APPLYING LEXICAL KNOWLEDGE FOR SEARCH AND NAVIGATION SUPPORT IN LEGAL DATABASES

Bernhard Waltl[1], Laura Altamirano Sainz[2], Florian Matthes[3]

[1]Research Associate, Technische Universität München, Department of Informatics, Software Engineering for Business Information Systems, Boltzmannstraße 3, 85748 Garching bei München, DE, b.waltl@tum.de;
[2]Student, Technische Universität München, Department of Informatics, laura.altamirano@tum.de,
[3]Professor, Technische Universität München, Department of Informatics, Software Engineering for Business Information Systems Boltzmannstraße 3, 85748 Garching bei München, DE, matthes@in.tum.de; https://wwwmatthes.in.tum.de/

Keywords: ***Search, Navigation, Lexical Knowledge, Ontology, GermaNet, Search Mechanism***

Abstract: *Searching and navigation in legal information databases are well studied tasks and highly relevant for practitioners and scientists. Modern legal information databases are the fundament of digital information retrieval, which becomes more and more relevant due to the fact that the amount of information increases continuously and rapidly. This paper investigates how the search and navigation process can be supported by the usage of lexical knowledge such as WordNet or GermaNet. Beside the lexical information of a language, those databases contain information about the linguistic relationships between words, e.g. hyponym and hypernym. An approach is proposed here about how these linguistic relationships can be used to support search and navigation in legal information databases. Analysis of the search term and proposing related hypernyms and hyponyms in the front-end enables users to expand or refine their search term depending on the selected word in the facet. From a legal science point of view, the suggestion of hypernyms or hyponyms has analogies to the well studied grammatical subsumption. The drawbacks, limitations and open issues regarding this approach are discussed at the end of this paper.*

1. Introduction

Data- and knowledge intensive tasks, such as the search processes in legal information databases consume plenty of valuable time within daily work of legal scientists and practitioners. Although it is already well studied, and plenty of high-quality articles covering multiple aspects of information retrieval (IR) on large datasets exist, this paper tries to unveil a further aspect that the integration of lexical information has. Thereby, the work is constrained to the IR of the legal domain. Consequently, some particularities, such as subsumption, exist which lead to the hypothesis that the usage of lexical knowledge and relations can be beneficial to the users during their search process. This article is an attempt for a better understanding of how lexical knowledge can be supportive to end users of legal information databases.

We have considered relevant literature from the IR domain (see Chapter 2) and made six expert interviews to empirically determine which lexical relations are important to the end-users (see Chapter 3). The architecture and components of the prototypical system is briefly introduced in Chapter 4. Finally, the limitations and drawbacks of our approach are stated out in Chapter 5 and the paper wraps up with a conclusion and outlook in Chapter 6.

1.1. Research Method and Objectives

The research objectives of this work address the question of how the consideration of lexical knowledge improves search in legal information databases. This is relevant to know how legal-database users can be supported and how query reformulation can be done more efficiently. Additionally, common mechanisms and methods in legal databases were investigated to help to derive a way to present lexical information to users by genuinely supporting them and enriching their experience. This work also covers a prototypical implementation for a support search mechanism integrated with lexical knowledge. According to Holtzblatt & Beyer (1997), "Systems are built to help people work better. They cannot be built well without understanding how people work". Therefore, a needs assessment with six experts has been performed with the goal of better understanding how potential target users of the system work and interact with such a system and whether they would find the proposal useful.

2. Related Work

2.1. Lexical Knowledge Databases

Lexical knowledge represents well-known information about words and relations between them (O'Hara, 2005). In the artificial intelligence approach, this is called "ontology" and pretends to construct a model that explains the relations of the entities (Pustejovsky & Bergler, 1992). As stated by Onyshkevich and Nirenburg (1995), an ontology is a model of the world, a body of knowledge of the world organized as a taxonomy. There exist many lexical relations between words; however, this research focuses only on a set of them (See Table 2.1).

