Tamper, Minna; Leskinen, Petri; Ikkala, Esko; Oksanen, Arttu; Mäkelä, Eetu; Heino, Erkki; Tuominen, Jouni; Koho, Mikko; Hyvönen, Eero

AATOS – A configurable tool for automatic annotation

Published in:
Language, Data, and Knowledge - First International Conference, LDK 2017, Proceedings

DOI:
10.1007/978-3-319-59888-8_24

Published: 01/01/2017

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Tamper, M., Leskinen, P., Ikkala, E., Oksanen, A., Mäkelä, E., Heino, E., Tuominen, J., Koho, M., & Hyvönen, E. (2017). AATOS – A configurable tool for automatic annotation. In *Language, Data, and Knowledge - First International Conference, LDK 2017, Proceedings* (Vol. 10318 LNAI, pp. 276-289). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 10318 LNAI). Springer. https://doi.org/10.1007/978-3-319-59888-8_24

AATOS – a Configurable Tool for Automatic Annotation

Minna Tamper, Petri Leskinen, Esko Ikkala, Arttu Oksanen, Eetu Mäkelä, Erkki Heino, Jouni Tuominen, Mikko Koho, and Eero Hyvönen

Semantic Computing Research Group (SeCo), Aalto University, Finland and
HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://seco.cs.aalto.fi, http://heldig.fi
firstname.lastname@aalto.fi

Abstract. This paper presents an automatic annotation tool AATOS for providing documents with semantic annotations. The tool links entities found from the texts to ontologies defined by the user. The application is highly configurable and can be used with different natural language Finnish texts. The application was developed as a part of the WarSampo[1] and Semantic Finlex[2] projects and tested using Kansa Taisteli magazine articles and consolidated Finnish legislation of Semantic Finlex. The quality of the automatic annotation was evaluated by measuring precision and recall against existing manual annotations. The results showed that the quality of the input text, as well as the selection and configuration of the ontologies impacted the results.

1 Introduction

Document databases are explored by users on a daily basis. The databases can be searched for different documents but it can be difficult to obtain satisfactory results easily. To improve the search results, search engines can utilize document metadata that contains descriptive keywords among other descriptive data about the document [4]. One way to enrich document metadata is by using Semantic Web technologies where relevant keywords would be identified from each document and linked to existing controlled vocabularies, giving the keywords semantic meanings. In the context of the Semantic Web this can be also called annotating.

Manually annotating or subject indexing each document is, however, laborious, costly, and time consuming work. [3,15] On the other hand, this is not a simple task for the computer either. Identification of terms from texts by extracting words can be inefficient and inaccurate. One word can mean many things. For example, it might be difficult to distinguish whether a word refers for example a person's name or a place. Futhermore, a referring expression may consist of multiple words; it can be difficult to identify a term if different chunks of words form a term separately and together. These tasks would require dedicated algorithms and possibly domain specific information extraction (IE) methods combined with Natural Language Processing (NLP) approach to identify terms with satisfactory precision.

[1] http://seco.cs.aalto.fi/projects/sotasampo/en/
[2] http://seco.cs.aalto.fi/projects/lawlod/en/

This paper presents a generic tool for automatic annotation that has been developed as part of the WarSampo and Semantic Finlex projects in the Semantic Computing Research Group (SeCo)[3]. The tool is used to annotate Finnish documents and is tested in two use cases: Kansa Taisteli magazine articles and the consolidated legislation of Semantic Finlex[4]. Kansa Taisteli magazine articles can be searched and explored in the WarSampo portal, which models the Second World War in Finland as Linked Open Data (LOD). [13] Kansa Taisteli is a magazine published by Sanoma Ltd and Sotamuisto association between 1957 and 1986. [24] The magazine articles cover the memoirs of WW2 from the point of view of Finnish military personnel and civilians. Semantic Finlex, on the other hand, is a service that offers the Finnish legislation and case law as Linked Open Data [7]. The results of the annotation process for both projects have been published in the Linked Data Finland service[5] [12].

