September 2019

Automatic Annotation Service: Utilizing a Named Entity Linking Tool in Legal Domain

Minna TAMPER [a], Arttu OKSANEN [c], Jouni TUOMINEN [a,b], Aki HIETANEN [d], and Eero HYVÖNEN [a,b]

[a] *Aalto University, Finland*

[b] *HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland*

[c] *Edita Publishing Oy, Finland*

[d] *Ministry of Justice, Finland*

Abstract. Texts referencing court decisions, statutes, and EU directives can be difficult to understand without context. It can be time consuming and expensive to find related statutes or to learn about context specific terminology. As a solution, we utilized an automatic annotation tool, Nelli, for extracting information and tailored it to a service that can automatically annotate legal documents to provide context to the readers. The service can identify and link named entities and references to legal texts to corresponding vocabularies and data sources by combining statistics- and rule-based named entity recognition with named entity linking. The results provide users with enhanced reading experience with contextual information and possibility to access related materials such as statutes and court decisions.

Keywords. Automatic annotation service, legal texts, named entity recognition, named entity linking

1. Introduction

Texts referencing legislative decisions, statutes, and EU directives can be difficult to understand without context. It can be time consuming and expensive to find correct versions of statutes, to learn about context specific terminology, or to find documents related to the same topic. To understand and interpret legal texts correctly, it is often important to get acquainted with other related contextual material.

The research hypothesis of this paper is that by annotating and linking legal texts it is possible to assist readers to understand the text and context by offering information about legislation, context, and terminology. To achieve this, we utilized an automatic annotation tool, Nelli [11], for identifying domain specific information and tailored it to a service[1] that visualizes automatically generated

[1] A demonstrator that is under development is available at `http://nlp.ldf.fi/nelli/`.

annotations and provides context through links to readers. The service can identify and link named entities and references to legal texts to corresponding vocabularies and data sources by combining statistics- and rule-based named entity recognition (NER) with named entity linking (NEL).

In the following sections, we will discuss the approach in more detail, starting with data in Section 2 and methods in Section 3. In Section 4, we present the application, and finally, in Section 5, we conclude with related work and discussion.

2. Data

Semantic Finlex[2] [8] is a web service that hosts the Finnish legislation and case law as linked open data. Currently, the data published in Semantic Finlex includes consolidated statutes with version history (approx. 2500 statutes), the original statutes as published in the official journal (approx. 50 000 statutes), the Judgments of the Supreme court (5500), and the Judgments of the Supreme administrative court (7500). In addition the data contains keywords used by the Supreme Court and the Supreme Administrative Court to annotate the court judgments. The judgments are also linked to judges and personnel contributing to the case. The original statutes are also linked to EU law and Finnish government bills. The service includes the legal texts in text, HTML, and XML formats. The documents are written in Finnish and Swedish.

3. Method

In order to automatically annotate the legal texts of Semantic Finlex, the Nelli tool [11] was utilized. Nelli is a combination of NER and NEL tools and disambiguates entities using a scoring scheme where the most popular named entity type, the longest string (Nokia Oyj (company) vs. Nokia (place)), and linked interpretation wins. However, initially Nelli was a command line tool that could be only used for annotating text documents. In order to annotate and provide context to legal texts, the tool was transformed into a web service with a restful API service. The number of input and output parameters was extended to support HTML, XML, and text formats and the output format was changed to JSON that returns the annotated document in the original form and a list of entities. Also, new tools were added in order to recognize more named entity types. The added tools were Regular expression-based named entity identifier[3] and Person-Name-Finder[4]. In addition, an existing tool, LINFER tool[5], was upgraded to identify

[2] `http://data.finlex.fi`

[3] The Regular expression based named entity identifier is a NEL tool that uses numerous regular expressions to identify named entities such as car plates and various registry numbers from text and links them to corresponding knowledge bases.

[4] The Person-Name-Finder service is a tool for identifying references to people with the help of ARPA [4] and a vast Finnish person name ontology, `http://light.onki.fi/henkilonimisto/en/`, based on the open data from the Population Information System of Finland, `https://vrk.fi/en/`.

