Informatik Spektrum (2020) 43:137–144
https://doi.org/10.1007/s00287-020-01259-2

HAUPTBEITRAG

Applying machine intelligence in practice

Selected results of the 2019 Dagstuhl Workshop on Applied Machine Intelligence

Bernhard Humm[1] · Hermann Bense[2] · Jürgen Bock[3] · Mario Classen[4] · Oren Halvani[5] · Christian Herta[6] · Thomas Hoppe[7,8] · Oliver Juwig[9] · Melanie Siegel[1]

Published online: 6 March 2020

Abstract
The relevance of Machine Intelligence, a.k.a. Artificial Intelligence (AI), is undisputed at the present time. This is not only due to AI successes in research but, more prominently, its use in day-to-day practice. In 2014, we started a series of annual workshops at the Leibniz Zentrum für Informatik, Schloss Dagstuhl, Germany, initially focussing on Corporate Semantic Web and later widening the scope to Applied Machine Intelligence. This article presents a number of AI applications from various application domains, including medicine, industrial manufacturing and the insurance sector. Best practices, current trends, possibilities and limitations of new AI approaches for developing AI applications are also presented. Focus is put on the areas of natural language processing, ontologies and machine learning. The article concludes with a summary and outlook.

Anwendungen maschineller Intelligenz in der Praxis

Ausgewählte Ergebnisse des Dagstuhl-Workshops zu angewandter maschineller Intelligenz 2019

✉ Bernhard Humm
bernhard.humm@h-da.de

Hermann Bense
hb@bense.com

Jürgen Bock
Juergen.Bock@kuka.com

Mario Classen
mario.classen@axa.de

Thomas Hoppe
thomas.hoppe@fokus.fraunhofer.de,
thomas.hoppe@htw-berlin.de

Oliver Juwig
oliver.juwig@axa.de

Melanie Siegel
melanie.siegel@h-da.de

[1] Hochschule Darmstadt—University of Applied Sciences, Haardtring 100, 64295 Darmstadt, Germany

[2] bense.com GmbH, Schwarze-Brüder-Str. 1, 44137 Dortmund, Germany

[3] KUKA Deutschland GmbH, Augsburg, Germany

[4] AXA Konzern AG, Colonia Allee 10–20, 51067 Cologne, Germany

[5] Fraunhofer Institut für Sichere Informationstechnologie SIT, Darmstadt, Germany

[6] Hochschule für Technik und Wirtschaft Berlin, Wilhelminenhofstr. 75a, 12459 Berlin, Germany

[7] Fraunhofer FOKUS, Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany

[8] Hochschule für Technik und Wirtschaft Berlin, Wilhelminenhofstr. 75a, 12459 Berlin, Germany

[9] AXA Konzern AG, Colonia Allee 10–20, 51067 Cologne, Germany

Introduction

The relevance of *Machine Intelligence, a.k.a. Artificial Intelligence (AI)*[1], is undisputed at the present time. This is not only due to AI successes in research but, more prominently, its use in day-to-day practice.

In 2014, we started a series of annual workshops at the Leibniz Zentrum für Informatik, Schloss Dagstuhl, Germany, initially focussing on Corporate Semantic Web and later widening the scope to Applied Machine Intelligence. In all workshops, we focussed on the application of AI technologies in corporate and organizational contexts. A number of books and journal articles resulted from those workshops [1–6].

This article presents selected results from the 2019 workshop. It is structured as follows: firstly, we present AI applications in selected domains, including industrial production, insurance and medicine. We then focus on practical aspects of three main AI areas: natural language processing, knowledge engineering and *machine learning* (ML). For those areas, we present best practices for developing AI applications. We conclude the article with a summary and outlook.

AI applications in selected domains

This section briefly presents the need for AI applications in the domains of medicine, industrial production and the insurance sector, as well as two cross-sector AI applications. Information on developing such applications will be given in the following sections.

Image-based medical diagnosis

In recent years, many successful *medical applications* have been developed, particularly for image analysis based on computer vision using deep neural networks. It is expected that the use of deep neural networks will revolutionize image-based medical diagnosis in the coming years, e.g. in pathology and radiology. In digital pathology, for example, there is a strong need for such systems due to a lack of experts and an increased volume of digitized images.

