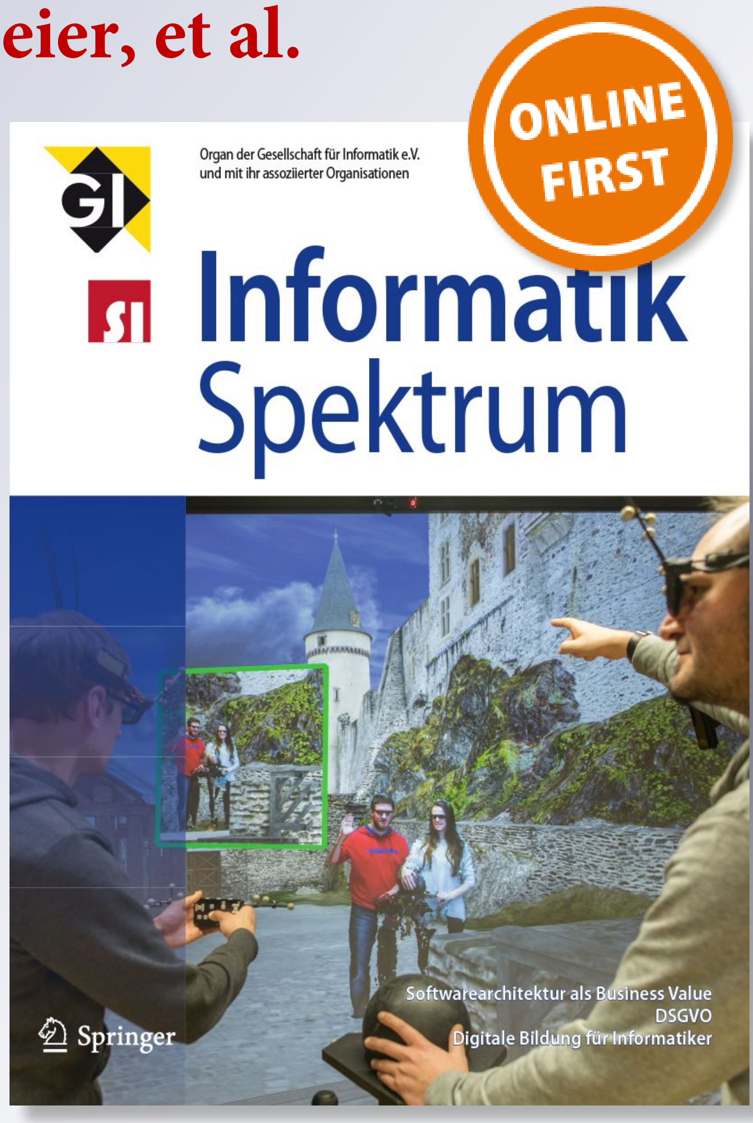
Current trends in applied machine intelligence

Bernhard G. Humm, Hermann Bense, Mario Classen, Stefan Geißler, Thomas Hoppe, Oliver Juwig, Adrian Paschke, Ralph Schäfermeier, et al.

Informatik-Spektrum
Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen

ISSN 0170-6012

Informatik Spektrum
DOI 10.1007/s00287-018-01127-0


ONLINE FIRST
GI
SI
Organ der Gesellschaft für Informatik e.V.
und mit ihr assoziierter Organisationen
Informatik
Spektrum
Springer
Softwarearchitektur als Business Value
DSGVO
Digitale Bildung für Informatiker

Current trends in applied machine intelligence

Bernhard G. Humm · Hermann Bense
Mario Classen · Stefan Geißler
Thomas Hoppe · Oliver Juwig
Adrian Paschke · Ralph Schäfermeier
Melanie Siegel · Frauke Weichhardt
Rigo Wenning

Introduction

Nearly a decade ago, the term “artificial intelligence” (AI) appeared to be a taboo word even though work on the semantic web was still being carried out. However, the recent successes of machine learning (ML), especially of artificial neural networks (ANN), have revived the interest in AI. Eventually, this has resulted in a new hype about AI, ML, and ANN, equating their meanings in popular and social media. This hype led to the suggestion to interpret AI as “alternative intelligence”, carrying the idea that in addition the blueprint of human intelligence, there are other forms of intelligence. Although it is not new, the term “machine intelligence”[1] is from our understanding better suited, since it emphasizes that we build machines with some kind of intelligence, whether that will be based on ANN, ML or GOFAI (“good old-fashioned, symbolic AI”).

In 2014 we started a series of annual workshops at the Leibniz Zentrum für Informatik, Schloss Dagstuhl, focusing on semantic applications and semantic technologies used in corporate context. A number of books and journal articles resulted from those workshops [3, 7, 9, 12, 13]. We found it worthwhile to adjust the scope of the workshops in order to connect semantic approaches to current approaches in the fields of data science and ML.

This article summarizes the main results of our 2018 workshop. It starts with intelligence applications in practice, followed by current trends in machine intelligence: natural language processing, combining symbolic and non-symbolic approaches, data quality, and processes and ontologies.

