Developing and Orchestrating a Portfolio of Natural Legal Language Processing and Document Curation Services

Georg Rehm[1], **Julián Moreno-Schneider**[1], **Jorge Gracia**[2], **Artem Revenko**[3], **Victor Mireles**[3], **Maria Khvalchik**[3], **Ilan Kernerman**[4], **Andis Lagzdins**[5], **Marcis Pinnis**[5], **Arturs Vasilevskis**[5], **Elena Leitner**[1], **Jan Milde**[1], **Pia Weißenhorn**[1]

[1]DFKI GmbH, Germany; [2]University of Zaragoza, Spain;
[3]Semantic Web Company, Austria; [4]K Dictionaries, Israel; [5]Tilde, Latvia

Corresponding author: `georg.rehm@dfki.de`

Abstract

We present a portfolio of natural legal language processing and document curation services currently under development in a collaborative European project. First, we give an overview of the project and the different use cases, while, in the main part of the article, we focus upon the 13 different processing services that are being deployed in different prototype applications using a flexible and scalable microservices architecture. Their orchestration is operationalised using a content and document curation workflow manager.

1 Introduction

We present a portfolio of various Natural Legal Language Processing and Document Curation services currently under development in the collaborative EU project LYNX, in which a consortium of partners from academia and industry develops a platform for the easier and more efficient processing of documents from the legal domain. First, our platform is acquiring data and documents related to compliance from multiple jurisdictions in different languages with a focus on Spanish, German, Dutch, Italian and English, along with terminologies, dictionaries and other language resources. Based on this collection of structured data and unstructured documents we create the multilingual Legal Knowledge Graph (LKG). Second, a set of flexible language processing services is developed to analyse and process the data and documents to integrate them into the LKG. Semantic processing components annotate, structure, and interlink the LKG contents. The LKG is incrementally augmented by linking to external data sets, by discovering topics and entities linked implicitly, as well as by using machine translation services to provide access to documents, previously unavailable, in certain languages. Finally, three pilots are developed that exploit the LKG in industry use cases.

The remainder of this article is structured as follows. Section 2 describes the different use cases, while, in the main part of the article (Section 3), we focus upon the 13 different processing services used in the prototype applications . The orchestration of the services is operationalised using a content and document curation workflow manager (Section 4). After a brief review of related work (Section 5) we summarise the article and take a look at future work (Section 6).

2 Use Cases

Within LYNX we work with three different use cases embedded in use case scenarios. In the following, we briefly sketch the three use cases.

The objective of the *contract analysis* use case is to enhance compliance with data protection obligations through automation, reducing costs, corporate risks and personal risks. The prototype analyses data protection legislation and case law from the EU and Member States and contracts between controllers, data subjects, processors, data processing policies and general contracts.

The *labour law* use case provides access to aggregated and interlinked legal information regarding labour law across multiple legal orders, jurisdictions, and languages. The prototype analyses labour legislation from the EU and Member States, and jurisprudence related to labour law issues.

The *oil and gas* use case is focused on compliance management support for geothermal energy projects and aims to obtain standards and regulations associated with certain terms in the field of

Proceedings of the Natural Legal Language Processing Workshop 2019, pages 55–66
Minneapolis, Minesotta, June 7, 2019. ©2019 Association for Computational Linguistics

geothermal energy. A user can submit a RFP or feasibility study to the system and is then informed which standards or regulations must be taken into consideration to carry out the considered project in a compliant manner. This scenario will innovate and speed up existing compliance related services.

3 NLLP Services

In the following main part of this article we describe many of the Natural Legal Language Processing services currently under development in our project: Term Extraction (Section 3.1), Lexical Resources (Section 3.2), Named Entity Recognition (Section 3.3), Concept Extraction (Section 3.4), Word Sense Disambiguation (Section 3.5), Temporal Expression Analysis (Section 3.6), Legal Reference Resolution (Section 3.7), Text Structure Recognition (Section 3.8), Text Summarisation (Section 3.9), Machine Translation (Section 3.10), Legal Knowledge Graph Population (Section 3.11), Semantic Similarity (Section 3.12) and Question Answering (Section 3.13). This set of services is heterogeneous: some of the services make use of other services, some services extract or annotate information (e. g., NER or Temporal Expression Analysis), while others operate on full documents (e. g., summarisation or machine translation), yet others provide a user interface (e. g., QA).

3.1 Term Extraction

To enable the creation of a taxonomy for a certain use case, domain or company, we use the cloud-based Tilde Terminology term extraction service[1]. It extracts terms from different corpora following the methodology by Pinnis et al. (2012). As a result, the platform creates a SKOS vocabulary containing terms, contexts and references to their source documents. Each term comes with a ranking score to describe the terms specificity in the source corpora compared to a general language corpus. The score is calculated based on TF-IDF (Spärck Jones, 1972) and co-occurrence statistics for multi-word terms (Pinnis et al., 2012). Once the term extraction workflow has been triggered, a corresponding online platform takes over. The workflow starts with plain text extraction from different file formats, then all plain-text documents are annotated, and a single collection of terms is created. As multiple surface forms of the same term may appear in the text, term normalisation is performed. This term collection is the first step towards, initially, creating or, later on, enriching the Legal Knowledge Graph. The collection can be used for creating hierarchical taxonomies augmented with multilingual information and linked to other knowledge bases.

