Knowledge Graph
Semantic Representation and Assessment of Innovation Ecosystems

Klaus Ulmschneider and Birte Glimm

Institute of Artificial Intelligence, Ulm University, Germany
{klaus.ulmschneider,birte.glimm}@uni-ulm.de

Abstract. Innovative capacity is highly dependent upon knowledge and the possession of unique competences can be an important source of enduring strategic advantage. Hence, being able to identify, locate, measure, and assess competence occupants can be a decisive competitive edge. In this work, we introduce a framework that assists with performing such tasks. To achieve this, NLP-, rule-based, and machine learning techniques are employed to process raw data such as academic publications or patents. The framework gains normalized person and organization profiles and compiles identified entities (such as persons, organizations, or locations) into dedicated objects disambiguating and unifying where needed. The objects are then mapped with conceptual systems and stored along with identified semantic relations in a Knowledge Graph, which is constituted by RDF triples. An OWL reasoner allows for answering complex business queries, and in particular, to analyze and evaluate competences on multiple aggregation levels (i.e., single vs. collective) and dimensions (e.g., region, technological field of interest, time). In order to prove the general applicability of the framework and to illustrate how to solve concrete business cases from the automotive domain, it is evaluated with different datasets.

1 Introduction

Continuous change is undeniably one of the main characteristics of our modern world. New technologies, techniques, business models, and processes are developed and evolve rapidly over time, primarily driven by the requirement to diversify from others and to cope with the pace of change. Individual and collective creativity, paired with existing knowledge and know-how, can be constituted as the backbone of this development and result in new knowledge or, consequently, in inventions. Effective R&D management, and, in particular, innovative capacity, is highly dependent upon knowledge and, above all, human individuals, and therefore inevitable linked with general business strategies [11]. Consequently, it is of increasing importance to be capable of analyzing knowledge occupants and specialists, who are driving the change. In principle, innovation is driven by mutual complementarities between individuals and organizations, i.e., the actual know-how carriers and problem-solving capacity. The interaction of knowledge occupants such as researchers, engineers, or organizations, as well as their created artifacts in consideration of time and location can be constituted as an innovation ecosystem. Thus, being able to capture, measure, locate, and assess such causal and contextual, local and global (competence) correlations, and making them explicit, can be a significant advantage from a strategic viewpoint (e.g., innovative impact, knowledge flows, knowledge gap identification, competitive assessments) and constitutes a competitive advantage when being capable of adapting rapidly to environmental transformation processes.

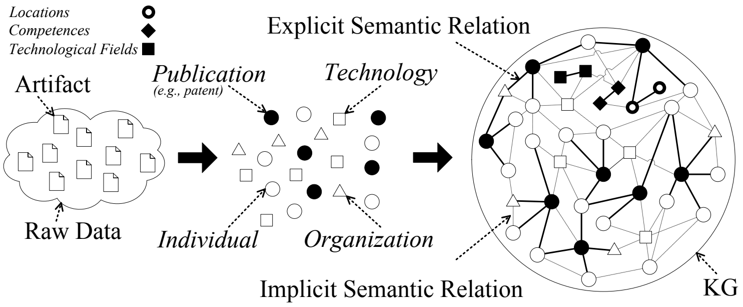

Fig. 1: Knowledge Graph Construction

In order to keep up with the process of continuous change, this paper introduces a novel methodology using several NLP, machine learning, and AI techniques combined with OWL reasoning. The aim is to gain multilevel competence information from publications on individual level (persons) and on multiple collective levels (e.g., department, institute, firm, university, industry, sector) as well as to capture their structural (e.g., institute belongs to university), temporal, and spatial (e.g., state, region, country, continent) arrangement in order to enable the analysis of an overall knowledge ecosystem (e.g., investigation of interactions and collaboration). The proposed framework allows for mutual exploitation of knowledge and know-how complementarities as well as knowledge flows and interactions between individuals and among organizations, or, their analysis with regard to regional (e.g., a firm's branches) or geopolitical (e.g., a nation's) aspects with the fundament of a graph-based representation.

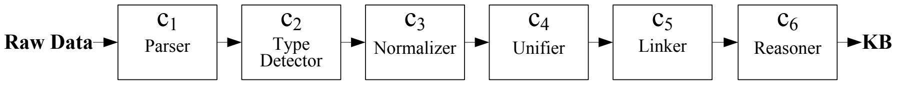

Fig. 2: Fundamental Processing Components

Specifically, knowledge occupants are identified, extracted, normalized, disambiguated, unified (if required), and brought into context with the help of the created *Knowledge Graph* (KG), which constitutes a (semantic) RDF graph and allows to store structure-lending entities as well as complex (semantic) analyses with regard to various dimensions. For example, temporal, spectral, spatial, topical features in the KG design allow to capture and react to weak signals (e.g., competence transitions) and to track evolutionary (knowledge) pathways as well as to generate innovation stimuli or support decision-making on future directions. To achieve this, a Semantic Pipeline (SP) [14], which processes raw data and stores the gathered information in the *Knowledge Graph*, is applied, primarily comprising the following tasks:

1. Parse raw data (e.g., patents, academic publications) into uniform objects
2. Identify and extract real-life entities which can occupy knowledge (candidates) and determine their type
3. Normalize the surface forms of names and store them together with extracted metadata in standardized properties
4. Disambiguate and, if referring to the same entity, unify competence occupant mentions
5. Semantically link (identified) competence occupant entities with each other, with related (business) artifacts (e.g., patents) as well as aligning them with, for example, topical, organizational, or spatial entities, and, on this basis, with entities which define a competence (an ontology)
6. Semantic analysis of knowledge interdependencies on multiple levels and in consideration of various dimensions (e.g., identification of competence clusters, knowledge flows, or competence paths by deducing implicit information by means of the semantic relations created in the previous step)

In particular, the presented framework allows for extracting and determining different types of entities from unstructured, semi-structured and structured textual content and transforming them into respective (structured) entities, and interlink them semantically by means of the KG. Further, the framwork can distinguish between different types of competence possessing entities, which enables the detection of competence ownership and analysis of competence clusters on several aggregation levels based on technological fields of interest. By enriching traditional (text) processing with components from the field of Semantic Web and Machine Learning, (knowledge) interdependencies and inferences of (implicit) information on multiple (competence) levels are enabled.

The benefits of the presented framework and its underlying semantic representation are illustrated with scientific publications as well as with patents. Patents are employed since they represent, besides academic publications, a large proportion of technological and procedural knowledge and occupy interesting (implicit) relational knowledge. Rapidly increasing filed numbers of patents as well as diverse patent application strategies and complex ambiguous writing styles make patent analyses a difficult and time-consuming task (see e.g., [13, 14, 15]). In contrast, a significant amount of intellectual capital is reflected in patents which make their analysis a worthwhile exercise. Patents are applied for by physical (individual) or legal (organizational) entities, i.e., the actual carriers of knowledge (e.g., skills, qualifications, experiences) and creative potential. The sheer amount of filed patents and the increasingly growing number of patent applications consequently result in high technological impact. Hence, patents are indicators of research and development efforts and the analysis of patenting activity constitutes one way of measuring intellectual capital (e.g., technological competence owned by an organization) [2]. Therefore, patents are a valuable (large scale) source of scientific and technological information for R&D processes as well as for strategic decision making (see e.g., [4]).

It has be shown that patents can be analyzed across various dimensions, e.g., with regard to technological progress, technology planning, technological forecasting, R&D portfolio management, or infringement analyses (see e.g., [5, 13, 3]) and that they can be aligned with technological fields of interest [14]. However, what has been neglected so far is to extend their exploitation with regard to new methodologies, processes, technologies, or materials, to the experts as well as the think tanks who possess this know-how (i.e., the physical and legal entities) in order to create value for multiple application scenarios, which include knowledge identification, knowledge acquisition, knowledge localization as well as the management and early development of resources with regard to market changes (e.g., human resource development, knowledge transfer, recruiting, headhunting, R&D management, competitor foresight, supplier identification).

The contributions of this paper are as follows. A framework is introduced, which combines various interdisciplinary research areas from several fields of computer science and an algorithmic model to accurately derive competence information in order to identify competence occupants and their interdependencies as well as allowing their analysis along multiple (aggregated) dimensions. For this reason the identified individual and collective competence occupant entities are integrated into the KG along with other types of entities (e.g., with the patents they applied for, patent classification systems as well as spatial, geopolitical, or topical entities) which, overall, formally represent a complex innovation ecosystem. On this basis, a reasoner is employed to exploit competences and their (semantic) interrelations and knowledge pathways being capable of deducing implicit information which provides more insights in depth and breadth. Various analysis techniques, such as text mining, network analysis, citation analysis or index analysis are combined to discover meaningful implications. The results can be reused in combination with gathered temporal information. Besides increased general transparency of non-obvious valuable information, various use cases are enabled, and therefore, the proposed framework can serve multiple stakeholders and application scenarios.

The presented framework overcomes the limitations of existing approaches with regard to the following key aspects: improved normalization, unification, and integration algorithms which identify, organize, and interlink competence occupants based on structured, semi-structured, or unstructured data comprising syntactic, lexical, semantic, relational, and machine learning techniques as well as aligning the identified competence occupants with conceptual systems and adding reasoning capability to the underlying (processed) data, i.e., the KG. The KG embodies a complex ecosystem, which is capable of representing competence occupants and their peripheral (i.e., established, structural, or interacting) entities (e.g., scientific publications, technological fields of interest, employers) in a semantic manner as well as evolutionary knowledge pathways and clusters (i.e., knowledge creation and diffusion over time). Therefore, this work extends the focus from analyzing information from the actual raw artifacts to the layer of processing, deducing, and analyzing derivative information, which is not directly observable, i.e., obtaining derivative information from artifacts of other purpose. For example, the purpose of patents is not primarily analyzing knowledge occupants and knowledge flows.