Intelligent applications in practice

Corporate applications pose different requirements to research prototypes, meeting business needs with high performance, usability, and maintainability. Our work and, hence, this article, are focused on corporate applications. In this section, we present exemplary requirements from two business sectors: insurance and medicine.

[1] Dating back at least to Donald Michie in the 1960s, England.

https://doi.org/10.1007/s00287-018-01127-0

Bernhard G. Humm · Melanie Siegel
Hochschule Darmstadt – University of Applied Sciences,
Haardtring 100, 64295 Darmstadt, Germany
E-Mail: {bernhard.humm, melanie.siegel}@h-da.de

Hermann Bense
bense.com GmbH,
Schwarze-Brüder-Str. 1, 44137 Dortmund, Germany
E-Mail: hb@bense.com

Mario Classen · Oliver Juwig
AXA Konzern AG
Colonia Allee 10–20, 51067 Köln, Germany
E-Mail: {mario.classen, oliver.juwig}@axa.de

Stefan Geißler
Expert System Deutschland GmbH,
Blumenstr 15, 69115 Heidelberg, Germany
E-Mail: skf.geissler@googlemail.com

Thomas Hoppe · Adrian Paschke
Fraunhofer FOKUS,
Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany
E-Mail: {thomas.hoppe, adrian.paschke}@fokus.fraunhofer.de

Ralph Schäfermeier
Freie Universität Berlin,
Berlin, Germany
E-Mail schaef@inf.fu-berlin.de

Frauke Weichhardt
Semtation GmbH,
Geschwister-Scholl-Straße 38, 14482 Potsdam, Germany
E-Mail: fweichhardt@semtalk.com

Rigo Wenning
World Wide Web Consortium,
2004 Route des Lucioles, 06902 Sophia Antipolis, France
E-Mail: rigo@w3.org

Published online: 29 November 2018

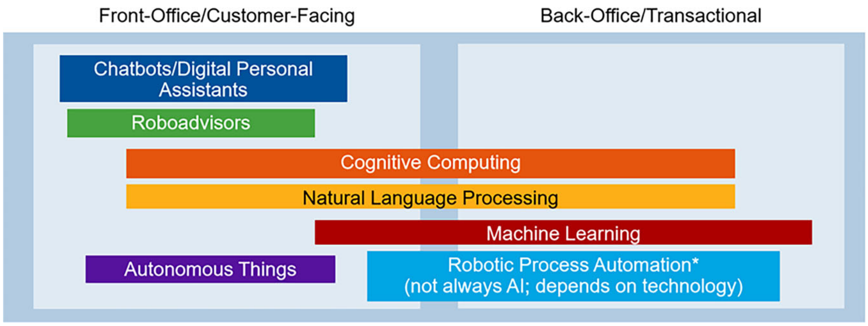

Front-Office/Customer-Facing
Back-Office/Transactional
Chatbots/Digital Personal Assistants
Roboadvisors
Cognitive Computing
Natural Language Processing
Machine Learning
Autonomous Things
Robotic Process Automation*
(not always AI; depends on technology)

***Fig. 1** AI applications in the insurance sector [11].*

Insurance

In the cost-sensitive insurance sector, shadow processing (German: "*Dunkelverarbeitung*") has traditionally been an important key performance indicator (KPI). Especially in personnel-intensive service areas, the quota of fully automated processing without human interaction or decisions is one of the most important KPIs for cost optimization. Today, back-office processing is dominated by batch processing and automated decision making, which for the most part are based on traditional algorithmic approaches.

It is not surprising that the promises of AI fall on fertile ground in the insurance sector. According to Gartner, there are many application areas for AI and ML in the value chain of insurance companies (see Fig. 1). In addition to fully automated processing in the back office, intelligent decision making and personalized behavior at customer touchpoints and in digital sales processes have seen an increasing importance in recent years.

ChatBots, especially with natural language processing (NLP), can still be regarded as a field of experimentation in the insurance sector. They are currently limited to automated FAQs. Other AI-based solutions, for example pre-classification of incoming mail, have already been successfully implemented and show promising results. Still, the introduction of AI-based technologies in the insurance sector is a time-consuming and costly process. The fact that most innovative AI frameworks and products are cloud based often conflicts with data protection needs for personnel data and other regulations.

Medicine

Medical treatment more or less adheres to the same chain of activities starting with the acquisition of the patient's medical record, followed by the collection of clinical findings (identification of symptoms), the generation of a diagnosis (identification of the presence of a syndrome), the derivation of an indication for an appropriate treatment, risk assessment, the execution of the actual treatment, success evaluation, and prognosis. Each of these process steps has a high potential for support by AI approaches.

Medical domain vocabulary is often ambiguous. For example, the term "cavity" might represent a hole in a tooth that has either been caused by caries or drilled by a dentist and would therewith be the outcome of a finding or a treatment activity instance, respectively. There are many relations that are hard to capture without background knowledge, e. g., the symmetry of the human jaw.