3.2 Lexical Resources for the Legal Domain

An essential aspect of the LKG is its capability to be easily adaptable across domains and sectors. It is based on both domain-dependent and domain-independent vocabularies, which will be accessible through a common RDF graph. The domain-dependent vocabularies account for particular terminologies coming from the legal sector and our use case domains (e. g., EuroTermBank[2]). The domain-independent vocabularies are taken from families of monolingual, bilingual and multilingual dictionaries published by one of our project partners, such as Global, Password, and Random House.[3] They contain various cross-lingual links for the five languages served by our platform (Dutch, English, German, Italian, Spanish). Besides their overall coverage of solely domain-independent vocabularies, they contain information on words and phrases that include also or only domain-dependent meanings (e. g., court for the former, lawyer for the latter). The motivation of relying on domain-independent dictionary data for the LKG is thus twofold: first, they provide a common substrate across domains that facilitates traversing semantically annotated documents coming from different specialised domains (e. g., Legal or Oil & Gas); second, they support certain NLP functionalities such as Word Sense Disambiguation and Word Sense Induction by providing a common catalogue of word senses. The data is being remodeled in RDF according to the Ontolex Lemon Lexicography Module Specification[4] and is accessed by the platform via a RESTful API. The LKG has a common core part (terminologies, sets of annotated legal corpora), but can be expanded to accommodate the necessities of particular use cases (e. g., to store private contracts).

[1] https://term.tilde.com

[2] http://www.eurotermbank.com

[3] https://www.lexicala.com

[4] https://jogracia.github.io/ontolex-lexicog/

3.3 Named Entity Recognition

The service for named entity recognition (NER) includes the elaboration of corresponding semantic classes and the preparation of a German language data set. Several state of the art models were trained, i. e., Conditional Random Fields (CRFs) and bidirectional Long-Short Term Memory Networks (BiLSTMs), and evaluated (Finkel et al., 2005; Faruqui and Padó, 2010; Benikova et al., 2014, 2015; Huang et al., 2015; Lample et al., 2016; Riedl and Padó, 2018, etc.). For training and evaluating the system we used a data set of German court decisions that was manually annotated with seven coarse-grained and 19 fine-grained classes: names and citations of people (person, judge, lawyer), location (country, city, street, area), organisation (organisation, company, institution, court, brand), legal norm (law, legal regulation, European legal norm), case-by-case regulation (regulation, contract), case law, and legal literature. The data set consists of approximately 67,000 sentences and around 54,000 annotated entities. For the experiment, two tools for sequence labeling were chosen. These are sklearn-crfsuite (CRFs)[5] and UKPLab-BiLSTM (BiLSTMs)[6] (Reimers and Gurevych, 2017). Three different models and two classifications each are developed for each of these model families (19 and seven classes, respectively). For CRFs these are (1) CRF-F with features, (2) CRF-FG with features and gazetteers, (3) CRF-FGL with features, gazetteers, and the lookup table. For BiLSTMs we used (1) BiLSTM-CRF without character embeddings, (2) BiLSTM-CRF+, and (3) BiLSTM-CNN-CRF with character embeddings generated by BiLSTM and by a CNN. In order to reliably estimate the performance of the models, we use stratified 10-fold cross-validation, which prevents overfitting during training. The stratification guarantees that the semantic classes are equally frequent in the test set relative to the size of the training set, which avoids measurement errors in the case of unbalanced data. The results were measured with precision, recall, and F_1-measure. The BiLSTM models performed better compared to CRF (see Table 1), the F_1 values were between 93.75–95.46 % for the fine-grained classes and between 94.68–95.95 % for the coarse-grained classes. By contrast, the CRF models reached 93.05–93.23 % and 93.11–93.22 %. Overall, the CRF models achieved about 1–10 % lower scores per class than the BiLSTMs. The models provide the best results in the fine-grained classes of judges, courts and laws; their F_1 values were 95 %. Performance was an F_1 value over 90 % in the classes countries, institutions, case laws and legal literature. The recognition of the classes persons, lawyers, cities, companies, legal regulations, European legal norms and contracts varied from 84 % to 93 %. In contrast, the values in the classes streets, landscapes, organisations and regulations were the lowest and amounted to 69–80 % with the CRF models and to 72–83 % with BiLSTM. The worst result was observed in the class brands. With CRF, a maximum F_1 value of 69.61 % was reached and with BiLSTM a maximum F_1 value of 79.17 %. The current NER tool is a working prototype. It already provides named entities locally, but is still being evaluated further. As of now, the service is available for German texts, but it can be easily adapted to other languages.

3.4 Concept Extraction

The LKG contains among its nodes entities from controlled vocabularies. These are typically expressed as SKOS concepts, which permits assigning to them multiple labels, i. e., various surface forms, in multiple languages. Furthermore, one can define relations between instances of concepts, such as hypernymy, to create a taxonomy. Taxonomies become useful when their concepts can be identified in documents, a process called Concept Extraction. A simple example would be taking the sentence “The tenants must pay the heating costs by themselves”, and identifying the presence of the concepts “tenant” and “heating costs”. If these are known to be instances of *Contractual parties* and *Energy costs*, respectively, a search for “energy costs” would point the user to this sentence. Thus, once concept extraction is performed, links between documents and elements of controlled vocabularies in the LKG can be established. While these relations are rather simple, they are the first step for enriching text fragments with knowledge from the LKG, as well as to enable further algorithms for the (semi-)automatic extension of the LKG. Importantly, the inclusion of labels in many languages allows linking of

[5] https://sklearn-crfsuite.readthedocs.io/en/latest/

[6] https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf

Table 1: F_1 values of the CRF and BiLSTM models for the coarse-grained classes.