In view of their critical nature, *Computer-Aided Diagnosis* tools have special requirements. Predictions made by AI systems should be *explainable* [7], i.e. there should be clues for a human as to why the system makes certain predictions. Another important requirement [8] is to get an *estimate of the reliability* of decisions, i.e. such systems should estimate their limits on their own. If, for instance, the characteristics of an image or parts of an image are particularly different from images in the training dataset, the system should give a high uncertainty score, which can be interpreted as "I don't know". Consequently, a medical expert should get feedback that the decision is unreliable and needs further inspection. In summary, both *explainability* and *uncertainty estimation* increase trust in medical diagnosis systems, which is important for the acceptance of such systems in medical practice.

Chatbots in the insurance sector

In times of online stores like Amazon, customers are used to near real-time processing of their requests. With these rapidly changing customer expectations, combined with strong competition on the insurance market, *Chat & Voice-Bots* are a way for insurance companies to offer customers a simple and understandable interaction channel.

With a service of this kind, customers can get instant information about products, their insurance coverage and so on. See Fig. 1 for an example.

Besides serving customers with relevant and easily accessible information without having to wait in service lines, the work for the agents is reduced by automating simple business transactions. Thus, they can focus on those cus-

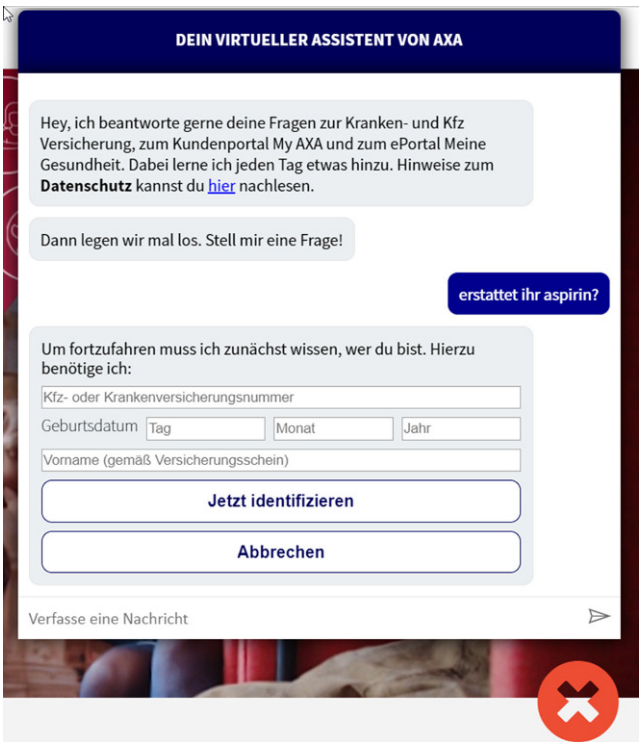

Fig. 1 Sample chatbot dialog by the AXA digital assistant (in German)

[1] 'We prefer the term Machine Intelligence to Artificial Intelligence (AI) in order to avoid interpretations of AI being an alternative form of intelligence equivalent to human intelligence. However, we will use both terms interchangeably.

tomer interactions in which knowledge and empathy for a human play an important role.

Of course, a bot is never as smart as a human with many years of experience. Therefore, it is a key success factor to be able to handoff the conversation to a human agent when needed. In the case of a misunderstanding, the customer does not need to change the communication channel, but get their information on the same channel. In addition, human conversation can be used to train the bot to make it even smarter.

Intelligent control in industrial production

On the way to *intelligent automation in industrial production systems*, it is essential that components of such a system are able to effectively communicate with each other. In this context, components are not only production resources such as devices, but also products being produced as well as processes that guide the production. In so-called *changeable production*, the ensemble of components must be enabled to react on changing requirements from any direction, e.g. a changed product specification in small-lot-size production, the replacement of a device, or the adaptation of the process in order to improve energy consumption or cycle time efficiency. Common to all of these scenarios is the requirement that all components can share information and "understand" each other. This motivates the use of semantic technologies when it comes to information modelling, and intelligent infrastructure elements that ensure storage, processing and communication of the semantic models.

Authorship verification

In today's rapidly changing world, huge amounts of data are generated every second, much of it in natural language. Text data often does not contain metadata that can be used to reveal its origin, i.e. authorship. *Authorship verification (AV)*, a research subject in the field of *digital text forensics*, suits this purpose. AV concerns itself with the question of whether two documents have been written by the same person.