[5] LINFER is a NER tool that utilizes the results of the Finnish dependency parser [3] and infers named entities by using linguistic information and dependency grammar.

more organizations from the texts. The service is currently only for the Finnish language documents but it is possible to configure Nelli for other languages.

4. Application

On top of the Nelli service a web application was built to visualize the results and to provide context and recommendations to the legal texts. The application form for annotating consists of an input field, input format (e.g., text, XML) selection, toggles for selecting what tools to use, and linking options. The linking options consist of ontologies and vocabularies located in a drop-down menu that have been configured in the ARPA tool [4]. The tool can form n-grams from the given text and linguistically manipulate it (e.g., lemmatize) to match it to the given ontology. Currently, the linking options have been set to link mentions in the text to common Finnish place names (YSO-places[6]), legal terminology (the consolidated vocabulary of Finnish legal terms (draft) [2], the Helsinki Term Bank for Arts and Sciences (TTP), DBpedia), and terms used by EU institutions (EuroVoc[7]), in addition to Semantic Finlex keywords, statutes, and case law. By default, the tools have been configured to identify a maximal number of entities without the linking option. After configuring the application, the user can press the "Annotate" button, and the application annotates the given input and retrieves similar court decisions using the Semantic Finlex case law finder [10]. The resulting annotations with an example text are presented in Fig. 1.

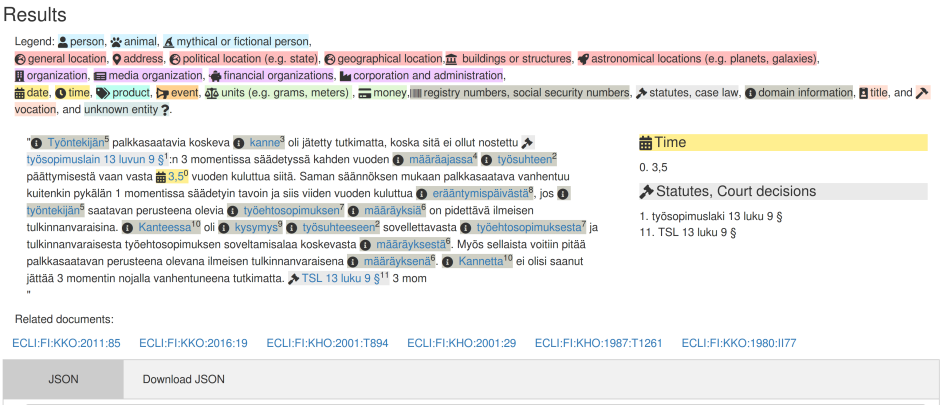

Figure 1. Results of annotating abstract of a case law.

The results are represented under the configuration interface accompanied by a legend that shows all possible named entity types and how they are shown in text (icon and color). Below the legend is the annotated text and on its right side a list of entities found in the text (by type). The linked entities are shown with links and by clicking them a popup appears and shows the description of the

[6] https://finto.fi/yso-paikat/en/
[7] http://eurovoc.europa.eu

given entity. Occasionally, when there is more than one option for an entity, all of them are shown in the popup and the user can select the correct one. In case the application has not found a matching entity, the user can use an autocompletion search field in the popup to query for suitable entities and link the entity manually. Below the text, there is also a list of similar or related documents that have been retrieved for the input text. At the bottom of the page, the JSON response shown that can be viewed or downloaded by clicking the tab.

In this example (Fig. 1), the application has managed to identify a reference to time, statutes, and multiple references to different contextual terms from an abstract of a case law. The statutes and times are identified but not linked whereas the domain information entities have been linked. The linking options were set to link legal terminology (i.e., domain information) to the consolidated vocabulary of Finnish legal terms and to the Helsinki Term Bank for Arts and Sciences (TTP). The Regular expression based named entity identifier can link statutes and case law to Semantic Finlex. However, currently the endpoint doesn't contain all the acronyms and alternative names for the statutes and therefore the linking often fails. Below the text, the application has retrieved six related court decisions. The user can click the links to read the related documents in Semantic Finlex.