Class	**CRFs**			**BiLSTMs**		
	F	**FG**	**FGL**	**CRF**	**CRF+**	**CNN-CRF**
Person	91.74	**92.20**	92.16	94.74	**95.41**	95.12
Location	89.26	89.45	**90.18**	91.68	**93.31**	92.57
Organisation	90.87	90.99	**91.11**	91.37	92.87	**93.21**
Legal norm	95.67	95.77	**95.86**	96.77	**97.98**	97.79
Case-by-case regulation	86.94	**86.96**	86.39	85.43	**90.61**	90.43
Case law	93.23	**93.25**	93.08	96.56	**96.99**	96.78
Legal literature	91.92	92.06	**92.11**	93.84	**94.42**	94.02
Total	93.11	**93.22**	**93.22**	94.68	**95.95**	95.79

documents in different languages, combining the knowledge derived from them, as well as multilingual search and recommendation. The Concept Extraction service works in as many languages as the taxonomies have labels in, and thus we can leverage multinational efforts for creating multilingual taxonomies such as EUROVOC[7] or UNBIS[8]. Furthermore, in the case where documents are in English, Spanish, Dutch, German, French, Italian, Czech or Slovak languages, additional linguistic processing increases the recall. The service can be used for production. It is available in most European languages.

3.5 Word Sense Disambiguation

To enable the use of incomplete KGs for automatic text annotations, we introduce a robust method for discriminating word senses using thesaurus information like hypernyms, synonyms, types/classes, which is contained in the KG. The method uses collocations to induce word senses and to discriminate the thesaurus sense from others. Its main novelty is using thesaurus information already at the stage of sense induction. The given KG enables us to cast the task to a binary scenario, namely telling apart the KG sense from all the others. This method does not require all possible senses of a word to be contained in the KG, which makes it especially useful in a production environment, where usually only incomplete KGs are available. We take as input a corpus, thesaurus information, and a concept from the KG, one of whose labels is found throughout the corpus (the target label). We want to distinguish, for each document in the data set, whether the target label is used in the thesaurus sense or not.[9] Thus, the end result is a partition of the corpus into two disjoint collections: “this” and “other”. The collection “this” contains the documents that feature the target label in the thesaurus sense, the collection “other” contains any other sense which does not match the domain captured in the thesaurus, which can be more than one. The experiments were conducted on two data sets[10] created specifically for this task: Cocktails and MeSH (Revenko and Mireles, 2017) (Table 3). This service is used for any kind of entity linking, especially after NER. This is done to correctly identify which named entities are indeed within the vocabulary scope of the LKG. The service is a working prototype. It is language agnostic, i. e., works for any language as long as the text can be tokenised correctly.

Table 2: Cocktails WSID accuracy scores

	Macro average	Micro average
Our Method	0.841	0.896
Baseline	0.725	0.737

Table 3: MeSH WSID accuracy scores

	Macro average	Micro average
Our Method	0.723	0.739
Baseline	0.680	0.735

3.6 Temporal Expression Analysis

Documents from the legal domain contain a multitude of temporal expressions that can be analysed, normalised (i. e., semantically interpreted) and further exploited for document and information mining purposes. We implemented a prototype for the analysis of time expressions in German-language legal documents, especially court decisions and legislative texts. Temporal expressions

[7] https://publications.europa.eu/en/web/eu-vocabularies/

[8] http://metadata.un.org/?lang=en

[9] Without loss of completeness we consider only the case when the target label is used in the same sense in all occurrences in a document. One can, of course, consider the context of every occurrence of the target label as a separate document, therefore extending the method to disambiguate every single occurrence of the target label.

[10] https://github.com/artreven/thesaural_wsi

Table 4: Comparison of the results of the original version of HeidelTime (HT) with the modified (HT nV) on the evaluation corpus. The last line indicates the improvement.

	strict			partial			strict+value				partial+value			
	P	R	F1	P	R	F1	P	R	F1	Acc	P	R	F1	Acc
HT	86.8	86.0	86.4	89.5	88.1	88.8	86.1	85.3	85.7	99.2	88.4	87.1	87.8	98.9
HTnV	94.9	92.0	93.5	96.6	93.5	95.0	94.0	91.1	92.5	99.0	95.3	92.3	93.8	98.7
+	8.2	6.1	7.1	7.1	5.4	6.2	7.9	5.8	6.9	0.2	6.9	5.1	6.0	0.2