In the medical domain, there exist numerous controlled vocabularies, thesauri, and ontologies. They contain medical terms and, potentially, additional information such as explanations, synonyms, hypernyms (broader terms), and domain-specific term relationships. Whereas some medical ontologies are commercial (e. g., Unified Medical Language System® Metathesaurus®, SNOMED-CT, etc.), many open source ontologies are available. For an overview, see, e. g., www.ontobee.org.

Personalized medicine is the practice of tailoring patient care based on the predicted response or risk of disease [1, 14]. For example, patients with the same broad diagnosis, such as breast cancer, can have varying forms of the disease, such as 'papillary carcinoma of the breast' or 'ductal carcinoma in situ', and will have major differences in their treatment plans. When the myriad of potential comorbidities, medications, symptoms, environmental factors, and demographics of patients are considered, there are even more differences in the optimal care paths for patients. Furthermore, since a genetic component is present in most diseases, the use of genome sequencing will allow an even more personalized approach to treatment. To quote Hippocrates, "It is more important to know what sort of person has a disease, than to know what sort of disease a person has" [19].

Medical consultants already face enormous challenges in keeping up to date with the rapid development of new treatments and medications, particularly for rare cases. Information providers offer evidence-based medical information services, continuously taking into account new publications and medical developments. Prominent examples are up-to-date (www.uptodate.com) and DynaMed Plus (www.dynamed.com).

Combining symbolic and non-symbolic approaches

Symbolic approaches are based on machine readable as well as human readable ("symbolic") knowledge representation. Symbolic approaches were dominant in AI research from the mid-1950s until the mid-1990s. The most successful applications of symbolic machine intelligence are expert systems that use production rules. Prominent technologies include classic AI languages like Prolog and Lisp, but also business rules engines and Semantic Web technology. For performance reasons, dedicated hardware was built, such as Lisp machines, which were able to execute Lisp as a native machine language. The Semantic Web standard OWL adopted description logic, allowing the use of reasoners to infer ontological information, and RIF enabling Web rule interchange based on RuleML.

A large AI hype was triggered by the Fifth Generation Computer Systems program, which was initiated by Japan's Ministry of International Trade and Industry in 1982. It led to the development of numerous AI technologies. A large community of AI researchers and developers conducted projects in domains such as diagnosis, consulting, rating, and many more [2]. In the 1990s, the hype vanished ("AI winter"). It can be assumed that the expectations were too high, and the results in terms of cost–benefit were not sufficient at this time.

Non-symbolic approaches have a similarly long history in AI research, dating back to the first ideas about neural computing in the 1940s. The most prominent examples are ANN, which simulate the fundamental physical (neural) processes in the brain. Enormous progress has been made in non-symbolic approaches and, in the meantime have become the dominant branch of AI. They include, e. g., ANN, deep learning, and ML models, such as support vector machines, Bayesian networks, Hidden Markov models, etc.

In recent years, a new AI hype seems to been partly founded on hardware advances, such as GPUs, which enable efficient, non-symbolic AI approaches such as deep learning. Companies and countries have identified ML as a field to have a competitive advantage in practically every IT application domain. Billions of dollars are invested into AI projects and into the strategic investment of companies and AI experts. As a result of the breakthroughs achieved recently, this is accompanied by an even more intense ethical discussion about the consequences of AI.

Both approaches, symbolic and non-symbolic, have advantages and disadvantages, which complement each other. Non-symbolic approaches are more robust against noise and provide better results, particularly in domains with little upfront knowledge. They usually require a lot of training data, but at the same time, scale up to Big Data more easily. However, they lack means for explaining a solution chosen. There lies the strength of symbolic approaches. Since they are based on explicit symbolic knowledge representation and reasoning, the chain of reasoning can be explained to human experts.

In sum, it is obvious that the combination of both approaches may provide benefits and has, therefore, been researched for quite some years.

Some approaches are hybrid in nature. A prominent example are Bayesian networks. A Bayesian network is a graph with states as nodes and conditional dependencies as edges. For example, a Bayesian network could represent the relationships between diseases and symptoms ("A cold may lead to a cough"). The Bayesian network is modeled by a human expert, here a doctor. This is the symbolic aspect of Bayesian networks. Then, observed statistics on the individual states and their combinations (e. g., "How many patients with a cold had a cough?") are added. This is the statistical, hence non-symbolic aspect. Based on Bayes' theorem on conditional probabilities, the Bayesian network can now be used to compute the probability of certain diseases, given the patient's symptoms.

Symbolic approaches can be enhanced by non-symbolic ML approaches. For example, production rules and association rules may be suggested by data mining approaches and validated by human experts, inductive logic programming supports rule learning, and text mining can be used as basis for ontology learning. It also works the other way. Non-symbolic ML approaches may be improved by the use of symbolic rule systems and ontologies, e. g., by enriching the training data with additional background knowledge or by testing, validating, and interpreting the learned models with symbolic (expert) knowledge. Logic-based rules can, e. g., interpret and reason about situations and make decisions based on learned event detection models. Background knowledge such as rules can be also used in combination with reinforcement learning. This approach has become particularly popular in recent years, e. g., by Alphabet's AlphaZero.