Table 2.1 Lexical relations (Princeton University, 2010).

Relation	Description
Synonym	X and Y are interchangeable in some context without changing the truth value of the preposition in which they are embedded. E.g. a weapon is synonym of a gun.
Hyponym	X is a hyponym of Y if X is a (kind of) Y. E.g. a pistole is a hyponym of weapon.
Hypernym	Y is a hypernym of X if X is a (kind of) Y. E.g. a weapon is a hypernym of pistole.
Meronym	Member of a constituent part of something. X is a meronym of Y if X is a part of Y. E.g. a trigger is a meronym of weapon.
Holonym	The name of the whole of which the meronym names a part. Y is a holonym of X if X is a part of Y. E.g. weapon is a holonym of trigger
Troponym	X is a troponym of Y if X is to Y in some manner, i.e. X is a particular way of Y. E.g. a loaded rifle is a weapon
Coordinate terms (siblings)	Nouns or verbs that have the same hypernym. E.g. a pistole is a coordinate term of rifle (both have the hypernym weapon)
Derivationally related forms	Terms in different syntactic categories that have the same root form and are semantically related. E.g. gunman is a derivationally related form of gun.

To retrieve this lexical information, it has to be stored in a database. For this project, this is done by GermaNet, which is a lexical semantic net for German developed at the University of Tübingen which aims to model the German's base vocabulary. Its framework is compatible with the Princeton WordNet since the same technology for the database format and the database compilation is used (Hamp & Feldweg, 1997). GermaNet version 9.0 release 2014 contains 93 246 synsets (University of Tübingen, 2014). A synset is a set of lexical units (i.e. words) and represents the relation of synonymy (Henrich & Hinrichs, 2010).

2.2. Ontologies as Search Support in Legal Databases

Divoli et al. (2008) made a study involving biologists in which they could show that the majority of them would like to see ontological relations with regard to their search queries. They followed the user-centered design approach and performed two questionnaires. The first one to find out which kind of information participants would like to see, and the second one, after designing 4 scenarios and adding them to the group's search engine, to find out with which one the participants felt more comfortable with.

Another project pertinent to mention is the CIRI (Concept-based Information Retrieval Interface) system (Airio, Järvelin, Saatsi, Kekäläinen, & Suomela, 2004). This is based on a three level model that allows to build queries directly from ontologies (chosen by the user), which is an advantage as the system will search directly from the lexical relations to retrieve the results.

Bast et al. (2007) also presented an ontological-related efficient system: ESTER (Efficient Search on Text, Entities and Relations). This proposal was built for combined full-text and ontology search. In their paper, the authors highlight the importance of performing semantic searches and provide the system with an entity recognizer. Their system is able to perform prefix search and suggest possible semantic interpretations while executing the one which is more likely to be searched. We should note that CIRI allows the user to browse the ontologies, which lets the users know the lexical relations of the word(s) they are searching for, whereas, ESTER integrates the ontologies in a very different way, which makes them invisible for the users.

A model for extracting information in the legal domain was developed in the University of Vale do Rio dos Sinos - UNISINOS (Araujo, Rigo, Muller, & Chishman). This model makes use of semantic information and integrates linguistic information obtained from studying legal documents. They created a Brazilian-legal-domain ontology and a set of knowledge acquisition rules; results indicate good prospects when applying them to search within a set of documents. The authors state that it is possible to improve legal information retrieval with ontologies.

Saravanan, Ravindran, and Raman (2009) created a framework that sets the ontological information between the user interface and the database (built upon legal information) and deduced that ontologies enable inferences that can ensure effective information retrieval. According to their results, ontology-based searches generate much better results than traditional methods.

