from the corpus and processed the corpus via the MAX ENTROPY classifier. The results of the classifier were then combined with the predictions by the LFR approach in a way that a classification to the type industry by either approach was counted as successful classification to that type. The result are shown in Figure 4. Using this approach Precision is reduced compared to the results by the MAX ENTROPY algorithm but level out towards the Precision values found for the LFR approach. Yet, we can observe a positive impact on Recall values. These benefited from an increase in training dataset size albeit at gradually decreasing returns. We also can observe that below the 6% mark of relative training dataset size Precision values fluctuate, which implies that below this threshold the actual sampling of the training data introduced a high level of uncertainty into the algorithm predictions. Above a threshold of 6%

Precision and Recall values stabilize at a gradual increase of Recall. Above a threshold of 10% this gradual increase flattens out at only very marginal level of Recall increase over size.

Figure 4: Precision and Recall values for combined LFR/Max Entropy classifier



To assess if the 6% and 10% threshold provide consistent results regarding high Recall values a further test-bed was constructed. In this case we set the training dataset size to approx. 6% of the corpus (22500) and 10% (36500). For each threshold we ran 100 iterations of the combined LFR/MAX ENTROPY approach. We then analyzed the distribution of Precision and Recall values and calculated quantiles as well as the average of Precision and Recall values (see Table 5).

Table 5: Quantiles and Means for Precision and Recall for 6% and 10% of corpus training dataset

		Min	5%	25%	50%	Mean	75%	95%	Max
Recall	6% Threshold	86.13	88.58	90.34	91.70	91.44	92.52	94.00	94.96
	10% Threshold	90.12	90.65	92.27	92.95	92.92	93.94	94.45	95.73
Precision	6% Threshold	60.22	65.88	75.33	77.97	76.23	79.62	81.45	83.77
	10% Threshold	67.95	69.79	77.76	79.19	78.27	80.29	81.88	82.98

The results show that overall for a 6% threshold of training dataset size Recall values of 91.44 can be expected. Yet, these values have a large range with 90% of all attempts reaching Recall values between 88.58% and 94.00%. For the 10% threshold the situation improves for both Recall and Precision albeit the impact of the larger training dataset size is more apparent for the Recall values. On average a Recall of around 93% can be expected. Increased training dataset size also had a positive impact on the range of potential Recall values with 90% of cases ranging between 90.65% and 94.45%. The minimum found for the 10% threshold suggest that Recall values can be expected to be at least 90%. Yet, readers have to be aware that this is specific to german data and might be different for other countries both positive and negative.

## Conclusions

As results suggest there is no “easy way out” regarding the problem of classifying companies within address data as listed in the Science Citation Index. Disambiguation, while not only solving the crucial problems attached to incomplete and messy data, as well as manual post-processing seem to be the only fruitful suggestions for arriving at publication-based indicators of science-industry linkage at very high levels of quality. First, disambiguation reduces the extent of the problem of classifying organizations into types (see Table 1 & Table 2). The main benefit for subsequent supervised learning or manual approaches is in the reduction of the problem to classify only a fraction of organizations into types.

Accepting the fact that manual effort can currently not be avoided, we propose that until data quality situations drastically improve a multi-method approach can provide some intermediate far from perfect step. These come with a couple grains of salt. First, computational resources for such a combined approach is substantial. Even though, timeliness might not be the most prevalent issue to solve the problem it should be taken into account, that such computations would have to be performed on a per country basis to arrive at comparable indicators on macro level. Similar holds true for achieving reference values for normalization of such indicators on organizational level. Second, the approach still involves substantial effort towards pre- and post-processing of the data. Even the combined approach features a high sweet-spot for training dataset size. To arrive at meaningful results, at least 10% of addresses should be coded per country. Same is true for post-processing. The low Precision values, even though counteracted by high Recall values, imply that manual post-processing has to be taken into account in terms of resources. Also, the number of organizations to code pre-classification at around 10% (36500) is significantly higher than coding organisation post-disambiguation (1702). Third, the solution is far from optimal and should not be considered as a final goal to solve the problem. Rather, the goal of near complete internationally comparable disambiguated data should be kept up. Fourth, there is little chance of assigning rare events, i.e. companies that are rarely involved in publishing and do not include any automatically identifiable term that label them as a company or from which the type industry might be inferred. This holds true for both the LFR approach as well as the ML-approaches utilized in that paper. For that matter, the author is sceptical that any automatized approach be it supervised or unsupervised can solve this problem without including further data such as company registers or data on university startup and spin-off company names. Fifth, due to the limitations of only operating on one country corpus we are unable to infer if the sweet-spot is to be taken as an absolute or relative threshold. A rule of thumb for replication of this approach therefore has to be to code at least around 30000 entries and at least 10% of the corpus. In the comparable example in the works of Morillo, it is suggested, that “[w]ith this set of analysis, it takes around one person/month (24 days) to encode records at the sector level, based upon our own experience.” (Morillo, 2013a). We suggest to at least twice this amount to account for at least two coders and some overhead for addressing intercoder validity. Sixth, we used Germany as a case, which features a large amount of extra-university research organizations, of which some are not *ceteris paribus* distinguishable by Legal Form from companies. Mileage of this approach may vary between countries with better results for countries in which the structure of the science system includes a low share of such organizations.