It will be interesting to see if the current success of non-symbolic approaches will be seconded and sustained by symbolic approaches in the future.

Trends in natural language processing

NLP has been a core area of research in and applications of AI for many decades. Inspired by early successes in computer sciences in the 1950s, scientists and funding organizations were quickly convinced that mastering natural language understanding and machine translation were the next logical step and should be reached within a few years. These challenges, however, proved to be much harder than originally anticipated. Truly robust, high quality, and generic NLP systems have been around for only a few years now.

The combined effects of the relentless increase in available computing power, the abundance of natural language data, as well as advances in algorithm design have profoundly transformed the field. Today NLP is a key element in strategic decisions in many large companies for its potential to support digital products and services and also the promise to help reduce costs in managing language-based information. A recent study [23] predicted the worldwide market for NLP powered applications and services to grow from ∼3bn$ in 2018 to over 25bn$ in 2025. Graduates in computer science, data science, and computational linguistics are likely to find themselves in a market that offers attractive prospects and opportunities.

A key aspect behind many of the breakthroughs of NLP systems since the 1990s is the increased adoption of non-symbolic approaches in the respective algorithms, which complement and sometimes replace traditional symbolic, rule-based approaches. Part-of-speech tagging and natural language recognition were among the first NLP subfields where these quantitative approaches became the de-facto standard, and as of today most other NLP disciplines such syntactic parsing, machine translation, semantics, text categorization, or information extraction are deeply influenced or dominated by them. Young et al. [24] lists recently reported results on a wide range of NLP tasks and deep learning powered systems. A technique specific to distributional semantics ("word embeddings", see the next section) has been labeled as the "secret sauce" of many NLP applications today [16]. A survey [24] of papers submitted to large NLP conferences over the last few years found that by 2016/2017, approx. 70 % discussed applications of deep learning. This further underlines the transformation that NLP has experienced in the recent years.

The impressive successes of deep learning based approaches may suggest that "traditional" knowledge based methods may no longer be needed. However, many experts expect that the need for explicit linguistic (symbolic) knowledge to be used in combination with non-symbolic, neural approaches still exists and that both approaches will continue to benefit from each other. This is especially true in industrial contexts, where the available corpora are often limited in size and where at the same time the subject to be analyzed comes with pre-existing structured knowledge bases. This background knowledge does not need to be learned, but instead can be imported and used right away.

For the sake of giving an example: Why trying to learn that "Hodgkin's lymphoma" is a kind of "neoplasm" when this is well represented in structured vocabularies in the medical/life science field? Also, that "bought" and "buys" are members of the same verbal paradigm will remain a task to be better addressed with a good morphology enabled lexicon, rather than hoping that this relation falls out of the analysis of their respective distributions in corpora. As Young et al. [24] formulate it: "[...] we expect to see more deep learning models whose internal memory (bottom-up knowledge learned from the data) is enriched with an external memory (top-down knowledge inherited from a knowledge base). Coupling symbolic and non-symbolic AI will be key for stepping forward in the path from NLP to natural language understanding."

Clustering word semantics

Word embeddings represent a computational approach to grasp word semantics by means of distributed representations. While traditional approaches represent the meaning of words in single components of high-dimensional vectors with up to hundreds of thousands of components, word embeddings use a different approach. They distribute the meaning over all components of smaller dimensional vectors (in the range of hundreds of components) and use a computational approach to determine the proper position of the word in the corresponding vector space. In the case of *Word2Vec* [18] and *GloVe* [20], the computational approaches used are based on word co-occurrences, and in the case of *fastText* [5], on n-grams. While the latter approach allows relating syntactically similar words, the former two are especially interesting from the semantic viewpoint.

The intuition behind *Word2Vec* and *GloVe* is that they push similar words into the same region and unsimilar words into different regions, thus imposing a structure on the vector space. Although their computation is purely based on co-occurrences, this results in vector space structures from which high-quality synonyms and closely related terms could be identified by their closeness (in terms of Euclidean distance) or their similarity (in terms of cosine vector similarity).

Such similarity information could be used for knowledge engineering or semantic processing purposes if there would not be a major obstacle: a distance or similarity can be computed between every pair of words. Thus, some way is needed to determine a sensible distance or similarity threshold up to which words should be considered related.

A corresponding threshold could simply be configured, but the threshold will always be arbitrary. However, first experiments have shown that an unsupervised graph-clustering approach can be used to derive clusters of useful synonyms and closely related super and sub terms. The basic idea behind this approach is to derive a similarity graph from a subset of important words determined by TF-IDF [21] and connect them with weighted edges if the angle between their word vectors similarities is smaller than 45° (cosine similarity > 0.7). In order to determine a tighter angle and to compare different clustering approaches, however, an approach to compare the quality of the clusters derived still needs to be found.

Named entity recognition

The set of named entities often offers a concise access to the content of a text. Knowing that a news article contains the named entities “Benjamin Netanjahu”, “Mahmud Abbas”, and “Heiko Maas” allows humans to quickly understand what this text might be about. Using the named entities of a text for indexing and reading enhancement, therefore, has a long tradition. Often, named entities are highlighted in the text or listed in dedicated sections, such that the reader of a business magazine or a yellow press tabloid can quickly verify whether and where the persons he or she might be particularly interested in are mentioned.