include dates, e. g., “1. Januar 2000” (1st January 2000), durations, e. g., “fünf Kalenderjahre” (five calendar years) and repeating time intervals, e. g., “jeden Monat” (every month). Such expressions should not only be identified, but also normalised by translating them into a standardised ISO format. Since no suitable data set existed, a text collection was prepared and annotated with temporal expressions using the TimeML standard. Previously, the automatic identification of temporal expressions (temporal tagging) has been mainly focused on English and domains such as news and narrative texts. Research showed that this task is domain- and language-sensitive, i. e., systems have to be adapted to the specific domain or language to ensure consistent performance. We can confirm this observation: during the annotation of the corpus deficits had become apparent, which concerned not only the annotation guidelines, but also the performance of the rule-based temporal tagger HeidelTime (Strötgen and Gertz, 2010), which was subsequently extended. One of the specifics of the domain are references to other legal texts which contain (alleged) dates (“Richtlinie 2008 / 96 /EG”, “Directive 2008 / 96 /EG”). Other peculiarities of the domain and/or language are the frequent use of compounds such as “Kalenderjahr” (calendar year), “Fälligkeitsmonat” (due month) or “Bankarbeitstag” (banking day), generic use of temporal expressions such as “jeweils zum 1. Januar” (1st January of each year) and event-anchored temporal expressions “Tag der Verkündigung” (proclamation day). Based on our new annotated corpus, HeidelTime was adapted to the domain. The evaluation showed that the adjustments made to HeidelTime significantly improved its performance (Table 4. Particularly noteworthy is the recall with an increase of approx. 10 percentage points. Normalisation remains problematic, which is also due to generic or event-based uses of temporal expressions as well as legal references.

3.7 Legal Reference Resolution

References to other documents are another class of expressions used in abundance in documents from the legal domain. The considered problem consists in recognizing and, ideally resolving, such references. Usually, editors attempt to be consistent and follow patterns to reference other documents. The developed methodology, currently implemented as a language-agnostic prototype, follows this assumption and attempts to discover patterns used in a semi-automatic manner. The discovered patterns are constructed from features that are either individual tokens (e. g., “Decision”, “EU”, etc.) or processed features (e. g., “DIGITS” as a placeholder for numbers). We use a seed collection of documents, where references have been manually annotated and resolved. For each reference we collect the tokens preceding the reference and analyse the features present in these tokens. Next we aggregate the most common combinations of features – these form “patterns”. Example of a pattern could be {“EU”, “Decision”, “DIGITS/DIGITS”} or {“the”, “data”, “subject”}. The second pattern is an example of a common combinations of tokens in text and does not necessarily indicate a reference. To filter out such irrelevant patterns from the seed documents we extract the strings containing the candidate pattern, but not containing a reference. If several such strings are found the pattern is discarded. In the next step the most common undiscarded patterns are presented to the user who can accept several patterns that are later used to discover new references, enabling the recursive improvement of patterns.

3.8 Text Structure Recognition

Knowing the structure of a document can drastically improve the performance of the analysis services applied to the text, as specialised fine-grained models and focused approaches can be integrated. In the legal domain, it is important to determine the structure of a document to identify sections, subsections, paragraphs, etc. correctly because many legal references also contain this type of information, ideally enabling automatically linking to the correct part of the text, instead of the whole document. Robust text structure recognition is still an open research question. Many approaches have been suggested in different fields, such as Optical Layout Recognition (OLR) for unstructured documents or markup-based approaches for structured documents. We try to cover both in our prototype.

Unstructured documents do not contain any structure information whatsoever, they are often provided in plain text. To process (plain) text, we start by applying a pattern based approach (regular expressions) that allows the identification of the title, headings and running text or paragraphs. After that, we apply topic detection to all those parts (both titles and running texts) in order to cluster sections with related topics.

Structured documents include structural information (e. g., markup). We consider two ways of analysing them: (1) defining a mapping between the elements important and relevant for the use cases addressed by our platform and the structural elements of the documents; and (2) extracting the plain text from the document and then applying the techniques for unstructured documents.

3.9 Text Summarisation

To enable our users to work with legal documents more efficiently, we experiment with summarisation services (Allahyari et al., 2017). While extractive summarisation has been popular in the past, the progress in neural technologies has renewed the interest in abstractive summarisation, i. e., generating new sentences that capture a document's meaning. This approach requires highly complex models and a lot of training data. In the absence of labeled training data, extractive methods are often used as the basis for abstractive methods, by assigning relevance scores to sentences in an unsupervised way. Abstractive summarisation is often augmented using word embeddings (Mikolov et al., 2013; Pennington et al., 2014) that provide a shared semantic space for those strongly related sentences that do no share the same but similar or related words. We develop two methods. The first tool is based on TF-IDF (Neto et al., 2000). This is a popular baseline as it is easy to implement, unsupervised, and language independent. Instead of using bag-of-words sentence representations, our approach tries to improve on this, by analysing the texts first. Searching the embedding space of all words used in the text, we cluster similar words so that morphological variants of a word like "tree" and "trees" or "eat" and "eating", but also synonyms like "fast" and "rapid" are considered as belonging to the same cluster. Based on these groupings we encode all documents and then calculate the weights for the sentences using TF-IDF. The second tool is based on the concept of centroids (Rossiello et al., 2017; Ghalandari, 2017) and benefits from the composability of word embeddings. Initially, keywords and concepts are extracted from the document. By composing their embeddings, the centroid is created, which represents the document's condensed meaningful information. It is then projected into the embedding space together with all sentence embeddings. Sentences receive relevance scores depending on their distance to the centroid in the embedding space. To avoid redundancy in the summary, sentences that are too similar to the ones already added to the summary are not used. Both tools can be used for multiple languages, single and multi-document summarisation. The current version of the centroid text summarisation is a working prototype. It already provides extractive summaries for single and multiple documents, but is still being tested and optimised. The service is only available for English but can be adapted to other languages by training new embeddings.