Beyond forensics, AV can be used for a variety of applications in many areas. In business related domains, for example, AV can be used for the purpose of multiple account detection on social platforms [9]. In information retrieval AV can be used to enhance search systems [10], making it possible to aggregate retrieved documents not only by their content but also by their underlying writing style. In education and research, AV can be applied to detect ghostwriting [11] or (intrinsic) plagiarism [12], whereas in digital humanities, AV can be used to authenticate historical writings [13]. Even in the healthcare domain AV can be used, e.g. to observe speech changes in individuals suffering from Alzheimer's disease [14].

Intelligent notification

Today, most news is spread with a broadcasting approach, using email newsletters and messengers like WhatsApp, FB Messenger, Instagram and Telegram. Usually, users cannot tailor the provided information to their specific interests; thus there is a demand for *intelligent notification services (INS)*, such as https://robb.ee.

This INS, for instance, provides a chatbot as an AI-based user front end for configuring the information need. Users can enter queries and user profiles using front ends like Apple Siri or Google Cloud Speech-to-Text. Using linked data from public and private sources, the query results may contain information on weather, stock exchange data, politics, culture, sporting events and more. The output language can be chosen. The INS system takes rules defined by the user to send messages only if the condition of the rules are satisfied. Thus the user is only informed about news which is relevant to them. Rules are of the kind "Send info if the stock of Lufthansa falls/increases more than 5% during the trading day", or "Send info if a Rolling Stones concert will take place at a maximum distance of 100 km of where I live", or "Send info if the authors Humm AND Schade have a new publication containing the topics AI and Ontology".

Natural language processing

Natural language processing (NLP) has found its way into our everyday lives, thanks to voice-controlled assistant systems, machine translation and other applications. The predominant technology in this research area is currently ML, especially deep learning. But there are also other promising technologies, as can be seen in the section on compression-based AV. NLP applications can be based on semantic dictionaries as described in the section on OdeNet. These are developed and extended with automatic procedures that include linguistic information.

Extending and applying OdeNet

WordNets are well-established lexical resources with a wide range of applications. They have been elaborately set up and maintained manually for more than 20 years. The most prominent example is the original Princeton WordNet of English (PWN) [15]. In recent years, there have been increasing activities for automatically extracting open WordNets for different languages from other resources and enriching these with lexical semantics information, thus building the so-called *Open Multilingual WordNet* [16]. These

WordNets were linked to PWN via shared synset identities (IDs) [17, 18]. In this context, a German lexical semantics resource with the name *Open German WordNet (OdeNet)* is being developed with the aim to be included as the first open German WordNet in the Open Multilingual WordNet. OdeNet is automatically created from different information sources. The resource is based on *Open Thesaurus*, a thesaurus consisting of entries created with crowd sourcing. It contains around 200,000 lexical entries in approximately 35,000 synonym groups. Syntactic categories were added by automatic part-of-speech (POS) tagging, links to the shared synset IDs by machine translation. With an analysis of German compounds, hyponym relations were automatically added. Further relations were taken from the English PWN, as well as definitions. A number of access tools for OdeNet have been implemented to extend the entries semi-automatically.

We have annotated basic German words and used OdeNet to mark complex synonyms of the basic German words in texts. This is done in the context of projects on simplified language. Another application is sentiment analysis: synonyms and antonyms for evaluating expressions can be used to expand sentiment dictionaries. Expressions for evaluated aspects can also be found in OdeNet.

COAV: a compression-based author-verification approach

Compression-based similarity detection can be seen as an alternative approach to traditional text classification tasks and has been widely used across many research fields. One of the greatest advantages of compression-based similarity detection is that the entire feature engineering process is performed internally by the underlying compression algorithm, such that a manual definition by a human domain expert can be avoided.

We adapted compression models to the field of AV and proposed a binary-intrinsic AV method named *COAV* [19], which yields competitive results compared to a number of state-of-the-art AV approaches, based on recurrent neural networks, support vector machines or random forests. In contrast to these, however, COAV does not make use of ML algorithms, NLP techniques, feature engineering, hyperparameter optimization or external documents (a common strategy to transform AV from a one-class to a multi-class classification problem). Instead, the only three key components our method relies on are a compressing algorithm, a dissimilarity measure and a threshold, where the latter is needed to accept or reject the authorship of the questioned document. Due to its compactness, COAV performs extremely fast and can be reimplemented with minimal effort. In addition, it can handle challenging AV cases where both the questioned and the reference document differ in terms of topic, genre or a long period of time over which they were written.