Person names are one subtype of named entities, places, or company and organization names are other examples. Named entities can be described as referring to one specific thing in our world. The expression “Elon Musk” refers to one specific person[2], while the common noun expression “table” refers to the set of all table-like objects.

Given this significance for searching and accessing textual information, it comes as no surprise that named entity recognition (NER) has been an active area of research in computational linguistics and data science for decades. NER comprises typically different sub-challenges that need to be addressed: The named entity must be detected in unstructured documents (“Is this expression a named entity or not?”), the recognized expression must be disambiguated (“Which one of the potentially many persons or places of the same name is referred to here?”), normalized (“What is the canonical way that this entity shall be referred to?” For example, “Neustadt a.d. Aisch” versus “Neustadt an der Aisch”) and ideally linked, i. e., tagged with a link to a database entry like a person’s entry in a corporate registry or to the URL of a company’s homepage.

It is important to understand that these steps are necessary for many high value tasks: Imagine your boss asks you to check the patent portfolio of the US pharmaceutical company Merck in the area of oncology. Typing “Merck AND oncology” into a naïve patent search engine will not only return

[2] Leaving aside the fact that there may still be several persons with this name.

hits of the German company Merck from Darmstadt and from the independent American company Merck, but will also miss potentially important hits, since some Merck subsidiaries do not even contain the string Merck in their name (e. g., Multilan AG or MedAdvisor Inc.). Considering that a patent portfolio can easily consist of many thousands of complex documents, it is evident that this seemingly simple research task can quickly turn out to be prohibitively complex and time consuming, unless we have access to powerful and accurate NER capabilities.

As with many similar NLP tasks, NER was traditionally predominantly addressed with lexicon-based and rule-based approaches in the past, while in recent years, ML approaches have complemented, and in some places replaced, these earlier methods. Checking the top entries for NER results at [22], we find nothing but deep learning powered approaches occupying the top positions of the list that reports results on the CoNLL 2003 NER task. While this underlines the dominance of deep learning approaches in scientific evaluations, it has been noted that the challenges in many NLP tasks differ between the scientific/academic world and industrial use. In [8] the authors explain some motivation for this position: While learning-based approaches often excel where sufficient training data are available, even in today's age of big data, a real-world project task in NLP may not come with the required amounts of data in many industrial settings. Also explainability is much more of an issue in the industrial world than in most academic settings, according to the authors. Industry users often require the ability to track why a certain decision has been made, which is still a lot harder in deep learning powered approaches today.

As a result, many promising approaches to NLP challenges attempt to benefit from combining NLP-inspired methods with the more recent quantitative approaches, such as deep learning. Examples for these hybrid strategies are luminoso.com, explosion.ai, or semiring.com.

Current NLP tools

It might be hard for a practitioner to select the best tool for a particular NLP use case. Since most NLP research is geared towards the English language, finding a good NLP tool to analyze non-English texts might be even harder. This section gives a brief overview of openly available NLP tools and development frameworks with a focus on multilinguality, based on the experience of practitioners and researchers.

A relatively new NLP tool is spaCy[3], an open-source NLP library written in Python and Cython. Its main advantages are its focus on speed and memory efficiency, as well as interoperability. spaCy allows the use of models trained by TensorFlow, PyTorch, scikit-learn, and Gensim. It follows the "convention over configuration" principle, which makes it easy to obtain first results. Another strength of spaCy is multilingualism because it provide statistical models for eight languages out of the box, as well as multilingual preprocessing and NER modules.

GATE[4] (General Architecture for Text Engineering) has been around for more than two decades. It is an NLP tool suite written in Java and comes with modules for tokenization, gazetteering, sentence splitting, POS tagging, NER, and coreference tagging. It supports twelve languages and comes with a GUI, which makes it easy to configure custom NLP pipelines. GATE comes with an API (called "GATE embedded"), which allows using the NLP components in applications.

UIMA (Unstructured Information Management Architecture) is an abstract development framework for NLP workflows and is an OASIS standard. An open-source reference implementation written in Java is available[5]. DKPro[6] is a collection of reusable UIMA components, including statistical tools, text similarity algorithms, word sense disambiguation, a Hadoop connector for Big Data analysis, and many more.

Some products reported to yield good results when dealing with German texts include Schmidt's TreeTagger[7], Brants' TnT statistical POS tagger[8], and WinRelan, which supports the GABEK ("Ganzheitliche Bewältigung von Komplexität", Holistic Mastery of Complexity) method[9]. It is reported to perform particularly well when it comes to resolving co-references across sentence borders.

[3] https://spacy.io.
[4] https://gate.ac.uk.
[5] https://uima.apache.org.
[6] https://dkpro.github.io.
[7] http://www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger.
[8] http://www.coli.uni-saarland.de/∼thorsten/tnt.
[9] https://www.gabek.com.