3.10 Machine Translation

To enable multilingualism and cross-lingual extraction, linking and search, we use the Machine Translation (MT) service Tilde MT[11]. In order to populate and process the Legal Knowledge Graph in a multilingual way, custom Neural Machine Translation (NMT) systems were trained for selected language pairs – English ↔ Spanish, English ↔ German, and English ↔ Dutch. In-domain business case specific legal data was gathered and processed prior to training the NMT systems on a mix of broad-domain and in-domain data to be able to translate both in-domain and out-of-domain texts. Marian was used for training (Junczys-Dowmunt et al., 2018). The translation service provides support for a runtime scenario as well as for asynchronous processes, i. e., support-

[11] https://tilde.com/mt

Table 5: Evaluation results of NMT systems

Language pair	Sentence pairs	BLEU	
		to EN	from EN
English ↔ Dutch	41,639,299	43.54	34.12
English ↔ Spanish	81,176,632	32.52	38.36
English ↔ German	24,768,821	38.73	44.73

ing background data curation processes. The synchronous translation service endpoint serves translation functionality for texts and documents annotated with the Natural Language Processing Interchange Format ontology (NIF) (Hellmann et al., 2013). The systems were automatically evaluated using BLEU (Papineni et al., 2002) on held-out evaluation sets. The sets were created from the in-domain parts of the parallel corpora used for training of the NMT systems. Table 5 contains statistics of the training data and the automatic evaluation results of the NMT systems.

3.11 Legal Knowledge Graph Population

For the definition of our Knowledge Graph, we benefit from predefined vocabularies such as EUROVOC. However, their knowledge is limited to that intended by their creators, and their level of specificity and focus will, in general, not match the ones required for an application. One possible option of extending existing knowledge resources is the large-scale analysis of documents in which entities contained in the knowledge resources have been identified, and to identify as well as to extract new relations, claims, or facts, explicitly mentioned in the documents. This approach mimics the process in which a human reads and understands documents. In our project we follow distant supervision (Ren et al., 2017). It takes a text corpus as input and 1) identifies sentences containing entity pairs for which relations are known, 2) uses machine learning to derive a statistical classifier to recognize these examples, and 3) applies this classifier to sentences that, by virtue of the classes of entities they contain, could also include an instance of a relation. The result is a list of sentences which have been annotated as containing a given relation. The relations included so far in this first experimental deployment, are of the type *person is located in location* and *location is contained in location*. They will be expanded with domain specific relations such as *activity requires permit* or *permit was issued on date*. So far, such relations can be recognized in English language texts, but training for German, Spanish and Dutch, using the same distant supervision approach is possible due to the multilingual nature of general purpose corpora and knowledge graphs (e. g., DBPedia).

3.12 Semantic Similarity

Using the services mentioned above, documents in the LKG are annotated to semantically describe their content and provenance. This added extra knowledge is useful for several applications, such as search, question answering, classification and recommendations, all of which rely on a notion of document similarity. Many such notions exist, and they are usually encoded in a function s that assigns, to every pair of documents, a number between 0 and 1, with 1 denoting the documents being identical (Gomaa and Fahmy, 2013). We use a hybrid type of similarity measure. First, the text entailed by the document, such as the resolution of temporal or geographical references, is performed. Second, similarity itself is computed using a linear combination of text-based and knowledge-based similarities. The former are encoded by cosine-similarity of TF-IDF vectors (of the documents or their translations), and the latter by the overlap (as measured by Jaccard coefficient of entities that the two documents either mention directly, or are linked in the LKG to mentioned ones. The overlaps are weighted depending on how far away in the LKG the entities mentioned in the document are, and the weight coefficients are determined, along with the coefficients of the linear combination, by training a linear regression classifier (Bär et al., 2012). This approach allows us to detect similarity between documents even if they have only few entities in common, by considering the knowledge about these entities. Additionally, by comparing mentions of entities instead of their surface forms, the multilingual nature of the LKG is exploited. The knowledge-based component of the similarity computation is language agnostic, while the text-based depends only on basic NLP tools (e. g., stemming, stop-word removal) which are available for English, German, Spanish and Dutch, among others. In order to compare documents in two different languages, machine translation between them, or to a third language must be available. The semantic similarity service is a prototype, requiring further testing and refining.

3.13 Question Answering

The Question Answering (QA) service accepts a natural language question and responds with an answer, extracted from a document in a given corpus. The end-to-end system consists of three components: 1) The Query Formulation module transforms a question into a query, which can be expanded using a domain specific vocabulary from the LKG. The query is then processed through an indexer to obtain matching documents from the corresponding corpora. 2) The Answer Generation module extracts potential answers from the retrieved documents from the LKG. 3) The Answer Selection module identifies the best answer based on various criteria such as local structure of the text and global interaction between each pair of words based on specific layers of the model. The QA service is the central component of two of our use cases. In these, a user asks a natural language question along with additional information such as an appropriate jurisdiction. The question is meant to trigger a query on a set of documents related to the specific jurisdiction. These additional parameters influence the search in the Answer Selection module by determining which subset of the documents should be used. With the help of the Machine Translation service, the QA service is able to return answers in languages different from the documents' language. The benefit from the service would be the time reduction in search for a relevant article in the legislation. For a question such as "How long is paternity leave?", the system returns relevant paragraphs and articles from the Labor Law that can be further processed by the lawyer. The QA service works for English and can be retrained and adopted to other languages.