We, as well as other researchers such as [20], evaluated our approach against publicly available benchmark datasets, which were used in three international AV competitions. Furthermore, we constructed additional corpora and evaluated our method against state-of-the-art AV approaches. In all cases, COAV achieved promising results, as can be seen in detail in [19] and [20].

Knowledge engineering of ontologies

Engineering ontologies can be a time-consuming and costly task. This section presents two approaches for constructing or enriching ontologies semi-automatically, highlighting possibilities and restrictions.

Enriching ontologies with Wikidata

In 2012, the Wikimedia Foundation started the *Wikidata* project with the intention of making factual information consistent in all different language versions of Wikipedia. Today, Wikidata has evolved into a huge, broad and multilingual fact base, which not only provides factual information about common entities in different domains, but which also contains translations of these facts. By integrating other kinds of relations, such as subClassOf, instanceOf, synonyms and other entity-specific relations, it has transformed into a *knowledge graph*. Via a SPARQL-based query interface and an application programming interface (API), this information can be retrieved comparatively easily. So the idea comes to mind of using this information to suggest augmentations during the knowledge engineering process.

Initial experiments with the retrieval of synonyms from Wikidata in the context of an industry project and the Qurator research project[2] have shown that a mean of three usable synonyms per case could be retrieved in up to 36% of cases. However, in a quarter of these cases, wrong synonyms were contained in Wikidata.

With regard to the retrieval of superclass information, the situation is even worse. Terminological *subclassOf* relations are frequently mixed up with assertional *instanceOf* relations, thus blurring the epistemological distinction between concepts and instance as well as between sets and elements. Additionally, terminological cycles where found, which make the uncontrolled extraction of subclass information from Wikidata impossible.

These findings indicate that the Wikidata community is not aware of basic principles of knowledge modelling. Although the information maintained by Wikidata is useful

[2] https://qurator.ai.

for its original goal, from the perspective of augmenting knowledge engineering processes, it should be used with care in human-guided curation processes.

Extracting ontological knowledge from semi-structured texts

In *Industry 4.0*, significant effort is put into developing new norms and standards for the description of components and information models, etc. There is already a large body of well-accepted existing industrial standards in the form of text documents. In order to achieve an increased level of autonomy in the interaction and interoperation of Industry 4.0 components, information models must be machine-interpretable instead of purely based on natural language norming documents. There is a chance that knowledge which is represented in these kinds of documents for standards and norms can be translated into a machine readable form, such as ontologies. This is possible since standards and norms are written with the purpose of being unambiguous, concise and explicit. This is a major difference to natural language documents from social media and other informal information sources, where a lot of research effort is currently being invested in analysing and interpreting those.

Classical and novel NLP approaches appear to be able to tackle the extraction of formal axioms and statements from industrial standards and norms, such as specifications, guidelines and even patents. Business rules or grammar-based systems, as well as classical Hearst patterns may be promising first attempts to achieve noticeable results. Moreover, formal technical glossaries can be used in order to maintain recurring terms and their definitions.

What is difficult is the use of background knowledge and contextual knowledge. Also, ML techniques might be difficult to apply in a domain where there is only a small number of documents available as training material, as is the case with norms and standards.

Practical aspects of machine learning

Since the advent of the digital age in around 2000, the rapidly increasing volume of digital data has intensified the research and development of ML and led to some breakthroughs, especially in the area of neural networks. The increased number of approaches available makes it difficult to maintain an overview of the field and to decide which approaches best fit a particular purpose. Under the label *democratization of AI*, some efforts have been made to open up ML to the masses. This requires, amongst other things, that the characteristics and properties of learning approaches can be described properly, that criteria for the selection of appropriate evaluation measures exist and that recurrent questions like "How many training examples do I need?" can be answered from an application-oriented viewpoint.

An ontology of machine learning

A German proverb says "The best cobblers have the worst shoes". Figuratively, the same holds true for us computer scientists. We probably all know or have heard of the Association for Computing Machinery (ACM) Computing Classification System. Its current version dates back to 2012. In terms of the development speed of computer science this is comparatively old. Although its granularity is coarse and its depth is somewhat limited, it covers different aspects of computer science useful for a gross classification of literature, but not sufficient for the description of particular systems. Unfortunately, as intellectual property of the ACM, it does not appear to be available in machine-processable form as an OWL or RDF file—only as an HTML-file.