The paper is organized as follows. Section 2 describes the central components of the presented framework. Section 3 demonstrates the applicability of the approach and presents the findings along with real-world business cases from the automotive domain. Finally, Section 4 discusses related work and Section 5 concludes with a summary and an outlook.

2 Competence Analysis

Identifying competence occupants and their mapping to actual competences and skills is a nontrivial task. In order to achieve high accuracy, multiple (pre-)processing steps need to be accurately performed. These processing steps include the determination of their (entity) type and structural level (e.g., individual vs. collective) as well as the normalization of occurring surface forms and their disambiguation. Then, the gathered competence occupants are semantically mapped with other entities (e.g., topical, spatial) to gain valuable insights from the *Knowledge Graph*. This section outlines inherent challenges of revealing competences, demonstrates the crucial processing steps to identify them accurately, and illustrates the capitalization of the gathered semantic graph for strategic purposes.

2.1 Entity Type Recognition and Disambiguation

In real-world, many types of entities exist. Examples include persons, organizations, locations, technologies, or materials. All of them can be helpful for knowledge analyses and are usually referenced within publications, such as patents, without being explicitly labeled with their (entity) type. In order to identify such named entities and to map them to their corresponding type for further processing (e.g., determination of competence level or interdependencies), they have to be identified and labeled correctly. Depending on the data source, entities must be parsed from structured or semi-structured sections of the respective artifacts, or extracted from unstructured content (with help of NLP components, i.e., *Named Entity Recognition* (NER)). In the context of patents, important information regarding competences is available in semi-structured formats and therefore can be parsed with reasonable effort. However, entities do not occur separated by their type. In case of scientific publications, algorithms exist to extract authors and their affiliations, but nowadays such information is usually available in structured or semi-structured formats as well. Having raw names of potential competence occupants available, their type is determined by using specific patterns and indicators which are employed and extracted during the normalization process illustrated in the next section.

2.2 Entity Name Normalization

Entity Name Normalization (ENN), also known as *Name Standardization*, refers to the standardization of name variants which can occur due to, for example, misspellings, abbreviations, or different naming conventions. Therefore, the process of name normalization attempts to transform surface forms, i.e., name variants, into a common format.

Competence occupants and their respective names, which are mentioned in artifacts such as scientific publications, frequently occur with several surface forms, might be incomplete, not formatted according to a common standard, extended with additional information not belonging to the actual name (e.g., titles, addresses, states), or acronyms are used. Hence, name normalization is a nontrivial task [1, 8, 9]. Table 1 illustrates some typical ENN challenges.

Challenge	**NameType**	**Group**	**Example #1**	**Example #2**
Orthography	Individual	Syntactic	Doe, John	JOHN DOE
Orthography	Organization	Syntactic	Wal-Mart	Wal*mart
Diacritical Marks	Individual	Syntactic	Ulf Lindström	Ulf Lindstroem (Ulf Lindstrom)
Diacritical Marks	Organization	Syntactic	Telefónica S.A.	Telefonica
Compound Names	Individual	Syntactic	Jean-Claude van Damme	van Damme, Jean-Claude
Transliteration	Individual	Writing Systems	Крассимир	Krassimir
Transliteration	Organization	Writing Systems	上海大学	Shànghâi Dàxué (Shanghai University)
Prefixes	Individual	Semantic	Mr. John Doe	Dr. John Doe
Prefixes	Organization	Semantic	Walt Disney	The Walt Disney Company
Suffixes	Individual	Semantic	John Doe Jr.	John Doe, MSc
Suffixes	Organization	Semantic	Exxon Mobil Corp.	Exxon Mobil Corporation
Acronyms	Individual	Semantic	John D. Doe	John Daniel Doe
Acronyms	Organization	Semantic	Univ. Beijing Technology	Beijing Tech Univ
(Other) Metadata	Individual	Semantic	John Doe, NY	John Doe, USA
(Other) Metadata	Organization	Semantic	Daimler AG, Stuttgart	DAIMLER 70567 STUTTGART DE

Table 1: Selected prevalent Entity Name Normalization challenges

The proposed normalization component covers such challenges by means of rule-based, syntactical, orthographical, semantical, statistical, lexical, and relational aspects. It attempts to reduce a raw name to its core components, i.e., removing all information not belonging to the actual name itself while extracting as much metadata and pieces of evidence as possible for further processing (e.g., academic or honorific prefixes for individuals or legal forms for organizations). Note that this process is entity type dependent and respective entities with their dedicated properties are created. For example, in case of an individual name the normalizer separates first name, middle names (if any), and last name into dedicated fields and is capable of taking several combinations, orders, or punctuation (e.g., Doe, John D.) into account.

