Automatic Annotation Service APPI: Named Entity Linking in Legal Domain

Minna Tamper[1,2][0000−0003−1695−5840], Arttu Oksanen[1,3][0000−0003−2327−6942], Jouni Tuominen[1,2][0000−0003−4789−5676], Aki Hietanen[4], and Eero Hyvönen[1,2][0000−0003−1695−5840]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
http://seco.cs.aalto.fi, firstname.lastname@aalto.fi
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://heldig.fi
[3] Edita Publishing Ltd.
http://www.editapublishing.fi
[4] Ministry of Justice, Finland
http://oikeusministerio.fi, firstname.lastname@om.fi

Abstract. Texts referencing court decisions and statutes 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 a named entity linking tool for extracting information and tailored it into a service, Appi, 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 the possibility to access related materials, such as statutes and court decisions.

Keywords: automatic annotation service · legal texts · named entity linking · linked data

1 Introduction

The research hypothesis of this paper is that by annotating and linking legal texts to knowledge bases it is possible to assist readers to understand the text and context by offering information about legislation, context, and terminology. To understand and interpret legal texts correctly, it is often important to get acquainted with other related contextual material. The linking of texts through similarity or references can aid in finding information. To support end users in close reading and to enable linking of legal texts, we created a service called Appi[5]. It utilizes a named entity linking tool, Nelli [13], for identifying domain specific information and to enable named entity linking of legal texts. As

[5] A demonstrator that is under development is available at http://nlp.ldf.fi/appi.

a result, the Appi service can identify and link named entities, terminology, 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). The end results can be edited in the Appi service's web application and they can be downloaded in JSON format. In this paper, the Appi tool is piloted for Finnish court decisions and legislation.

2 Data

Semantic Finlex[6] [9] 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), Judgments of the Supreme court (5500), and Judgments of the Supreme administrative court (7500). In addition, the data contains keywords for the statutes, and 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 legislation 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 [13] was developed further. Nelli is a combination of NER (FiNER [11,4], LINFER [13]) and NEL (ARPA [5]) tools and it disambiguates entities using a scoring scheme where popularity of the interpretation of the named entity type, the string length, and successful linking are taken into account. 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 restful API service. The support for input formats was extended to 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 recognized entities. Also, new tools were added in order to recognize more named entity types: FinBERT[7] [15], a regular expression-based named entity linker tool called Reksi[8], and Person Name Finder[9].

FinBERT is a Finnish version of the Google's BERT [1] deep transfer learning model. It was added to improve identification of named entities in text by adding a deep learning based method to complement the two rule based NER tools

[6] `http://data.finlex.fi`
[7] `http://turkunlp.org/FinBERT/`
[8] `http://nlp.ldf.fi/reksi`
[9] `http://nlp.ldf.fi/name-finder`

(FiNER, LINFER) to find more named entities that go unnoticed with the rule-based tools. Reksi is a NEL tool that uses numerous regular expressions to identify named entities, such as registry numbers, references to statutes, and case law from the text and links them to corresponding knowledge bases. It utilizes the regularity of the forms of the entities in texts and formats them to find the matching entities from the target ontologies. It was developed to enable better identification of named entities that appear in a form that is easy to identify using regular expressions. These entities are common in legal documents, such as court decisions, where, for example, references are made to earlier court decisions and statutes, and punishments (sentence times and fines) are given. The Person Name Finder service is a tool for identifying references to people by linking the names to the Finnish person name ontology HENKO[10] [14]. The tool was added to improve identification of person names that are mentioned in the texts. In addition, an existing tool, LINFER, was upgraded to identify 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

The Appi web application was built on top of the results of the Nelli service to visualize them and to provide context and recommendations to the legal texts by linking the given text to different ontologies and to other legal texts in the Semantic Finlex dataset. For this purpose, the application form for annotating consists of an input field, input format (e.g., text, XML) selection, toggles for selecting what tools to use in Nelli, and linking options. The linking options consist of a list of ARPA configurations for ontologies and vocabularies located in a drop-down menu. Based on the selected configuration, the ARPA 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 enable linking of mentions in the text to common Finnish place names (YSO places[11]), legal terminology (the consolidated vocabulary of Finnish legal terms (draft) [3], the Helsinki Term Bank for Arts and Sciences[12], DBpedia, and terms used by EU institutions (EuroVoc[13]) in addition to Semantic Finlex keywords, statutes, and case law. The user can retrieve textually similar court decisions by selecting the option to enable fetching of recommendations from the Semantic Finlex case law finder [12].

The Appi tool can be used as follows. Firstly, the application is given an input, e.g., an abstract of a Finnish court decision in text format. Next, the application is configured to identify and link named entities, e.g., using FinBERT, Reksi, and ARPA. The ARPA tool can be used by selecting a linking option, e.g., a configuration for domain information such as legal terminology. Lastly,

[10] `http://light.onki.fi/henko/en/`
[11] `https://finto.fi/yso-paikat/en/`
[12] `https://tieteentermipankki.fi/wiki/Termipankki:Etusivu/en`
[13] `http://eurovoc.europa.eu`

the user can enable the fetching of recommendations using the case law finder. After configuring the application, the user can click the "Annotate" button, and Appi annotates the given input and retrieves recommendations. The results are presented as shown in Fig. 1.