What holds true for computer science in general holds true for ML in particular: there is no *ontology of ML* yet. Such an ontology could be useful for a variety of purposes:

- **Systematization of ML:** To classify ML approaches and technologies and to explicate their relationships
- **Teaching support:** To gain a better overview and understanding of the field by describing the preconditions, limitations and features of ML approaches; for the development of learning paths for teaching ML
- **Machine-processable description of ML components**: To support automated orchestration, thereby supporting ML engineers

How could such an ontology of ML be built as an explicit, shared formal model of a conceptualization [21]? The answer is: by the application of good knowledge engineering practice, i.e. by specifying use cases, by developing competence questions [22] for these use cases, by deriving important conceptual categories from these competence questions and by research, extraction and transformation from existing knowledge sources. The integration and augmentation of these resources needs to be done by a group of experienced ML practitioners. We are currently forming such a group. If you are interested in participating, please contact Thomas Hoppe.

Performance measure awareness

Before ML-based applications can be used in practice, a comprehensive *prediction performance assessment* must be carried out in advance. However, a question that arises is: *which performance measure is suitable for a specific ML task*?

In general, we can distinguish between two evaluation approaches: (1) *Threshold-dependent* and (2) *threshold-independent* performance measures. With regard to (1), *single-number metrics* represent the most common choice. They can be derived from a standard *confusion matrix* (in the case of classification) or from *error metrics* (in the case of regression). Common classification measures are, for instance, *Accuracy, F1 (including precision and recall), Kappa, Matthews Correlation Coefficient (MCC), Equal Error Rate, Youden's Index, Likelihood Ratio* and *(Adjusted) Geometric Mean*, while for regression, common metrics include *Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)* and *R-squared*. Regarding (2), one of the most common measures is the *AUC*, which represents the area under the *ROC curve*.

Regardless of which measure is considered, it should be ensured that it is appropriate for the specific application, given the fact that each performance measure has its own limitations. When focusing on the performance of a classifier that is going to be used in production, we should consider measures linked to fixed thresholds (1). If, on the other hand, we are only interested in the theoretical performance of a classification model, we can consider (2) instead. What also affects the choice of the measure is whether it is suitable for the evaluation of imbalanced datasets, which are frequently found in practical scenarios such as sentiment detection. Here, measures such as Accuracy or MCC are unreliable, while other metrics such as Kappa, Geometric Mean or Youden's Index make more sense. Note that F1 can also be used to assess the performance of an ML application on imbalanced datasets. However, since F1 is affected by the changes in the class distribution [23] and also suffers from other deficiencies as described in [24], a thorough consideration is recommended before using this metric.

Finally it should be emphasized that insufficient knowledge of the respective performance indicators can have devastating consequences, especially in critical applications.

How many trainings samples do I need?

In application domains where data acquisition is expensive, such as medical research or field explorations, there is a need for reliable estimations of the sample size for training ML models. Methods like power analysis, widely used in statistics and medical research, are not applicable for ML, since important information like effect size and statistical power are usually not known. While in the ML community the analysis of the prediction performance of a model is an established field (e.g. using accuracy, F1 etc.; see previous section), estimating the sample size needed ex ante is rarely done and little information can be found in the community.

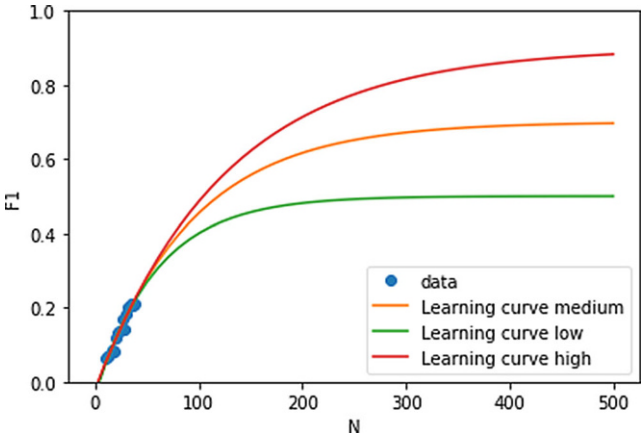

Fig. 2 Learning curve analysis

Consider the following use case: ML shall be used to predict an impending drop-out of a patient in psychotherapy. Input data are questionnaires and free texts. The open question is: how many patients and questionnaires are needed in order to achieve an F1 score of 0.8 or better?