Additionally, names are normalized with regard to three normalization stages: the raw normalized name (core), a human display name (e.g., uniform casing, acronyms expanded) and one form for machine processing (e.g., punctuation, diacritical marks, special characters, stop words removed). Moreover, the structural (aggregation) level, in particular with regard to competences, is determined as well (i.e., single competence vs. collective competence, e.g., institute vs. university) and respective associations are created. Note that additional information and (extracted) metadata (e.g., NY=New York, Jr.=Junior) is, if possible, normalized and filed in the same name object during this processing step as well. Table 2 illustrates a successful normalization process for two individual and three organization names. Note that the overall entity type detection and normalization process is implemented with a feedforward and feedback loop to update potential incorrect assignments. For example, individuals are usually not mentioned together with legal forms, however person names can be the same as or part of a company name.

Raw Name	Normalized Name Object (Extract)
Dr. John Francis D. Smith Jr.	[FirstName=John,middleNames[Francis,D],FamilyName=Smith,prefixes[Doctor],suffixes[Junior]]
Smith, Dr. John F. D.	[FirstName=John,middleNames[F,D],FamilyName=Smith,prefixes[Doctor],suffixes[]]
North Texas University	[Name=North Texas University,type=Academic]
University of North Texas	[Name=North Texas University,type=Academic]
Nike, Inc.	[Name=Nike,type=Business,legalForm=Incorporated]

Table 2: Exemplary initial rudimentary entity name normalization steps

2.3 Entity Unification

Entity Unification (EU), also known as *Entity Linking*, *Entity De-Duplication*, *Reference Normalization*, *Instance Unification*, *Record Linkage*, *Coreference Resolution*, or *Entity Resolution*, refers to the process of determining whether two entities (i.e., name mentions) refer to the same object (e.g., a person) in real-world, and, if referring to the same entity, mapping them to a canonical unambiguous referent (see e.g., [1, 10]).

This processing step, which is building up on the *ENN* component, is essential, because competence occupants occur in various forms and the process of disambiguation and mapping therefore can have strong effect on the accuracy of single and collective competence assignments [7, 9], and, in particular, on higher level deductions. In consequence, the (normalized) surface form of potential competence occupants, such as individual or organization names, need to be disambiguated and, in case of ambiguity, unified in preparation for further processing steps and, eventually, the population of the *Knowledge Graph* to accomplish the competence analysis task. Table 3 illustrates some typical EU challenges.

Challenge	NameType	Group	Example #1	Example #2
Spelling Mistakes/OCR Errors	Individual	Semantic	I.F. Kennedy	John Fitzgerald Kemedy
Spelling Mistakes/OCR Errors	Organization	Semantic	Tesla Inc.	Telsa Inc.
Similar Names	Individual	Semantic	Jonathan Meier	Jonathan Meyer
Similar Names	Organization	Semantic	TLG Immobilien	TAG Immobilien
Acronyms	Individual	Semantic	J.F. Kennedy	John Fitzgerald Kennedy
Acronyms	Organization	Semantic	IBM	International Business Machines Corporation
Translations	Individual	Multilinguality	Franz	Francis
Translations	Organization	Multilinguality	Universidad de Chile	University of Chile
Marriage/Name Changes	Individual	Temporal	Hillary Diane Rodham	Hillary Diane Rodham Clinton
Mergers/Splits/Acquisitions	Organization	Temporal	Mannesmann	Vodafone Group
One Name - Multiple Entities	Individual	Semantic/Temporal	John Doe (Dover)	John Doe (New York)
One Name - Multiple Entities	Organization	Semantic/Temporal	Merck & Co., Inc.	Merck KGaA
Multiple Names - Same Entity	Individual	Semantic/Temporal	President Trump	Donald Trump
Multiple Names - Same Entity	Organization	Semantic/Temporal	Daimler AG	Daimler-Benz AG
Ambiguities/Missing Information	Both	Semantic	Trump	Trump
Structural (level)	Organization	Semantic	Institute of Artificial Intelligence, Ulm University	Ulm University

Table 3: Selected prevalent Entity Unification challenges

Entity unification is achieved using a fuzzy matching approach which combines several techniques and matching rules. For example, phonetic (distance-based), feature-based, and probabilistic similarity measures are employed. Furthermore, meta information as well as graph-based (i.e., implicit and explicit references) and statistical indicators, which can be derived from the source artifacts, are incorporated. In particular, the algorithms combine exact, partial and approximate matching (experiments were conducted with several similarity measures such as Cosine Similarity, Jaro-Winkler, Levenstein) on all normalization stages (RawProcessed, Display, Machine). Further, peripheral features, i.e., explicitly and implicitly (derived) references, are examined. Examples include topical and spatial associations, citations, references to (patent) classification systems, or the analysis of coreferences (e.g., co-authorship among authors in case of scientific publications or among inventors and assignees in case of patents as well as individual-organization associations). One important factor, which is often neglected, is time. Names can change due to marriage, mergers, splits, and acquisitions. If such information is available, it is reused for the unification process as well. Note that, in case of unification, all references are updated and every known surface form (variants, including the raw name and all normalized forms) is stored with the unified object (entity) and reused for further normalization and unification tasks. The best normalized name (in format and without metadata) is selected as reference name.