Another important project is the one developed by Dini et al. (2005). The LOIS (Lexical Ontologies for legal Information Sharing) project aims to build a semantically enriched and multilingual terminological database that follows the WordNet framework (Fellbaum, 1998., Miller, 1995). These authors stated that if knowledge is modelled by using ontologies or enhanced thesauri, then the ability to extract and exploit information from documents is enhanced by the explicit links that can be established among related items.

Expanding search queries has also been in the focus of the research done by Schweighofer and Geist (2007). They developed a methodology to improve Boolean search by using lexical ontologies and user relevance feedback. The former refers to query expansion with a lexical ontology; and the latter refers to collect relevance information from documents from an issued query to issue a second one.

Results when the system expands the query with synonyms (which are established by the system) were quite good. They admitted "that users may not be able to find a proper search term for the knowledge base as common language may use a different term", which still is a hard problem.

In 2015, Jörn Erbguth presented the improvements of search mechanisms in Swiss Lex focussing on the translation of search terms and the integration or exclusion of undesired terms. Thereby, the Thésaurus de Droit Suisse (TDS) in combination with linguistic tools have been used (Erbguth, 2015).

In summary, the ontology domain has been explored in many ways to be combined with databases. Prior research pursues to integrate the lexical knowledge in the background of the search systems, making it invisible to the user. On the other hand, there is also some work that requires the interaction of the user with the ontologies. Both approaches have produced promising results and consider that ontological information can improve the search process.

3. Evaluation of Search, Navigation and Exploration Mechanisms

3.1. Expert Interviews for Search in Legal Databases

Before performing the needs assessment, a comparison between different databases was done. This involved five legal databases, in which their search support mechanisms were analysed to investigate which mechanisms could be improved by using lexical information. The results are part of a Master's thesis in the field (Altamirano, 2015) and showed that mechanisms like the autocompletion, faceted navigation or automated term suggestions/query expansion could be enhanced by lexical information.

The needs assessment involved six people, all experienced in the law domain and all use a computer to perform their job more than 50% of their time (self-declared). During the interviews, they also mentioned that they review legal literature very often. All participants agreed that legal information databases will become more important in the future and that formulating the query in the search system is one of the most important procedures while searching. Answers were not so uniform when participants were asked whether they think that recommendations (offered by a system) of query reformulation could support the efficiency and effectiveness of their search: While three agreed completely, the other three were doubtful about the type of recommendations.

Four types of recommendations or tools to perform searches were presented to the interviewees. 100% of the participants answered that the selection of information source is completely relevant for the process. Whereas, the autocomplete tool was not as favoured.

Afterwards, three imaginary situations were selected to extract the ontological information (of the main word involved) from WordNet (lexical database) and to request the target users to rate whether they would find each relation to the word useful to give motivation or foundation to each of the cases.

In regard to the first situation, participants were asked the following question: *Imagine you are searching for the term "Abortion", what kind of information would you like the system to suggest? (Scale from 1 to 5, 5 being the highest)*. Figure 3.1 shows the interviewee's ratings to each of the lexical relations presented.

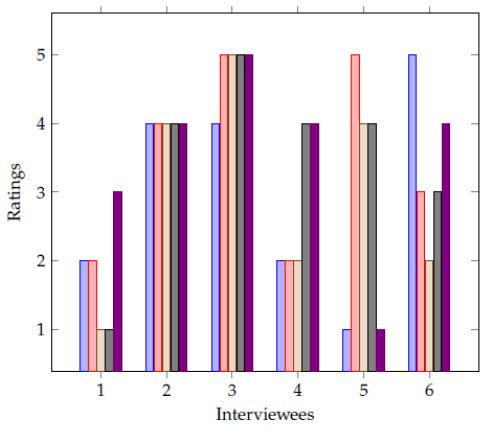

Figure 3.1 Ratings to the question: Imagine you are searching for the term "Abortion", what kind of information would you like the system to suggest? (Scale from 1 to 5, 5 being the highest).