One option is to perform a *learning curve analysis,* which is becoming more and more popular amongst ML practitioners (Fig. 2).

In the example, data of $N = 48$ patients is available where nine patients are in the class drop-out. For a learning curve analysis, you train a model with fewer samples than available over and over again, e.g. for $N = 10$, 20, 30, 40, 48. Each time, the prediction performance (here: F1) is measured. The sample size and prediction performance are entered in a coordinate system. Using regression, the prediction performance for higher training sample rates is predicted (Fig. 2).

This approach has two disadvantages. Firstly, you do need at least *some* training data to start with (here: 48). Secondly, with only a few training samples, the prediction is most vague. In Fig. 2, this is indicated by three different curves with estimated F1 scores ranging from 0.4 to 0.9 for 500 training samples.

In the field of *Computational Learning Theory*, different lower and upper bounds on the number of needed training examples were derived, which assure with high confidence that the generalization error lies below a given threshold. This is known as *PAC learning* (*Probably Approximate Correct learning*). The upper bounds are worst case estimates. They are not applicable in practice, since they overestimate the sample size for risk minimization. Although under the label *PAC-Bayes*, bounds are derived that account for the mean case, they do not appear to be applicable in practice either, since they are based on prior assumptions about the sample distribution, which is usually unknown in advance like the sample size.

Abu Mostafa[3] explains in the context of PAC learning that the number of samples needed only needs to increase linearly with the number of Vapnik–Chervonenkis (VC) dimensions to maintain a certain quality level. However, this is not really helpful, since the VC dimension is only known for a comparatively small number of ML approaches.

Finally, various *rules of thumb* derived from experience are used in the ML community to estimate the sample size, e.g.:

- Ca. 10 samples for each feature
- Ca. 150 samples for each class

An advantage of rules of thumb is that an estimate can be given without having a single training sample. However, at the beginning of an ML project, the number of relevant features is often unknown or not yet clear, particularly when dealing with texts or if feature engineering still needs to be performed. These estimates are also very rough and not specific to the concrete problem.

To conclude, none of these approaches are really satisfying. However, combining various approaches mentioned above may increase confidence in a sample size prediction. Whatever the case may be, it is advisable to continuously perform a learning curve analysis while collecting more and more training samples.

Conclusions

AI applications are in everyday corporate use. This article presents a number of AI applications from various application domains, including medicine, industrial manufacturing and the insurance sector. We have presented best practices, current trends, possibilities and limitations of new AI approaches for developing AI applications. We focused on the areas of NLP, ontologies and ML.

The selection of approaches presented is by no means comprehensive. It reflects a subset of topics that were discussed during the 2019 Dagstuhl workshop on Applied Machine Intelligence. We have written an article, "5 Years of Semantics Workshops in Schloss Dagstuhl: it connects!" [25] (article published in same issue of *Informatik Spektrum* and also available in German), which gives an impression of the spirit of those workshops. We will continue to share our experiences in Applied Machine Intelligence in Dagstuhl workshops and to publish our results. If you work on intelligent applications in corporate contexts, you are cordially invited to participate in next year's workshop (contact: Bernhard Humm, Thomas Hoppe).

[3] https://work.caltech.edu/telecourse.html Lecture 7: The VC Dimension.

Funding Open Access funding provided by Projekt DEAL.