4 Orchestration of Individual Services

Our technology platform is based on microservices, which provide multiple advantages, especially in a collaborative project with a distributed set of partners from academia and industry, including use case partners. Microservices are small and autonomous and can be developed more efficiently than a monolithic, integrated system. In addition, the development and deployment of microservices can, to a very large extent, be automated, also facilitating the monitoring of individual services. A crucial advantage is concerned with the scalability of systems based on microservices, which is a lot easier than scaling monolithic systems. The communication between different services is executed over HTTP, the interfaces are documented using the OpenAPI specification. For the deployment of the containerised microservices we use OpenShift; alternative technologies such as, among others, Kubernetes, could also be used.

In our project we conceptualise the specific requirements of the different use cases as content curation workflows (Schneider and Rehm, 2018b; Bourgonje et al., 2016a,b; Rehm et al., 2018). Workflows are defined as the execution of specific services to perform the processing of one or more documents under the umbrella of a certain task or use case. The specification of a workflow includes its input and output as well as the functionality it is supposed to perform: annotate or enrich a document, add a document to the knowledge base, search for information, etc. The project offers compliance-related features and functionalities through common services and data sets included in the LKG. Workflows make use of these services to implement the required functionality. The content curation workflows for the different use cases that we prototypically implement in the project have been defined as follows. We performed a systematic analysis of the microservices, developed in parallel, and matched them with the required functionalities for each use case. First, we determine the principal elements involved in each use case, i. e., the services, input and output. Second, we define the order in which the services have to be executed. Third, we identify the shared components in the different workflows. Currently we have defined five different workflows. Two are defined for the acquisition and population of the LKG with new information, the other three are defined to address the requirements of each use case (see Figures 1 to 3 in the appendix). We currently work with two alternative workflow management implementations that take care of the orchestration of services. The first one is based on Camunda BPM,[12] including a logic layer that manages the various processes (including NIF annotations). The second approach is based on RabbitMQ,[13] an open-source message broker, on top of which we developed our own solution for parallelising processing steps and improving the performance of the overall orchestration. In both cases, the main concept of the service orchestration is focused on the use of queuing systems, so that most of the processes could be executed in parallel and either synchronous or asynchronously.

[12] https://camunda.com

[13] https://www.rabbitmq.com

5 Related Work

There are several systems, platforms and approaches that are related to the technology platform, which is under development in the project LYNX. In the wider area of legal document processing, technologies from several fields are relevant, among others, knowledge technologies, citation analysis, argument mining, reasoning and information retrieval. A literature overview can be found in (Schneider and Rehm, 2018a) and (Agnoloni and Venturi, 2018).

Commercial Systems and Services – The LexisNexis[14] system is the market leader in the legal domain; it offers services, such as legal research, practical guidance, company research and media-monitoring as well as compliance and due diligence. WestLaw is an online service that allows legal professionals to find and consult relevant legal information.[15] One of its goals is to enable professionals to put together a strong argument. There are also smaller companies that offer legal research solutions and analytic environments, such as RavelLax,[16], or Lereto[17]. A commercial search engine for legal documents, iSearch, is a service offered by LegitQuest.[18] The Casetext CARA Research Suite allows uploading a brief and then retrieving, based on its contents, useful case law.[19] There is also a growing number of startup companies active in the legal domain.

Research Prototypes – Most of the documented research prototypes were developed in the 1990s under the umbrella of Computer Assisted Legal Research (CALR) (Span, 1994). In the following we briefly review several of these systems, which usually focus on one very specific feature or functionality. One example is the open source software for the analysis and visualisation of networks of Dutch case law (van Kuppevelt and van Dijck, 2017). This technology determines relevant precedents (analysing the citation network of case law), compares them with those identified in the literature, and determines clusters of related cases. A similar prototype is described by (Agnoloni et al., 2017). (Gifford, 2017) propose a search engine for legal documents where arguments are extracted from appellate cases and are accessible either through selecting nodes in a litigation issue ontology or through relational keyword search. Lucem (Bhullar et al., 2016) is a system that tries to mirror the way lawyers approach legal research, developing visualisations that provide lawyers with an additional tool to approach their research results. Eunomos is a prototype that semi-automates the construction and analysis of knowledge (Boella et al., 2012).

6 Summary and Future Work

This article presents the technology platform currently under development in the project LYNX, focusing upon processing services. These serve two main purposes: 1) to extract semantic information from a large and heterogeneous set of documents to ingest the extracted information into the Legal Knowledge Graph; 2) to extract semantic information from documents that users of the platform work with. In addition to the semantic extraction of information and knowledge, we provide services for the processing and curation of whole documents (summarisation, translation) with the goal of mapping extracted terms and concepts to the LKG, and services that aim at accessing the LKG (question answering). We currently experiment with two different curation workflow managers to make specific sets of services available for specific use cases. Future work includes completing development work on the services, adapting the services to all languages required in the project's use cases, implementing the prototype applications and developing the freely accessible web interface of the platform.