OdeNet: the open German WordNet

The Open-de-WordNet (OdeNet) initiative provides a German resource in the multilingual WordNet initiative. WordNet links the concepts (the synsets) of languages. The resources are available under an open-source license, which will be included in the NLTK language processing package [4]. WordNet resources are largely used in NLP projects all over the world. The idea of the language technology group at Hochschule Darmstadt – University of Applied Sciences is to create a German resource that starts from a crowd-developed thesaurus. The resource will be open and included in the NLTK package, such that it will be further developed by researchers while using the resource for their NLP projects.

For the first version, existing resources were combined: The OpenThesaurus German synonym lexicon (https://www.openthesaurus.de), the Open Multilingual WordNet English resource (http://compling.hss.ntu.edu.sg/omw20/omw), the Princeton WordNet of English (PWN) [10] and the OMW data [6], which was made by matching multiple linked wordnets to Wiktionary, and the Unicode Common Locale Data Repository.

The first version of OdeNet was created in spring 2017 as an experimental project at Hochschule Darmstadt – University of Applied Sciences. It was created fully automatically. Manual corrections were made to OdeNet on lexical entries in the domains of project management and business reports. German definitions were introduced, relations were corrected and supplemented, and ili links (links to the multilingual concepts) were added. In autumn and winter of 2017, we worked on the syntactic categories. The main focus was on correcting the POS tags of multiword lexemes. The next step in winter 2017/18 was the annotation of basic German words, as listed in https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Grundwortschatz. We annotated all lexical entries (except for function words) of this list with *dc:type="basic_German"* in OdeNet, added missing entries, and corrected synsets manually. Then, we implemented an analysis of German nominal compounds and used this information for the addition of hypernym relations.

The resulting WordNet resource contains about 120,000 lexical entries in about 36,000 synsets. About 20,000 of these synsets are linked to synsets in the English WordNet PWN and then to the multilingual ili IDs. There are 2,664 antonym relations and 1,053 pertainym relations linking lexical entities, as well as about 10,000 hyponym relations and about 1,000 meronym relations linking synsets. The resource is available on github at https://github.com/hdaSprachtechnologie/odenet.

Processes, models, and ontologies

Process models describe how people and machines interact and which results are created in which situation. Describing a process in a formalized way makes it easier to understand and to discuss it. Often, process models are enriched with context information such as forms or transactions that are to be used while executing the process. This suddenly turns the model into a means of navigation. Enriching process models dynamically with semantically enriched information creates a basis for understanding the context in which an instance of the process has been executed.

Process models are sometimes looked at as a nice drawing exercise or as a graphical programming way. If you combine process models and ontologies, another aspect of process modeling arises: understanding the process context and thus allowing for an even deeper understanding of the behavior of an organization.

What does this mean? Process modelers can build their models based on a controlled vocabulary instead of using plain text to name their model elements: Combine concepts (classes) and verbs (methods) to create tasks, and concepts and participles or attributes (states) to create events. A task named "create invoice" could be made up of a class "invoice" that is assigned a method "create" resulting in an event "invoice done" based on this class "invoice" that is also assigned a state "done". Concepts can form an ontology or just a kind of vocabulary, meaning that they can be part of a formal ontology or just elements, e. g., of an informal list.

How does this help us to understand the process context? If you have a defined concept that your task is made of you can match this concept with similar concepts of given information systems or metadata of documents, e. g., the tags that are assigned to a document. As there might not be a direct match with a given document or, e. g., a transaction defined in an ERP system, you can use semantic technologies to create parameters for a search either in the document set or, e. g., in a semantically enriched information system. In our example, the process model would offer an example invoice or the most recent invoice that has been prepared. The user could use this as a template. This would then allow collecting more context data while an instance of the process is executed and help us understand why this instance was carried out this way (in this case, which template was used and why this invoice was the result of this process and not a different one).

Another use case could be the match with elements of an organizational knowledge graph. Linked data mechanisms could be used to adapt the actual process execution by, e. g., changing the fields of a form that was assigned to a process task in a process model according to the change in the knowledge graph. A form that is to be used in the process contains a group of associations that contain the address of a person. This information is semantically matched to an element of a knowledge graph that contains a similar group of associations. If the element of the knowledge graph is changed (e. g., add the association "has pet"), the form could also be enhanced by the new association.

Data quality via metadata

The W3C PROV Standard

While members of the semantic community actively worked with W3C standards, such as RDF, RDF-S, OWL, SPARQL, RIF, etc., it seems that one W3C standard was not yet widely recognized by the semantic community. PROV is a W3C recommendation (finalized in 2013) with a declarative first-order logic interpretation for the representation, recording, sequentialization, and querying of *provenance* information. It is described as a lean task ontology consisting of three major concepts (agent, activity, and item) and seven properties relating those concepts. These properties allow describing the items used by and generated by an activity, the agent to whom items and activities are attributed or associated. Three additional reflexive relations allow describing agents acting on behalf of other agents, which items are generated from other items, and which activities are informed by other activities.