	Average rating	%
Synonyms	3	50%
Hyponyms	3.5	58.3%
Hypernyms	3	50%
Coordinate terms (Siblings or related terms)	3.5	58.3%
Derivationally related terms	3.5	58.3%

Table 3.1 Summary of average ratings for the first situation.

Table 3.1 summarizes the average ratings. It can be seen that participants are more interested in seeing hyponyms of the word or its siblings and derived terms rather than synonyms or hypernyms. However, it is noticeable that there is not a big numerical difference between the ratings.

With respect to the second situation *(see ratings in Figure 3.2),* we can identify, in the average ratings (Table 3.2), that the target users would like to see troponyms as part of the suggestions. Although, they found coordinate terms (related terms) also relevant.

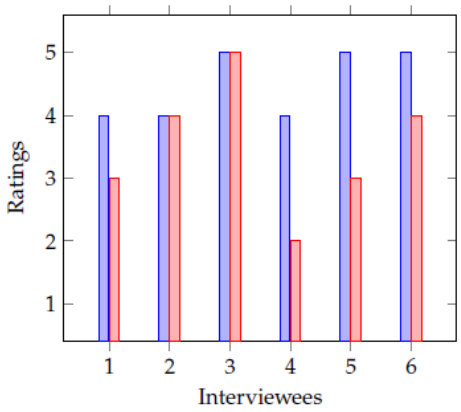

Figure 3.2 Ratings to the question: Suppose you got a case about discrimination against handicap. How useful would you find:...? (Scale from 1 to 5).

	Average rating	%
Troponyms	4.5	75%
Coordinate terms (Siblings or related terms)	3.5	58.3%

Table 3.2 Summary of average ratings for the second situation.

Table 3.3 shows the arithmetic mean of the ratings for the lexical relations involved in this question. These results were really interesting since, according to the answers, the presented relations appear to be useful only in certain law domains. While in others, they can be useful only in giving a broader overview to the lawyers and making them better understand the case but not to add facts to it. Holonyms' average rating in the third situation was only of 44.4%. One of the participants, who gave this relation a rating of only one, mentioned that in this case, law is only interested if the arm can be healed, how long it would take or if it is a partial or total loss. Therefore, such information would not be helpful.

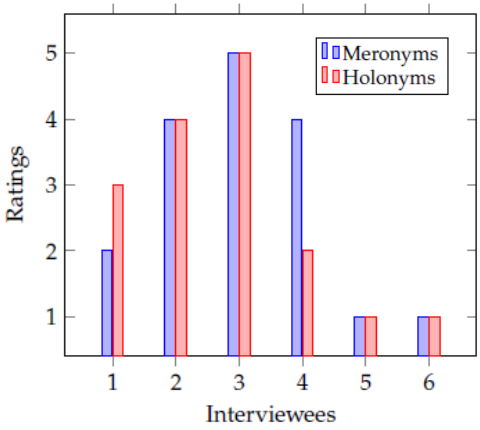

Figure 3.3 Ratings to the question: If you got a case about a person having an accident during work and as a result the right hand of the worker was seriously injured. How useful would you find a system that suggests:...? (also in a 1 to 5 scale).

	Average rating	%
Meronyms	2.8	47.2%
Holonyms	2.7	44.4%

Table 3.3 Summary of average ratings for the third situation.

Some interviewees mentioned that integrating this kind of knowledge in a search engine was a great idea and that this information would definitely support legal searches when used in the correct way. Furthermore, some of them said that the crucial aspect in this field is to restrict the results to the context and that this is precisely what many databases are missing nowadays.

On the whole, we can conclude that hyponyms, troponyms, sibling/related terms and derivationally related terms were the relations that the target users find more helpful. If we imagine the ontological tree of a certain word, we can say that hyponyms, troponyms and derivationally related terms are terms that are placed "below" the word, while siblings are placed in its same level. Therefore, to build the prototype system, hyponyms were selected to exemplify the integration of the lexical information to the system, as well as hypernyms combined with sibling terms.