Acknowledgments

This work has been partially funded by the project LYNX, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 780602. For more information please see http://www.lynx-project.eu.

References

Tommaso Agnoloni, Lorenzo Bacci, Ginevra Peruginelli, Marc van Opijnen, Jos van den Oever, Monica Palmirani, Luca Cervone, Octavian Bujor, Arantxa Arsuaga Lecuona, Alberto Boada García, Luigi Di Caro, and Giovanni Siragusa. 2017. Linking european case law: BO-ECLI parser, an open

[14] https://www.lexisnexis.com

[15] http://legalsolutions.thomsonreuters.com/law-products/westlaw-legal-research/

[16] http://ravellaw.com

[17] https://www.lereto.at

[18] https://www.legitquest.com

[19] https://casetext.com

framework for the automatic extraction of legal links. In (Wyner and Casini, 2017), pages 113–118.

Tommaso Agnoloni and Giulia Venturi. 2018. Semantic processing of legal texts. In Jacqueline Visconti, editor, *Handbook of Communication in the Legal Sphere*, pages 109–138. De Gruyter, Berlin, Boston.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A brief Survey. *arXiv preprint arXiv:1707.02268*.

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 435–440. Association for Computational Linguistics.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pad. 2014. GermEval 2014 Named Entity Recognition Shared Task: Companion Paper. In *Proceedings of the KONVENS GermEval workshop*, pages 104–112, Hildesheim, Germany.

Darina Benikova, Seid Muhie Yimam, Prabhakaran Santhanam, and Chris Biemann. 2015. Germaner: Free open german named entity recognition tool. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2015, University of Duisburg-Essen, Germany, 30th September - 2nd October 2015*, pages 31–38.

Jagjoth Bhullar, Nathan Lam, Kenneth Pham, Adithya Prabhakaran, and Albert Joseph Santillano. 2016. *Lucem: A Legal Research Tool*, 63. Computer Engineering Senior Theses.

G. Boella, L. di Caro, L. Humphreys, L. Robaldo, and L. van der Torre. 2012. Nlp challenges for eunomos, a tool to build and manage legal knowledge.

Peter Bourgonje, Julian Moreno-Schneider, Jan Nehring, Georg Rehm, Felix Sasaki, and Ankit Srivastava. 2016a. Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer. In *The Semantic Web*, number 9989 in Lecture Notes in Computer Science, pages 65–68. Springer. ESWC 2016 Satellite Events. Heraklion, Crete, Greece, May 29 – June 2, 2016 Revised Selected Papers.

Peter Bourgonje, Julián Moreno Schneider, Georg Rehm, and Felix Sasaki. 2016b. Processing Document Collections to Automatically Extract Linked Data: Semantic Storytelling Technologies for Smart Curation Workflows. In *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, pages 13–16, Edinburgh, UK. The Association for Computational Linguistics.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Semantic Approaches in Natural Language Processing: Proceedings of the 10th Conference on Natural Language Processing, KONVENS 2010, September 6-8, 2010, Saarland University, Saarbrücken, Germany*, pages 129–133. universaar, Universitätsverlag des Saarlandes / Saarland University Press / Presses universitaires de la Sarre.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 363–370. The Association for Computer Linguistics.

Demian Gholipour Ghalandari. 2017. Revisiting the Centroid-based Method: A Strong Baseline for Multi-Document Summarization. *arXiv preprint arXiv:1708.07690*.

Matthew Gifford. 2017. Lexridelaw: an argument based legal search engine. In *ICAIL '17*.

Wael H Gomaa and Aly A Fahmy. 2013. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using Linked Data. In *12th International Semantic Web Conference*. 21-25 October.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F T Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. *arXiv preprint arXiv:1804.00344*.

Dafne van Kuppevelt and Gijs van Dijck. 2017. Answering legal research questions about dutch case law with network analysis and visualization. In (Wyner and Casini, 2017), pages 95–100.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Joel Larocca Neto, Alexandre D Santos, Celso AA Kaestner, Neto Alexandre, D Santos, et al. 2000. Document Clustering and Text Summarization.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, and Tatiana Gornostay. 2012. Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012)*, June 2012, pages 193–208, Madrid, Spain.

Georg Rehm, Julián Moreno Schneider, Peter Bourgonje, Ankit Srivastava, Rolf Fricke, Jan Thomsen, Jing He, Joachim Quantz, Armin Berger, Luca König, Sören Räuchle, Jens Gerth, and David Wabnitz. 2018. Different Types of Automated and Semi-Automated Semantic Storytelling: Curation Technologies for Different Sectors. In *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 232–247, Cham, Switzerland. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348. Association for Computational Linguistics.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1015–1024. International World Wide Web Conferences Steering Committee.

Artem Revenko and Víctor Mireles. 2017. Discrimination of word senses with hypernyms. In *Proceedings of the 5th International Workshop on Linked Data for Information Extraction co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017.*, volume 1946 of *CEUR Workshop Proceedings*, pages 50–61. CEUR-WS.org.