The technical infrastructure of the application is based on OpenShift container application platform[8] where the tools are run as individual loosely coupled services. This allows a dynamically scalable configuration, by replicating computationally intensive processes, e.g., dependency parsing[9] and named entity recognition[10], for higher throughput.

5. Related Work and Discussion

The main inspiration of the application has been the contextual reader application CORE [5] that was created to link text into ontologies in real-time to provide related materials and context. This application was initially utilized in the Semantic Finlex portal [8], configured to use content-related ontologies to provide context for the user. However, the tool does not have a powerful disambiguation system when compared with other named entity linking tools, e.g., DBpedia Spotlight[11] [7] and Gate Cloud[12] [6]. For this purpose Nelli was created and with it the first contextual reader in the BiographySampo portal [11]. With Nelli, the entities are not extracted in real time but in a preprocessing phase for more robust semantic disambiguation similarly to [9,1].

The initial demo application using Nelli manages to identify, highlight, and link named entities from the text. The initial annotation accuracy using Nelli was approx. 80% [11] for people and places in biographical texts. The service has been upgraded and the initial results show promise but it still needs a formal evaluation which will be carried out in future. The current version is still under

[8] `https://www.openshift.com`
[9] `https://github.com/SemanticComputing/finnish-dep-parser-docker`
[10] `https://github.com/SemanticComputing/finer-docker`
[11] `https://www.dbpedia-spotlight.org/demo/`
[12] `https://cloud.gate.ac.uk`

development and more work needs to be done in order to achieve the goals of the project, such as creating intelligent search applications that can utilize annotated legal texts. The demo application presents how by annotating documents it is possible to cater information and related documents to provide context to the reader automatically.

Acknowledgments This work is part of the ANOPPI project[13] funded by the the Ministry of Justice in Finland. Thanks to Saara Packalén, Tiina Husso, and Oili Salminen of the Ministry of Justice, and Risto Talo, Jari Linhala, and Sari Korhonen of Edita Publishing Ltd. for collaboration. CSC – IT Center for Science, Finland, provided us with computational resources.

References

[1] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM, 2010.

[2] M. Frosterus, J. Tuominen, and E. Hyvönen. Facilitating re-use of legal data in applications—Finnish law as a linked open data service. In *Proceedings of the 27th International Conference on Legal Knowledge and Information Systems (JURIX 2014)*. IOS Press, December 2014.

[3] K. Haverinen, J. Nyblom, T. Viljanen, V. Laippala, S. Kohonen, A. Missilä, S. Ojala, T. Salakoski, and F. Ginter. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48:493–531, 2014. Open access.

[4] E. Mäkelä. Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In *Proceedings of the ESWC 2014 demonstration track*, pages 424–428. Springer-Verlag, 2014.

[5] E. Mäkelä, T. Lindquist, and E. Hyvönen. CORE – a contextual reader based on linked data. In *Proceedings of Digital Humanities 2016, Krakow, Poland (long papers)*, pages 267–269, 2016.

[6] D. Maynard, I. Roberts, M. A. Greenwood, D. Rout, and K. Bontcheva. A framework for real-time semantic social media analysis. *Journal of Web Semantics*, 44:75–88, 2017.

[7] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM, 2011.

[8] A. Oksanen, J. Tuominen, E. Mäkelä, M. Tamper, A. Hietanen, and E. Hyvönen. Semantic Finlex: Transforming, publishing, and using Finnish legislation and case law as linked open data on the web. In G. Peruginelli and S. Faro, editors, *Knowledge of the Law in the Big Data Age*, volume 317 of *Frontiers in Artificial Intelligence and Applications*, pages 212–228. IOS Press, 2019.

[9] F. Piccinno and P. Ferragina. From TagME to WAT: A New Entity Annotator. In *Proceedings of the first international workshop on Entity recognition & disambiguation*, pages 55–62. ACM, 2014.

[10] S. Sarsa and E. Hyvönen. Searching case law judgements by using other judgements as a query, 2019. Submitted article under evaluation.

[11] M. Tamper, E. Hyvönen, and P. Leskinen. Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICling 2019)*. Springer-Verlag, April 2019. Forthcoming.

[13] `https://seco.cs.aalto.fi/projects/anoppi/en/`