References

1. Bense H, Gernhardt B, Haase P, Hoppe T, Hemmje M, Humm B, Paschke A, Schade U, Schäfermeier R, Schmidt M, Siegel M, Vogel T, Wenning R (2016) Emerging trends in corporate semantic web—selected results of the 2016 Dagstuhl workshop on corporate semantic web. Informatik Spektrum 39(6):474–480
2. Busse J, Humm B, Lübbert C, Moelter F, Reibold A, Rewald M, Schlüter V, Seiler B, Tegtmeier E, Zeh T (2015) Actually, what does "Ontology" mean? A term coined by philosophy in the light of different scientific disciplines. J Comput Inform Technol 23(1):29–41. https://doi.org/10.2498/cit.1002508
3. Ege B, Humm B, Reibold A (eds) (2015) Corporate Semantic Web – Wie Anwendungen in Unternehmen Nutzen stiften. Springer, Heidelberg (in German)
4. Hoppe T, Humm B, Schade U, Heuss T, Hemmje M, Vogel T, Gernhardt B (2015) Corporate semantic web—applications, technology, methodology. Informatik Spektrum 39(1):57–63. https://doi.org/10.1007/s00287-015-0939-0
5. Hoppe T, Humm B, Reibold A (eds) (2018) Semantic applications—methodology, technology, corporate use. Springer, Berlin
6. Humm BG, Bense H, Classen M, Geißler S, Hoppe T, Juwig O, Paschke A, Schäfermeier R, Siegel M, Weichhardt F, Wenning R (2019) Current trends in applied machine intelligence. Informatik Spektrum 42(1):28–37. https://doi.org/10.1007/s00287-018-01127-0
7. Samek W, Wiegand T, Müller K (2017) Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. CoRR, abs/1708.08296
8. Leibig C et al (2017) Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep 7(1):17816
9. Hosseinia M, Mukherjee A (2017) Detecting sockpuppets in deceptive opinion spam. CoRR abs/1703.03149
10. Rexha A, Kroll M, Ziak H, Kern R (2017) Extending scientific literature search by including the author's writing style. In: Mayr P, Frommholz I, Cabanac G (eds) Proceedings of the fifth workshop on bibliometric-enhanced information retrieval (BIR) co-located with the 39th European Conference on Information Retrieval ECIR 2017, Aberdeen, 09.04.. CEUR Workshop Proceedings, vol 1823, pp 93–100 (CEUR-WS.org)
11. Stavngaard M, Sorensen A, Lorenzen S, Hjuler N, Alstrup S (2019) Detecting ghostwriters in high schools. CoRR abs/1906.01635
12. Stein B, Lipka N, Prettenhofer P (2011) Intrinsic plagiarism analysis. Lang Resour Eval 45(1):63–82
13. Kestemont M, Stover JA, Koppel M, Karsdorp F, Daelemans W (2016) Authenticating the writings of Julius Caesar. Expert Syst Appl 63:86–96
14. Hirst G, Feng WV (2012) Changes in style in authors with alzheimer's disease. Engl Stud 93:357–370

15. Fellbaum C (ed) (1998) Wordnet: an electronic lexical database. MIT Press, Cambridge
16. Bond F, Paik K (2012) A survey of WordNets and their licenses. In: Proceedings of the global wordnet conference
17. Bond F, Foster R (2013) Linking and extending an open multilingual WordNet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics Sofia, pp 1352–1362
18. Bond F, Vossen P, McCrae JP, Fellbaum C (2016) Cili: the collaborative interlingual index. In: Proceedings of the Global WordNet Conference, vol 2016
19. Halvani O, Winter C, Graner L (2017) On the usefulness of compression models for authorship verification. In: Proceedings of the 12th international conference on availability, reliability and security ARES '17. ACM, New York, pp 54:1–54:10
20. Bevendorff J, Stein B, Hagen M, Potthast M (2019) Generalizing unmasking for short texts. In: NAACL-HLT, vol 1, pp 654–659
21. Studer R, Benjamins R, Fensel D (1998) Knowledge engineering: Principles and methods. Data Knowl Eng 25(1–2):161–198
22. Grüninger M, Fox MS (1995) Methodology for the design and evaluation of ontologies. In: Proceedings of the workshop on basic ontological issues in knowledge sharing IJCAI-95, Montreal (https://www.researchgate.net/publication/2288533_Methodology_for_the_Design_and_Evaluation_of_Ontologies)
23. Tharwat A (2018) Classification assessment methods. Appl Comput Inform. https://doi.org/10.1016/j.aci.2018.08.003
24. Powers DMW (2015) What the F-measure doesn't measure: features, flaws, fallacies and fixes. CoRR. http://arxiv.org/abs/1503.06410. Accessed 2019/11/01
25. Schade U, Fillies C, Humm B, Reibold A, Schumann F, Weichhardt F 5 Years of Semantics Workshops in Schloss Dagstuhl: it connects! https://www.researchgate.net/publication/336798165_5_Years_of_Semantics_Workshops_in_Schloss_Dagstuhl_it_connects_4_Day_Workshop_-30_People_-1_Common_Cause (respectively in German: https://www.researchgate.net/publication/335621540_5_Jahre_Semantik-Workshops_im_Schloss_Dagstuhl_das_verbindet). Accessed 2019/11/01