Further research employing this approach could focus on increasing Recall even further, putting less attention to the impact on Precision to a certain extent. Choice of heterogeneous and not directly fit-for-the-purpose algorithms could be tested and evaluated for their potential to increase Recall to a level that can be considered acceptable as such. The goal here should be to aim at combining the results of algorithms that feature a low correspondence in prediction.

All in all, even though these results suggest an approach that might have some potential, the author suggests that an internationally disambiguated data source will provide both benefit for the approach described here as well as hold further advantages in regard of normalization issues.

## References

- Bador, P., & Lafouge, T. (2005). "Rédaction des adresses sur les publications. Un manque de rigueur défavorable aux universités françaises dans les classements internationaux". *La Presse Médicale*, 34(9), 633–636.
- Blei, D.M., McAuliffe, J. (2007) "Supervised topic models. Advances". *Neural Information Processing Systems 20 (NIPS 2007)*.
- Boser, B. E.; Guyon, I. M.; Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers". *Proceedings of the fifth annual workshop on Computational learning theory – COLT '92*.
- Bourke, P., Butler, L. (1996), "Standards issues in a national bibliometric database: The Australian case", *Scientometrics* 35, 199–207.
- Breiman, L. (1996). "Bagging Predictors". *Machine Learning*, 24(2), 123–140.
- Breiman, L. (2001). "Random Forests". *Machine Learning*, 45(1), 5–32.
- Butler, L. (1999). "Who 'owns' this publication? Problems with assigning research publications on the basis of addresses". In *Proceedings of the seventh conference of the international society for scientometrics and informetrics*.
- De Bruin, R. E., Moed, H. F. (1990), "The unification of addresses in scientific publications". In: *Informetrics 1989/90*, 65–78.
- Friedman, J.H. (1999). "Stochastic Gradient Boosting". Technical Report, Department of Computer Science, Stanford University.
- Friedman, J., Hastie, T. Tibshirani, R.. (2008). "Regularization Paths for Generalized Linear Models via Coordinate Descent" *Journal of Statistical Software*, 33(1), 1-22.
- Galvez, C., Moya-Anegón, F. (2006). "The unification of institutional addresses applying parametrized finite-state graphs (P-FSG)" *Scientometrics*, 69(2), 323-345.
- Galvez, C., Moya-Anegón, F. (2007). "Standardizing formats of corporate source data." *Scientometrics*, 70(1), 3-26.
- García-Zorita, C., Martín-Moreno, C., Lascrain-Sánchez, M. L., & Sanz-Casado, E. (2006). "Institutional addresses in the Web of Science: the effects on scientific evaluation". *Journal of Information Science*, 32(4), 378–383.
- Huang, S., Yang, B., Yan, S., & Rousseau, R. (2014). "Institution name disambiguation for research assessment". *Scientometrics*, 99(3), 823–838.
- Nigam, K., Lafferty, J., McCallum, A. (1999). "Using maximum entropy for text classification." *IJCAI-99 Workshop on Machine Learning for Information Filtering*.
- Morillo, F., Aparicio, J., González-Albo, B. und Moreno, L. (2013a) Towards the Automation of Address Identification. *Scientometrics*, 94 (1), 207–24.
- Morillo, F., Santabarbara, I., Aparicio, J. (2013b). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95(3), 953-966.

Niu, L., Wu, J., & Shi, Y. (2012) "Entity disambiguation with textual and connection information". *Procedia Computer Science*, 9, 1249–1255.

Perianes-Rodriguez, A., Chinchilla-Rodriguez, Z., Vargas-Quesada, B., Olmeda-Gomez, C., & Moya-Aregon, F. (2009). "Synthetic hybrid indicators based on scientific collaboration to quantify and evaluate individual research results". *Journal of Informetrics*, 3(2), 91–101.

Van Raan, A. F. J. (2005). "Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods". *Scientometrics*, 62(1), 133–143.