3.2. Extending Search, Navigation and Exploration Mechanisms with GermaNet

Based on interviews and a study of search mechanism used in legal information databases, we developed a prototypical search system that is able to use lexical information for query reformulation (see Figure 3.4). User entered search terms are going to be processed to find semantically related terms (Step 1). Thereby, corresponding hypernyms and hyponyms are queried from GermaNet (Step 2). Afterwards, the occurrences of the hypernyms and hyponyms throughout the corpus are determined (Step 3). Those terms that occur at least once in the corpus are going to be presented in facets. The ones that are semantically related but never occur in the corpus won't be displayed in the user interface (Step 4 and 5). Finally, the resulting facets are presented to the users, which can then interact with this ontological information and expand or narrow their search queries in the system.

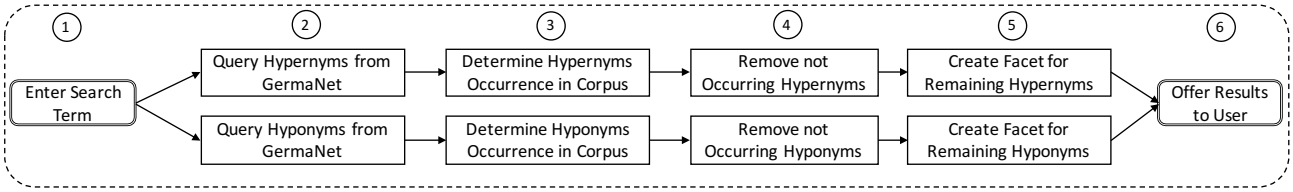

Figure 3.4 Search Query Reformulation Process

4. Software Concept and Prototypical Implementation

4.1. System Architecture

During the implementation of the system, fundamental software design principles such as low coupling of components were considered. The solution resulted in a combination of building two applications, a client-side application using AngularJS and a server-side component based on the Play Framework. This way, GermaNet can be integrated in a Java environment and Elasticsearch search engine integrated with AngularJS.

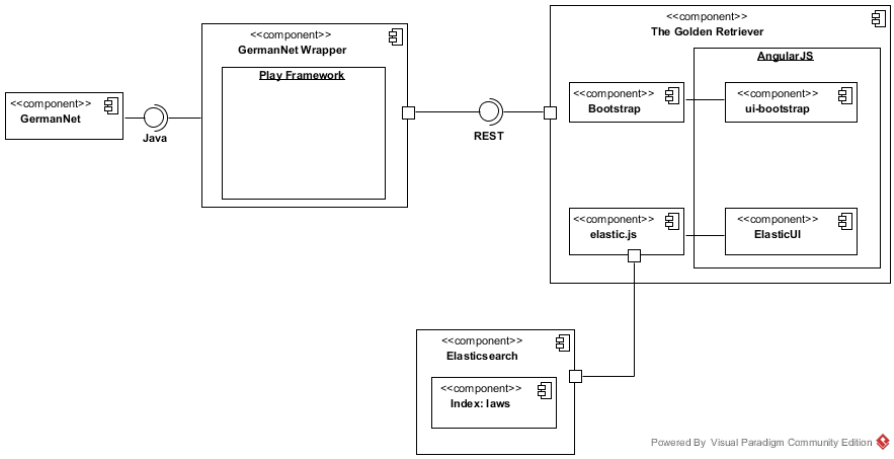

Figure 4.1 Overall reference architecture of the system fostering easy adaptability and reuse of components.

This architecture supports adaptability, since the component involving GermaNet (i.e. GermaNetWrapper) is separated from the other components. Although GermaNet covers a broad vocabulary, it has limitations regarding the legal vocabulary. In the future, this module could be replaced by a more specific ontological database focusing on the legal domain. Therefore, the separation and low coupling allows easy adaption and integration. It is important to notice that with this solution there will be three servers running in parallel: the AngularJS web-server (npm - node package manager (Joyent, Inc., 2015)), the Play framework server (through Typesafe Activator (Typesafe, Inc., 2015)) and an Elasticsearch instance. Here is a component description:

GermaNet and the GermaNetWrapper: is the component containing lexical information and serving as a lexical information retriever. It accesses the GermaNet database and creates the appropriate methods to return the lexical relations (i.e. hyponyms and hypernyms) of a given word.