Martin Riedl and Sebastian Padó. 2018. A named entity recognition shootout for german. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 120–125. Association for Computational Linguistics.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based Text Summarization through Compositionality of Word Embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21.

Julian Moreno Schneider and Georg Rehm. 2018a. Curation Technologies for the Construction and Utilisation of Legal Knowledge Graphs. In *Proceedings of the LREC 2018 Workshop on Language Resources and Technologies for the Legal Knowledge Graph*, pages 23–29, Miyazaki, Japan. 12 May 2018.

Julian Moreno Schneider and Georg Rehm. 2018b. Towards a Workflow Manager for Curation Technologies in the Legal Domain. In *Proceedings of the LREC 2018 Workshop on Language Resources and Technologies for the Legal Knowledge Graph*, pages 30–35, Miyazaki, Japan. 12 May 2018.

Georges Span. 1994. Lites: An intelligent tutoring system shell for legal education. *International Review of Law, Computers & Technology*, 8(1):103–113.

Karen Spärck Jones. 1972. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, 28:11—-21.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 321–324, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adam Z. Wyner and Giovanni Casini, editors. 2017. *Legal Knowledge and Information Systems - JURIX 2017: The Thirtieth Annual Conference, Luxembourg, 13-15 December 2017*, volume 302 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.

A Appendix: Examples of Use-Case-specific Curation Workflows

The following three figures are additional material with regard to Section 4. They provide illustrative examples of use-case-specific processing service and curation workflows.

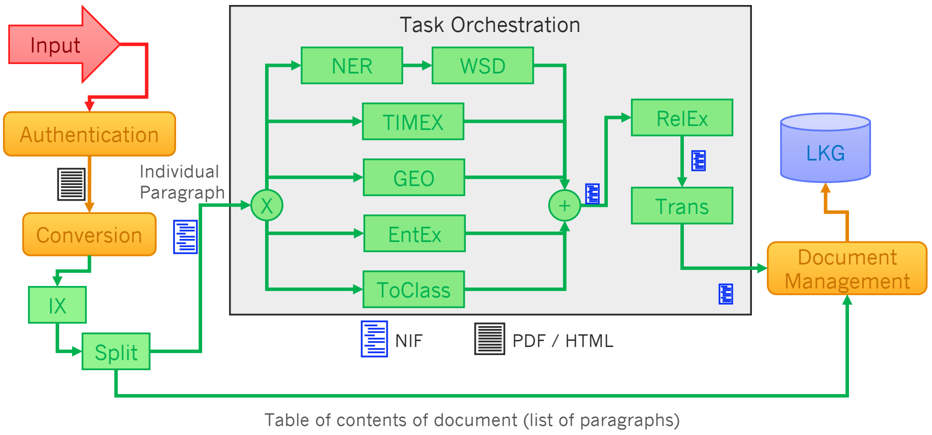

Figure 1: Legal Knowledge Graph population workflow

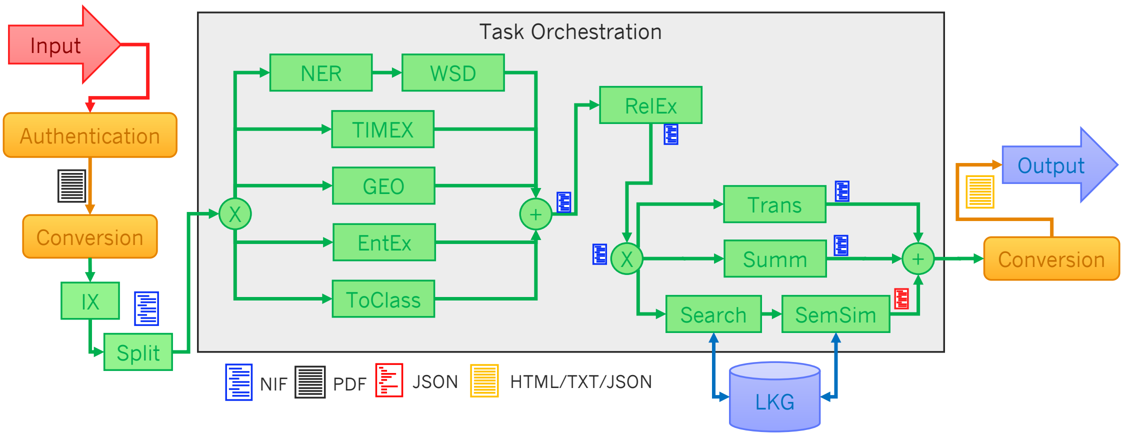

Figure 2: Workflow of the Contract Analysis use case

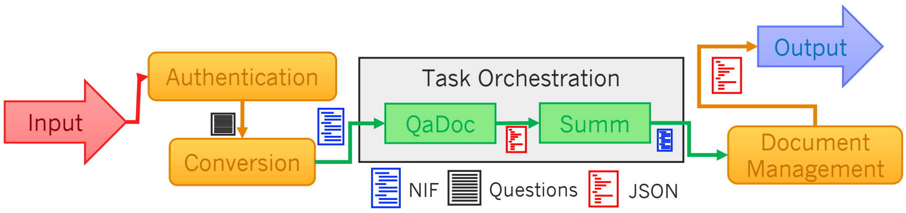

Figure 3: Workflow of the Labour Law use case