2.4 Multidimensional Competence Assessment

The main challenges of effective competence assessment are to accurately determine the possession and location of individual and collective competences as well as the capability to track their temporal evolution.

In order to cope with these challenges, the presented framework allows to identify, structure, (semantically) interlink, and therefore measure and analyze competence occupants with regard to multiple dimensions and in consideration of temporal aspects. The idea behind the approach is that all types of entities have causal relationships with each other (cf. Figure 3), i.e., competences are mutually dependent or influenced from other entities and several inverse deductions can be made.

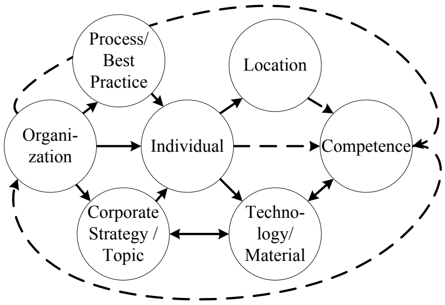

Fig. 3: Exemplary Causal Competence Relationships in a Business Context

In particular, the framework semantically links the gathered normalized and unified competence occupants with their established artifacts (i.e., their publications) as well as with other entities such as other business-related artifacts (e.g., technology fact sheets) and conceptual systems such as technological fields of interest, locations, or patent classification systems (see [14] for more information). On this basis, the competence occupants are aligned with entities which (formally) describe their actual competence(s) analogously. The semantic relations include, among other properties, the (relation) type (e.g., 'hasTopic', 'hasInventor', 'hasAssignee', 'belongsTo[Company]', 'isLocated') and respective temporal information (e.g., point in time of person-company association). Note that skills and competences, which are hierarchically organized and interlinked with each other, are further integrated with other knowledge-related entities (e.g., entities representing topical information) using semantic relations. Hence, they bridge the gap between the competence occupants and the entities which describe the actual skills and competences. Thus, 'hasCompetence' relations, competence clusters, and knowledge pathways in the KG can be inferred (e.g., by applying transitivity rules) on multiple levels using a reasoner, thus, enabling competence assessments. As example, consider multiple employees having associations with a competence such as Artificial Intelligence. If such a pattern is detected, the respective competence can be attributed to the overall company or academic institution of the competence holder. Hence, by analyzing the overall KG, such competences can be deduced and become explicit on higher level associations. The same applies for the localization of competences, temporal analyses (e.g., competence development paths) and dedicated technological disciplines.

One important factor of the *Knowledge Graph* design is the concept of subsumption. In previous processing steps the aggregation level of competences is implicitly derived and transformed to respective (causal) relations, i.e., single competences are associated with collective competences (competence clusters). As example, consider research groups, institutes, R&D teams, laboratories, medical centers, think thanks, or organizational units, which can be further aggregated, e.g., with regard to research projects, firm or university level, or depending on several dimensions, with respect to an industry's, a region's, a nation's, topical, or higher competence levels (e.g., Artificial Intelligence → Computer Science).

In contrast to related work, dedicated entities from processed information are created and interlinked with other explicitly or implicitly gathered entities. This conceptual and representational difference allows the framework to assess competences on multiple levels, to drill up and down, and to answer complex (business-related) questions based on other related entities, which are also associatively and hierarchically organized. Moreover, individual and collective (competence) interactions can be captured by utilizing temporal features.

Specifically, with the gathered *Knowledge Graph*, the framework is further capable of identifying competences which can assist to solve a given problem, detecting potential knowledge gaps, tracking changing competence strategies of competitors or which nations are building up competences in a certain technological field of interest.

The reasoner can, in combination with SPARQL queries, which consider the established semantic relations as well as the concepts of transitivity and subsumption, deal with such business cases and provide answers to concrete questions. For example, the reasoner is able to derive institutes which are associated with a university as well as their employees and therefore deduce the competence range of the university based on the employees' competence relationships.

3 Validation

This section presents the research design and experimental results conducted with the proposed framework using several general datasets as well as patent datasets of interest within the automotive domain.

3.1 Research Design

The analysis framework is written in Java and extends the (semantic) processing pipeline used by Ulmschneider and Glimm [14] with normalization-, unification-, and competence-related components (see previous sections). The *Knowledge Graph* (KG), which is constituted by RDF triples and its defining ontology (OWL 2 RL profile), is prepopulated with several conceptual systems such as competences,[1] technological fields of interest, locations, and several patent classifications systems (IPC, CPC, USPC). All of them are hierarchically organized and semantically interlinked. In order to evaluate the presented framework, multiple datasets are used. The preprocessing components are evaluated with the following datasets:

- University names from the literature [10, 6]
- World universities (all university names of the world)
- Large companies with high impact (companies listed on major stock market indices)
- Abstracts of scientific publications (KDDCup hep-th papers 1992-2003[2]), containing metadata such as titles, authors, affiliations, dates