The Golden Retriever: is the main module of the whole system. It is built with AngularJS framework and uses Bootstrap and UI Bootstrap to build the front-end. Furthermore, this module integrates the ElasticUI directives and elastic.js to retrieve information from Elasticsearch engine.

Elasticsearch: is the search engine, storing and indexing all legal documents.

4.2. Prototypical User Interface

Figure 4.2 shows the main user interface of the system. Users can perform queries and results will be filtered accordingly. After entering a search term, the facets at the left-hand side of the interfaces present the user with search support mechanisms including the ontological relations.

Within "Suche eingrenzen" (Narrow your search) in **Fehler! Verweisquelle konnte nicht gefunden werden.** the hyponyms of a given search term are shown. These suggestions should help the user to narrow the query. Additionally, we can see another search mechanism within *„Suche erweitern"* (Expand your search). This presents the hierarchy tree of the word's hypernyms. These elements are

also presented as a breadcrumb in the top of the search results. Using these mechanisms, a query reformulation can be performed by the users. The number of potential results is shown in the brackets right next to the search corresponding search term, e.g. hypernym, respectively hyponym. The popover that is shown in the query field shows the whole list of hypernyms of the word with the purpose to serve as a visual tool to reinforce the visual recognition of the hierarchy tree.

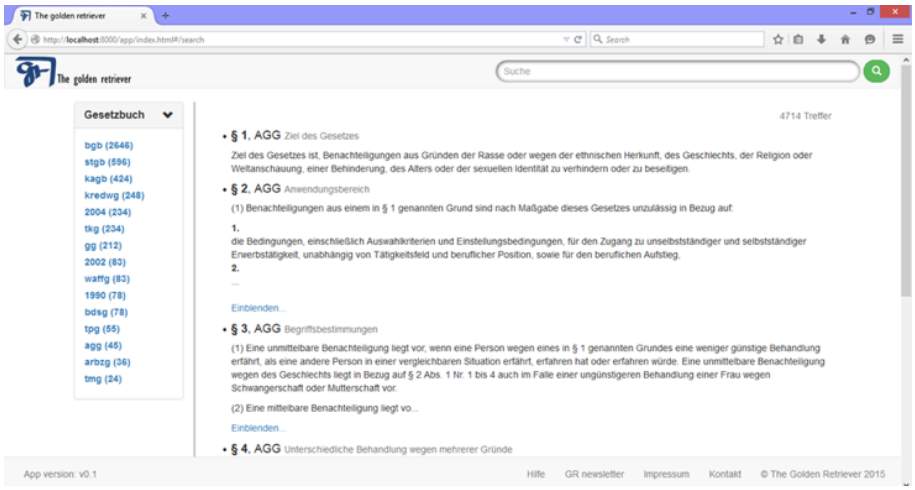

Figure 4.2 Main user interface

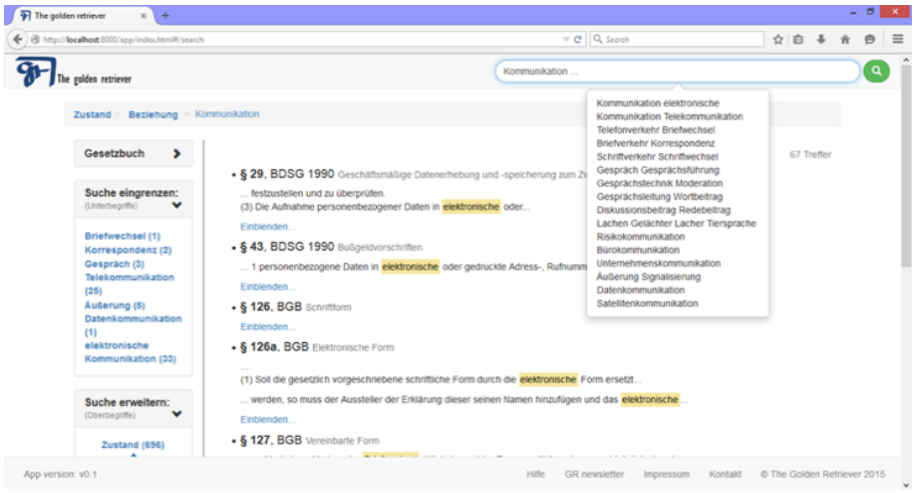

Figure 4.3 System features.

5. Critical Reflection and Drawbacks

One expert interview was conducted to evaluate the system. The main interest was to get feedback whether the system is clear for the users and to find out whether the interviewee thinks that the implemented search support mechanisms can actually help a legal database user. Results were favourable. On the one hand, the user interface was overall rated as very clear and clean. Moreover, the search support mechanisms presented seemed not confusing to navigate through the lexical terms and to display useful information for the search. On the other hand, some suggestions to improve the system were implied. These suggestions included simplifying some views and allowing customization. The overall impression that the interviewee got of the system was good but there is potential for the interface that would make it more useful.

The following limitations are faced by the current prototypical system regarding its implementation:

Constraining to single search terms can be considered as an obvious limitation of the current approach and implementation. At the current stage it is not possible to reformulate a more elaborate search query containing complex or longer phrases (n-grams). A possible workaround would be to consider only the nouns in the search query but this then leads to other implications, such as which term to replace (expand) with what information from the lexical database.