This simple recommendation has a broad range of possible uses for intelligent applications. Provenance information can be used to derive statements about the quality, reliability, or trustworthiness of agents and information items, e. g., for data driven journalism. The documentation of the origin of collected and derived information, and of the processes used during the derivation, is not only useful meta-information for data management, but also for data science in general and ML in particular to document and explain derived insights. Additionally, provenance information can be used to derive dependency information for causal inference. Of course, process models are a natural source for the generation of provenance information, which describe activities consuming and producing items and the agents to whom they are related.

However, the application of this standard is not only restricted to the information domain or applications in computer science. PROV itself is open and in addition to the description of information, it also allows the description of provenance information itself and, more importantly, the recording of the provenance of real-world entities. This makes it applicable for traceability of products in production processes, supply chains and logistic transport applications, investigations on copyright management, money laundering and private equity, and for analyzing cultural influences and impacts in digital humanities, to name just a few applications.

Privacy and linked data

Kirrane and Wenning [15] describe a system using data annotations with Linked Data. The annotations are used to add semantics about policy and permissions to data in order to inform about privacy properties of a given data set. In several technical variants, this approach is also known as the "sticky policy" paradigm. With the policy information about a data set at one's fingertips, this information can be used to control data flows.

Imagine a shop service. While browsing, the service is aware which policy applies and writes log data. However, once the log data has been written, the context is lost.

For fear of privacy violations, large amounts of data in data warehouses and data lakes remain unused. Data from the past was often just stored in silos without policy information having been recorded. These data can only be used via anonymization, meaning the loss of a lot of relevant information.

In order to control data flows across company borders, all interested parties have to understand those semantics. This is a typical task for standardization. The goal is to cover the most commonly used semantics in one or more standards. This way,

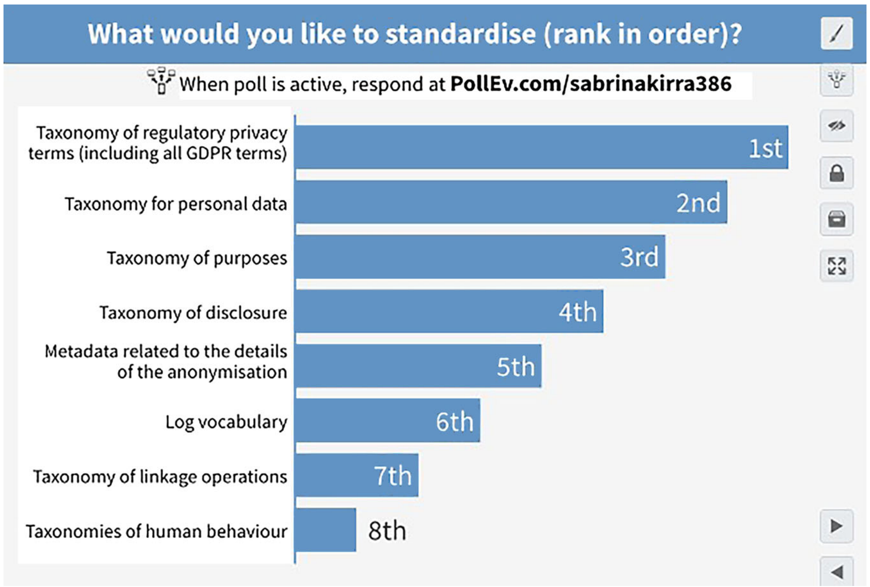


***Fig. 2** Tasks in privacy and linked data from the W3C Workshop on Data Privacy Controls and Vocabularies*

only edge cases need burdensome handcrafted customization. One example of such a standard is the W3C Recommendation PROV (see the previous section). Once the origin of data has been stored, algorithms can benefit from this information to decide whether or not to use the data in a certain context.

However, why use linked data? This has to do with the semantics in the policy, especially in with respect to privacy. The privacy policies of today are typically pages of legalese, where a service and its functionality is described. The policy then has rules and information about the use, storage, and processing of the data. To be able to write such a policy, the entire system needs to be described. This makes such policies extremely hard to read [17]. With linked data, the granularity changes. It is no longer necessary to describe an entire system and then describe the use of data in general. Instead, the policy information may link to the data directly. To be extensively usable, this link has to be bidirectional. The policy information can now point to an object that it applies to. It does not point to a description of the data, but to the data itself. If the policy information points to a data record, we end up with a semantically complete phrase that can be used in a legal context.

In order to have complete sentences that provide sufficient semantics for a legal context, semantics have to be established and agreed upon. As for provenance, we need vocabularies for purposes, retention, etc. A W3C workshop on data privacy controls and vocabularies in April 2018[10] explored these needs of interoperability. At the end of the workshop, a poll with suggestions for next steps was created (see Fig. 2). The next steps with the highest priority were to develop a taxonomy of privacy terms (which include, in particular, terms from the GDPR) and a taxonomy of personal data. Classification of purposes (i. e., purposes for data collection), of disclosures, and of methods of data anonymization were other next steps.

Subsequently, the W3C Data Privacy Vocabularies and Controls Community Group (DPVCG) was created[11]. The DPVCG started to work on the GDPR taxonomy and further use cases.