Further, two multilingual patent datasets from two emerging technological fields of interest, which are relevant for the automotive domain, are used for evaluating the overall competence recognition and analysis framework with regard to business-related questions:

- Alternative Mobility Concepts (AMC)
    - Electro Mobility (EM)
    - Hydrogen Mobility (HM)
- Artificial Intelligence (AI)

The two integrated patent datasets contained more than 13,600 patents from the areas of Artificial Intelligence (28.38%), Electro Mobility (41.08%) and Hydrogen Mobility (30.54%). Most patents were applied for in the United States (51.4%), followed by Japan (25.35%), Germany (7.17%), and China (3.56%). All other patents were filed in other countries (12.52%), whereas the distribution of identified languages (textual content) was 71.13% (English), 16.01% (Japanese), 5.47% (German), 3.67% (Chinese), and 3.72% (other languages).

Note that all datasets are based on real-life data. The general datasets are used as baseline to evaluate entity type detection, name normalization, and entity unification whereas the integrated patent dataset is employed for detecting

[1] Incorporates a domain-specific competence taxonomy combined with the ESCO ontology (European Skills, Competences, Qualifications and Occupations, see `http://ec.europa.eu/esco/` for more information)

[2] `http://www.cs.cornell.edu/projects/kddcup/datasets.html`

and evaluating expertise for the technological fields of AMC and AI on multiple dimensions. In order to analyze patents and align them with technological fields of interest the above-mentioned extended semantic processing pipeline (SP) was employed. Remember that the SP was enhanced by competence-specific components. Based on assignments to technological fields of interest as well as associations from other derived features competence occupant profiles were then semantically interlinked with actual competences.

As evaluation metrics accuracy A, which we define as the percentage of correct results R_c out of all non-null results R_w as well as coverage C, defined as the percentage of non-null results to total results, are used. Hence, coverage incorporates null results R_n as well.

$$A = \frac{R_c}{R_c + R_w} \qquad C = \frac{R_c + R_w}{R_c + R_w + R_n}$$

Additionally, the success rate ($SR = A \times C$), which indicates how likely the framework succeeds to generate a correct result, as well as the F-measure (F_1 score), which constitutes the harmonic mean of accuracy and coverage ($F_1 = 2 \times A \times C/(A + C)$), are calculated and measure the overall performance of the respective processing steps.

3.2 Findings

The most important evaluation step constitutes the detection of individual and collective competences and determining their correct type. Since, to the best of our knowledge, no curated data is available for patents as a source of competence allocation and the creation of such a dataset is very labor-intensive, this step is evaluated with alternative datasets containing academic institution names, company names as well as person names. For academic institution names, a baseline dataset containing UK universities as used in Liu et al. [10] and Jacob et al. [6] is employed. Further, we extended this dataset with all other university names of the world, resulting in almost 12,000 instances. Note that both datasets exclusively include academic institution names. In order to evaluate the same with business organizations, we employed a dataset containing around 600 companies (i.e., their names) which are listed on major stock market indices. Table 4 lists the evaluation results.

Dataset	A_t	C_t	SR_t	F_{1t}	A_{n1}	C_{n1}	SR_{n1}	F_{1n1}	A_{n2}	C_{n2}	SR_{n2}	F_{1n2}
UK Universities	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
World Universities	1.00	0.99	0.99	0.99	0.78	0.99	0.77	0.87	0.85	0.99	0.84	0.91
Large Companies	1.00	0.93	0.93	0.96	0.97	0.93	0.90	0.95	1.00	0.93	0.93	0.96

A = Accuracy, C = Coverage, SR = Success Rate, F_1 = F-measure, (t) = type detection, $(n1)$ = name normalization (standard), $(n2)$ = name normalization (considering all normalization stages)

Table 4: Evaluation Results

Overall, the performance was satisfying with regard to accuracy and coverage for the name type detection and the name normalization task. We further evaluated whether the type of organization was identified correctly. For both, the UK and the World University dataset, the accuracy and coverage was close to 100% with regard to the determination of their organization type (academic) and structural level (e.g., institute, faculty, university). For business organizations the coverage of the organization type was slightly lower (91%) but also with an accuracy reaching almost 100%. Only the accuracy of identified legal forms was relatively low with 86%. After evaluating datasets containing collective competences we shifted our attention to single competences. Therefore, we utilized the KDDCup dataset (scientific publications) and extracted the included author information. After parsing, we received 62,664 person names and further processed them to finally retrieve unambiguous 19,993 person names. We then evaluated whether the framework is able to correctly assign the authors to their respective type (individual) and achieved 89% accuracy. Additionally, more than 2,200 associated organizations could be derived from author affiliation strings. Finally, we processed the integrated patent dataset and obtained 75,715 potential competence owner mentions. After processing the patents with the framework, 48,046 name mentions were annotated as individuals and 22,103 as organizations. For 5,566 mentions the name type could not be determined. Among the recognized organizations 20,625 were detected as business organizations and 858 as academic institutions, both with their respective kind (e.g., legal forms such as 'incorporated', academic types such as 'university'). For 620 organizations the type could not be uniquely determined. Moreover, the conducted experiments revealed that accurately determining entity types and profound name normalization can increase the accuracy of unification tasks for more than 9%, which, in consequence, improves the overall quality of the analyses to be conducted on basis of the KG.