Word sense disambiguation is a further limitation for this work. This problem limits the system from extracting the precise word sense from the lexical database of the introduced search term by the user.

Combination of the search terms at its current implementation is done via reformulation of search terms. Thereby, the search queries get expanded by adding more terms to the overall search query. Future implementation might enable users to not only combine search terms with a logical OR but with more elaborate logical connections, such as AND or NOT.

GermaNet represents a limitation as well. The Golden Retriever depends on the powerfulness of this database at the moment, regarding the lexical content. We know that GermaNet contains an extensive amount of German words; nevertheless, it is not particularly tailored to the legal domain.

6. Conclusion and Outlook

This paper explores the usage of lexical knowledge stored in ontologies, e.g. GermaNet, to enhance search mechanisms of legal information databases. We examined the different types of lexical relations, which are provided by lexical knowledge databases, such as hyponym, hypernym, meronym, etc. Using those relations, it is possible to find related words to the one given in the search query. In a next step, we empirically investigated the usefulness of offering related words to searching users. For this, we conducted six expert interviews (legal practitioners and legal scientists). Based on the feedback of the experts we prototypically implemented a search system that allows for example, full-text search in German laws. Additionally, we integrated common search mechanisms (facets) offering lexical information hypernym and hyponym to end users. Using those facets, the end user can expand the search query with given and related terms. E.g., if a user enters the word “Schusswaffe”, the system automatically offers him the term “Waffe” to reformulate the entered query. In this case, “Waffe” is a hypernym (the more general term) of “Schusswaffe”, which is automatically detected by the prototype using the lexical knowledge database GermaNet.

A few drawbacks and limitations regarding our approach remain. The current approach only allows the reformulation of search queries with just one term and does not provide any mechanism for word sense disambiguation. Additionally, the current implementation only adds new search terms to the entered search query, but in reality a more complex combinations of search terms using logical operators may be required. Lastly, GermaNet is not tailored to the legal domain, such that the hypernyms and hyponyms relations may not reflect the relations between legal terms.