Fig. 1. Results of annotating an abstract of a court decision.

The results are presented under the configuration interface accompanied by a legend that shows available named entity types and how they are shown in text. Below the legend is the annotated text and on its right side a list of entities found in the text (by type). The recognized entities are shown in text with links, and by clicking them a popup appears and shows the description of the 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 documents that have been retrieved for the input text. At the bottom of the page, the results are presented in JSON form that can be viewed or downloaded by clicking the tab.

In this example (Fig. 1), Appi has identified a reference to time, statutes, and references to different contextual terms in an abstract of a court decision. 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. The Reksi tool links statutes and case law to Semantic Finlex. However, currently the endpoint doesn't contain all the alternative names for the statutes and the linking fails for missing names. 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.

5 Related Work and Discussion

The Appi service provides easy access to related legal texts and helps to understand the terminology. The inspiration for the application has been the contextual reader application CORE [6] 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 [9], configured to use content-related ontologies to provide context for the user. However, the tool does not have a powerful disambiguation system like other named entity linking tools, e.g., DBpedia Spotlight[14] [8] and Gate Cloud[15] [7]. For this purpose, Nelli was created, and based on it, a contextual reader was implemented for the BiographySampo portal [13] for Finnish biographical texts. In BiographySampo, the entities are not extracted with Nelli in real time but in a preprocessing phase that ensures robust semantic disambiguation similarly to [10,2]. The results are recorded in RDF format and visualized by the contextual reader by querying them from the BiographySampo endpoint. The Nelli tool was modified to serve better the needs of Semantic Finlex and used to build Appi application that can disambiguate in real time, visualize the results in a contextual reader, and function as an annotation tool.

The initial demo application, Appi, manages to identify, highlight, and link named entities from a text. The annotation accuracy using Nelli was approximately 80% [13] for people and places in biographical texts. The service has been upgraded and the results are promising but it still needs a formal evaluation, which will be carried out in the future. The recommendations and legal text references can be identified with varying accuracy partially due to lack of document metadata. The current version is still under development and more work needs to be done so that it can be utilized to extract all references to legislative texts such as EU statutes and link them to the CELLAR system[16]. The Appi demo 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[17] funded by the the Ministry of Justice in Finland. CSC – IT Center for Science, Finland, provided us with computational resources.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
2. Ferragina, P., Scaiella, U.: 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. pp. 1625–1628. ACM (2010)

[14] `https://www.dbpedia-spotlight.org/demo/`
[15] `https://cloud.gate.ac.uk`
[16] `https://data.europa.eu/euodp/data/dataset/sparql-cellar-of-the-publications-office`
[17] `https://seco.cs.aalto.fi/projects/anoppi/en/`

3. Frosterus, M., Tuominen, J., Hyvönen, E.: 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). pp. 115–124. IOS Press (2014)
4. Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., Löfberg, L.: Old Content and Modern Tools-Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910. arXiv preprint arXiv:1611.02839 (2016)
5. Mäkelä, E.: Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In: Proceedings of the ESWC 2014 demonstration track. pp. 424–428. Springer-Verlag (2014)
6. Mäkelä, E., Lindquist, T., Hyvönen, E.: CORE – a contextual reader based on linked data. In: Proceedings of Digital Humanities 2016, Krakow, Poland (long papers). pp. 267–269 (2016)
7. Maynard, D., Roberts, I., Greenwood, M.A., Rout, D., Bontcheva, K.: A framework for real-time semantic social media analysis. Journal of Web Semantics **44**, 75–88 (2017)
8. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding light on the web of documents. In: Proceedings of the 7th international conference on semantic systems. pp. 1–8. ACM (2011)
9. Oksanen, A., Tuominen, J., Mäkelä, E., Tamper, M., Hietanen, A., Hyvönen, E.: Semantic Finlex: Transforming, publishing, and using Finnish legislation and case law as linked open data on the web. In: Peruginelli, G., Faro, S. (eds.) Knowledge of the Law in the Big Data Age, Frontiers in Artificial Intelligence and Applications, vol. 317, pp. 212–228. IOS Press (2019)
10. Piccinno, F., Ferragina, P.: From TagME to WAT: A new entity annotator. In: Proceedings of the first international workshop on Entity recognition & disambiguation. pp. 55–62. ACM (2014)
11. Ruokolainen, T., Kauppinen, P., Silfverberg, M., Lindén, K.: A Finnish news corpus for named entity recognition. Language Resources and Evaluation **54**(1), 247–272 (2020)
12. Sarsa, S., Hyvönen, E.: Searching case law judgements by using other judgements as a query (2019), Aalto University, paper: `https://seco.cs.aalto.fi/publications/2019/sarsa-et-al-csfinder.pdf`
13. Tamper, M., Hyvönen, E., Leskinen, P.: 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 (2019), forthcoming
14. Tamper, M., Leskinen, P., Tuominen, J., Hyvönen, E.: Modeling and publishing Finnish person names as a linked open data ontology. In: 3rd Workshop on Humanities in the Semantic Web (WHiSe). CEUR Workshop Proceedings (2020), accepted
15. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S.: Multilingual is not enough: BERT for Finnish (2019), `http://arxiv.org/abs/1912.07076`, arXiv preprint arXiv:1912.07076