After integration of all processed profiles, we applied the reasoner on the KG and studied several business cases (cf. Section 2). For example, we found that Microsoft attempt to concentrate their core competences (including cooperation partners) regarding Artificial Intelligence in the US. Further, regarding their patent strategy, they massively reduced patent applications in this area beginning from 2009 which might be a weak signal that the company will focus on other technologies (and competences) in the future. In contrast, Daimler increased their efforts with regard to electric storage systems in the past view years and protect their corresponding inventions globally, which can be interpreted as that they are building up and protecting their (core) competences in this area worldwide.

Summing up, the experiments revealed that considering metadata, such as spatial, temporal, topical, or relational information as well as implementing self-improvement mechanisms, can increase the accuracy of the overall gathered and computationally represented innovation ecosystem. Further, the capability of answering complex (business-related) queries on top of the KG makes the framework a considerably powerful tool. However, it must be noted that the overall processing and, in particular, the entity unification process with pairwise examinations is computationally extensive (i.e., the creation of the *Knowledge Graph*). In contrast, the upstream name normalizer turned out to be inexpensive and accurate. Nonetheless, we did not compare the results with (other) machine learning approaches so far and leave this as a future task.

4 Related Work

As illustrated in the previous sections, the presented competence analysis framework is interdisciplinary and combines as well as enhances techniques from several research areas. Therefore, related work is partitioned based on the fundamental components of the framework: Named Entity Detection and Name Normalization, Named Entity Disambiguation and Unification, and competence-, skill-, or expert- related analyses.

The task of entity name normalization and unification has been studied extensively. Solutions range from rule-based, dictionary-based, or string matching techniques to machine learning and hybrid approaches based on several types of data (e.g., (domain-specific) databases, websites) and application scenarios (newspaper articles, genes, diseases, employers, job postings, academic institutions) (e.g., see [1, 5, 6, 7, 8, 9, 10]). Some combine Named Entity Recognition (NER) with ENN, but few consider multilinguality, temporal aspects, or (hierarchical) dependencies (e.g., institute vs. university, Germany vs. Europe). Most approaches, however, have in common, that important information, such as metadata and relationships, are neglected. Many authors examine the problem of normalization and disambiguation as one single, isolated task (e.g., normalized strings vs. objects with extracted and normalized meta information and their relations with each other). Moreover, the differentation between structural (competence) levels and the obligatory process to use the gathered information with regard to higher level associations, including respective (business-related) analyses, receives almost no attention. In contrast, this work shifts the document-centric view (e.g., search engines, cross-document person name normalization) to the actual (multiple types of) entity mentions within documents, their inline references and cross-references among artifacts, which are, altogether, transformed into a respective graph-based representation.

Studies dealing with competences have several purposes. Some focus on creating a thesaurus or taxonomy, e.g., to improve search engines. Others create visualizations (e.g., competence maps), mostly based on quantitative techniques and for different purposes (e.g., see [2, 11]). For example, Moehrle et al. [11] create inventor competence maps from patents with focus to HR management and Barirani et al. [2] create competence maps based on patent citations to assess national and firm-level competences. They are able to identify and locate the largest invention communities in a given technological discipline. However, the approach requires patent citations as a prerequisite. While graphical presentations allow for deriving insights with regard to the big picture (e.g., to understand interdependencies on higher levels), concrete (qualitative) questions cannot be answered based on the underlying data (e.g., with regard to specific competences, competence occupant interactions, or competence developments). Zhao et al. [16] propose a system to recognize and normalize professional skills from resumés and matching them with a taxonomy created from Wikipedia categories and resumé sections. However, exact matches between taxonomy entries and extracted skills from textual content are required. Ronda-Pupo and Guerras-Martín [12] analyze collaboration correlations by measuring scientific output and impact of institutions in the academic community by employing graph-based metrics (degree centrality) to derive insights about an institution's relevance within a collaboration network in the discipline of management.

5 Summary and Outlook

Continuous change and the requirement to diversify from others to remain competitive requires highly qualified specialists who possess cutting-edge intellectual capital and who are capable of transforming ideas, technological know-how, and constraining specifications into business value. However, accurately identifying and allocating such expertise is a challenging exercise.

Hence, this paper presents an integrated framework to detect competence occupants in publications such as patents, and to represent them, along with the actual publications, conceptual systems and other business-related artifacts, as a semantic graph (KG). The resulting KG allows for their topical, structural, and spatial analysis and supports inferences on multiple dimensions while considering individual as well as several collective competence levels.