We consider our approach as a step towards the integration of lexical knowledge into search processes of legal scientists and legal practitioners. Although various challenges remain this could serve as a base line for upcoming research and implementations.

7. References

Airio, E., Järvelin, K., Saatsi, P., Kekäläinen, J., & Suomela, S. (2004). Ciri-an ontology-based query interface for text retrieval. In *Web Intelligence: Proceedings of the 11th Finnish Artificial Intelligence Conference* .

Altamirano Sainz, L. (2015) Applying lexical knowledge to improve search quality for a German legal information database. Master's thesis at Technische Universität München.

Araujo, D. A. d., Rigo, S. J., Muller, C., & Chishman, R. Automatic Information Extraction from Texts with Inference and Linguistic Knowledge Acquisition Rules. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (pp. 151–154).

Bast, H., Chitea, A., Suchanek, F., & Weber, I. (2007). ESTER: Efficient Search in Text, Entities, and Relations. In C. Clarke, N. Fuhr, & N. Kando (Eds.), *30th International Conference on Research and Development in Information Retrieval (SIGIR'07)* . ACM.

Dini, L., Peters, W., Liebwald, D., Schweighofer, E., Mommers, L., & Voermans, W. (2005). Cross-lingual legal information retrieval using a WordNet architecture. In G. Sartor (Ed.), *The 10th international conference on Artificial intelligence and law* (p. 163).

Divoli, A., Hearst, M. A., & Wooldridge, M. A. (2008). Evidence for showing gene/protein name suggestions in bioscience literature search interfaces. In *Pacific Symposium on Biocomputing* (Vol. 13, pp. 568–579).

Erbguth, Jörn; Bloch, Marc Sommer: Neue Suche bei Swisslex. In: Erich Schweighofer, Franz Kummer und Walter Hötzendorf (Hg.): Tagungsband des 18. Internationalen Rechtsinformatik Symposions IRIS 2015, Bd. 2015: OCG – Oesterreichische Computer Gesellschaft 2015.

Fellbaum, C. (1998.). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Hamp, B., & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (pp. 9–15).

Henrich, V., & Hinrichs, E. (2010). GernEdiT - The GermaNet Editing Tool. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, . . . D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* . Valletta, Malta: European Language Resources Association (ELRA).

Holtzblatt, K., & Beyer, H. (1997). Contextual design. In A. Edwards & S. Pemberton (Eds.), *CHI '97 extended abstracts* (p. 184).

Joyent, Inc. (2015). *Node.js*. Retrieved from https://nodejs.org/

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39–41. doi:10.1145/219717.219748

O'Hara, T. P. (2005). *Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions* (Ph. D. Thesis). New Mexico State University, New Mexico.

Onyshkevich, B., & Nirenburg, S. (1995). A lexicon for knowledge-based MT. *Machine Translation, 10*(1-2), 5–57. doi:10.1007/BF00997231

Princeton University. (2010). *Princeton University "About WordNet." WordNet.: WordNet 3.0 Reference Manual - Glossary of WordNet terms*. Retrieved from https://wordnet.princeton.edu/man/wngloss.7WN.html

Pustejovsky, J., & Bergler, S. (1992). *Lexical semantics and knowledge representation: First SIGLEX Workshop, Berkeley, CA, USA … 1991 : proceedings.*

Saravanan, M., Ravindran, B., & Raman, S. (2009). Improving legal information retrieval using an ontological framework. *Artificial Intelligence and Law, 17*(2), 101–124.

Schweighofer, E., & Geist, A. (2007). Legal Query Expansion using Ontologies and Relevance Feedback. In *LOAIT* (pp. 149–160).

Typesafe, Inc. (2015). *Typesafe: Activator*. Retrieved from http://typesafe.com/activator

University of Tübingen. (2014). *GermaNet - A German Wordnet: GermaNet - An Introduction*. Retrieved from http://www.sfs.uni-tuebingen.de/GermaNet/