Conclusion

Today, AI is a hype topic, and its usefulness is hardly disputed. However, hype leads to exaggerated expectations. To suggest more realistic expectations, we favor the alternative term "machine intelligence".

In this article, we focused on industrial application use cases. We presented examples from the business sectors insurance and medicine. The current AI hype focuses mainly on ML and especially on the recent successes of ANN. Without disputing the importance of ML approaches, we regard this an unnecessary restriction. We see the combination of symbolic approaches with non-symbolic approaches

[10] https://www.w3.org/2018/vocabws/.
[11] https://www.w3.org/community/dpvcg/.

(like ML) as an important current trend. This trend is particularly visible in approaches and applications of NLP. We presented some details on word embeddings, named entity recognition, and word nets.

Finally, we emphasized that for developing data-intensive applications in a corporate context, it is also important to consider privacy and provenance issues. In this article, we have also given insights in current developments like the W3C PROV standard.

We will continue to share our experiences in developing intelligent corporate applications in Dagstuhl workshops and to publish our results. If you work on intelligent applications in corporate contexts, you are cordially invited to participate next year's workshop (contact: Bernhard Humm, Thomas Hoppe).

References

1. Academy of Medical Sciences (2015) Stratified, personalised or P4 medicine: a new direction for placing the patient at the centre of healthcare and health education (technical report). Academy of Medical Sciences. May 2015. https://acmedsci.ac.uk/viewFile/564091e072d41.pdf. Accessed 23 Sept 2018
2. Bense H, Bodrow W (1995) Objektorientierte und regelbasierte Wissensverarbeitung. Spektrum Akademischer Verlag, Heidelberg
3. 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
4. Bird S, Klein E, Loper E (2009) Natural language processing with Python. O'Reilly Media, Sebastopol, CA
5. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146, http://aclweb.org/anthology/Q17-1010, last access: 24.9.2018
6. 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
7. 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 Informat Technol (CIT) 23(1):29–41, https://doi.org/10.2498/cit.1002508
8. Chiticariu L, Li Y, Reiss FR (2013) Rule-based information extraction is dead! Long live rule-based information extraction systems! in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, 18–21 October 2013. Association for Computational Linguistics, Stroudsburg, pp 827–832
9. Ege B, Humm B, Reibold A (eds) (2015) Corporate Semantic Web – Wie Anwendungen in Unternehmen Nutzen stiften. Springer, Heidelberg (in German)
10. Fellbaum C (ed) (1998) WordNet: An electronic lexical database. MIT Press, Cambridge
11. Harris-Ferrante K (2017) To the Point: Leveraging AI for Success in Digital Insurance. In: Presentation Gartner Symposium ITXPO, Nov. 5–7, 2017, Barcelona, Spain
12. 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
13. Hoppe T, Humm BG, Reibold A (eds) (2018) Semantic Applications – Methodology, Technology, Corporate Use. Springer, Berlin
14. Humm BG, Walsh P (2018) Personalised clinical decision support for cancer care. In: Hoppe T, Humm BG, Reibold A (eds) Semantic Applications – Methodology, Technology, Corporate Use. Springer, Berlin, pp 125–143
15. Kirrane S, Wenning R (2018) Compliance using metadata. In: Hoppe T, Humm BG, Reibold A (eds) Semantic Applications – Methodology, Technology, Corporate Use. Springer, Berlin, pp 31–45
16. Manning CD (2015) Computational linguistics and deep learning. Comput Linguist 41(4):701–707
17. McDonald A, Cranor L (2008) The cost of reading privacy policies. I/S J Law Policy Inf Soc. 2008 Privacy Year in Review issue. http://aleecia.com/authors-drafts/readingPolicyCost-AV.pdf, last access: 20.11.2018
18. Mikolov T et al (2013) Efficient estimation of word representations in vector space. https://en.wikipedia.org/wiki/ArXiv https://arxiv.org/abs/1301.3781, last access: 24.9.2018
19. Murugan R (2015) Movement towards personalised medicine in the ICU. Lancet Respir Med 3(1):10–12
20. Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation. in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 25–29 October 2014, Doha, pp 1532–1543. https://nlp.stanford.edu/pubs/glove.pdf, last access: 23.9.2018
21. Robertson S (2004) Understanding inverse document frequency: On theoretical arguments for IDF. J Doc 60(5):503–520, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.2284&rep=rep1&type=pdf, last access: 28.8.2018
22. Ruder S (2018) Repository to track the progress in Natural Language Processing (NLP). https://github.com/sebastianruder/NLP-progress, last access: 23.9.2018
23. Tractica.com (2017) Natural language processing market to reach $22.3 billion by 2025, August 21. https://www.tractica.com/newsroom/press-releases/natural-language-processing-market-to-reach-22-3-billion-by-2025, last access: 23.9.2018
24. Young T, Hazarika D, Poria S, Cambria E (2017) Recent trends in deep learning based natural language processing. IEEE Comput Intell Mag 13(3):55–75