Accordingly, the pro-active exploitation and management of competences can be achieved for multiple application scenarios and HR managers, procurement managers, technical engineers, innovation managers, patent analysts, researchers, existing think thanks, or business analysts are capable of utilizing competence intelligence according to their specific needs.

Controlled experiments with multilingual patents on emerging technological disciplines, which are emphasized along with real-world application scenarios and business cases from the automotive domain, demonstrate the feasibility of extracting, processing, and analyzing expertise on multiple dimensions. In particular, we have shown how to identify individuals and organizations referenced in (scientific) publications (e.g., patents) and how to map them with indicators of expertise (e.g., related topical information) on multiple aggregation levels.

The conducted controlled experiments emphasize that the illustrated improved processing techniques can indeed increase the accuracy of identification and disambiguation of competence occupants and their alignment with other entities on individual and organizational level. Moreover, the implemented techniques allow for accurately determining the type of collective competence occupants (e.g., academic institution, business organization) and their structural level (e.g., institute vs. university). Hence, additional implicit competence information can be deduced using a reasoner which is capable of analyzing the structure and pathways (e.g., by traversing the KG) as well as higher level associations based on the semantic graph-based representation.

With this work we demonstrate how to detect, normalize, unify, aggregate, and interrelate competences in a structured and analyzable form. In order to enhance the proposed framework and add additional value, we will integrate further complementary types of relevant (business) artifacts (e.g., technology fact sheets, invention reports) to the KG and extend the KG with further semantics.

Altogether, the framework and its analysis pipeline will be further developed and enhanced with focus on integrated interlinked views on competences and complementary entities as well as their (latent) interdependencies targeting a broader view and allowing additional predictive features based on the representation of the gathered innovation ecosystem (e.g., competence requirement foresight, (competitor) competence activity predictions).

References

1. Aswani, N., Bontcheva, K., Cunningham, H.: Mining information for instance unification. In: Proc. 5th International Semantic Web Conference. vol. 4273 (2006)
2. Barirani, A., Agard, B., Beaudry, C.: Competence maps using agglomerative hierarchical clustering. Journal of Intelligent Manufacturing 24(2), 373–384 (2013)
3. Ernst, H.: Patent information for strategic technology management. World Patent Information 25(3), 233–242 (2003)
4. Giereth, M., Stäbler, A., Brügmann, S., Rotard, M., Ertl, T.: Application of semantic technologies for representing patent metadata. In: Informatik 2006. vol. 1, pp. 297–304 (2006)
5. Huang, S., Yang, B., Yan, S., Rousseau, R.: Institution name disambiguation for research assessment. Scientometrics 99(3), 823–838 (2014)
6. Jacob, F., Javed, F., Zhao, M., Mcnair, M.: sCooL: A system for academic institution name normalization. In: Proceedings 15th International Conference on Collaboration Technologies and Systems (CTS'14). pp. 86–93 (2014)
7. Jijkoun, V., Khalid, M.A., Marx, M., de Rijke, M.: Named entity normalization in user generated content. In: Proceedings 2nd Workshop on Analytics for noisy unstructured text data (AND '08). pp. 23–30 (2008)
8. Jonnalagadda, S., Topham, P.: NEMO: Extraction and normalization of organization names from PubMed affiliation strings. Journal of Biomedical Discovery and Collaboration 5, 50–75 (2010)
9. Khalid, M.A., Jijkoun, V., de Rijke, M.: The impact of named entity normalization on information retrieval for question answering. In: Proceedings 30th European Conference on IR Research (ECIR'08). vol. 4956, pp. 705–710 (2008)
10. Liu, Q., Javed, F., Mcnair, M.: CompanyDepot: Employer name normalization in the online recruitment industry. In: Proceedings 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). pp. 521–530 (2016)
11. Moehrle, M.G., Walter, L., Geritz, A., Müller, S.: Patent-based inventor profiles as a basis for human resource decisions in research and development. R&D Management 35(5), 513–524 (2005)
12. Ronda-Pupo, G.A., Guerras-Martín, L.Á.: Collaboration network of knowledge creation and dissemination on management research: Ranking the leading institutions. Scientometrics 107(3), 917–939 (2016)
13. Tseng, Y.H., Lin, C.J., Lin, Y.I.: Text mining techniques for patent analysis. Information Processing and Management 43(5), 1216–1247 (2007)
14. Ulmschneider, K., Glimm, B.: Semantic exploitation of implicit patent information. In: Proceedings 7th IEEE Symposium Series on Computational Intelligence (SSCI'16) (2016)
15. Zhang, L., Li, L., Li, T.: Patent mining: A survey. SIGKDD Explorations 16(2), 1–19 (2014)
16. Zhao, M., Javed, F., Jacob, F., Mcnair, M.: SKILL: A system for skill identification and normalization. In: Proceedings 29th Conference on Innovative Applications of Artificial Intelligence (AAAI'15). pp. 4012–4017 (2015)