2 The Annotation Model

Due to the monotonous and costly nature of manual annotation, it is important to design annotation tools where the annotation process can be performed as swiftly as possible. The entrance barrier to annotation can be lowered with a generic annotation tool because it would reduce development costs and preparatory work. [26]

One example of an automatic annotation system is the DBpedia Spotlight service[6]. DBpedia Spotlight is an open source service that recognizes DBpedia resources in natural language text. It is a solution to linking unstructured information sources to the Linked Open Data cloud [6]. In a generic automatic annotation tool, the text can ideally be linked to multiple ontologies. In addition to linking documents, the application needs to be able to select the best describing keywords for a document. This is not a simple task and it needs natural language processing methods in addition to linking text correctly to ontologies.

In natural language processing, *named entity linking (NEL)* [9,2] is the task of determining the identity of named entities mentioned in a text, by linking found named entity mentions to strongly identified entries in a structured knowledge base. In general, NEL consists of *named entity recognition (NER)*, followed by *named entity disambiguation (NED)* [16,9]. NER [20,8] recognizes the occurrence or mention of a named entity (e.g., people's names, organizations, locations) in a text and NED [2,25,5] identifies which specific entity it is. A further refinement to this formulation is suggested by Hachey et al. [9], which divides NEL into *extraction*, *searching* and *disambiguation* steps.

The automatic annotation tool (AATOS)[7] presented in this paper has been designed by taking into consideration the use cases and the background of the field. In order to annotate Finnish texts, it requires specific tools designed for the Finnish language. In addition to the NLP approach, it needs to identify relevant concepts and named entities

[3] http://seco.cs.aalto.fi
[4] http://data.finlex.fi
[5] http://www.ldf.fi
[6] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Introduction
[7] https://github.com/SemanticComputing/aatos

and link them to controlled vocabularies with matching terms. Based on both of the requirements mentioned, a general model for annotation has been created and implemented using Python.

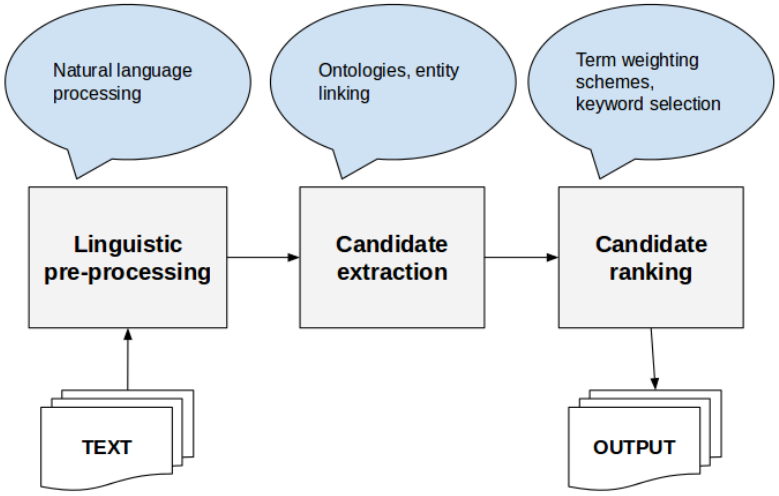

Fig. 1. Model of annotation.

As shown in Figure 1, AATOS consists of the following components or phases: 1) linguistic preprocessing, 2) candidate extraction, and 3) candidate ranking. These components process the input in the given order and produce an output.

The application needs by default its input in text format. It is possible to give the input in HTML format and the application is able to extract all text from the body element to use as an input. The user needs to define input and output formats, their locations (URL or file path), file extensions, an output file, and possible a source file in RDF format that can be populated with the results. The input text is written in a natural language that can be processed using linguistic tools and methods. The linguistic preprocessing transforms words on the textual data into their base form (lemma), and extracts textual entities that can be linked to ontologies in the next phase. The candidate extraction component, given vocabularies, can link the text with other resources and identify keyword candidates to create more accurate descriptions of the documents. For these purposes the SPARQL ARPA tool can be used. ARPA is a configurable automatic annotation tool that uses LAS (Lexical Analysis Services), SPARQL, and ontologies to identify entities from a text document, and in return gives suggestions for annotating texts [18].

In order to use ARPA the user needs to define ARPA configurations file that can be given to the annotation tool from command line. Each ARPA configuration is defined separately and the set of configurations are shown in Table 1. The tool executes the given

ARPA configurations and produces a list of concepts (URI references). Disambiguation strategies are utilized to identify the best concepts from the result list (Primary, Second, Third) and they can be used to identify the best property for linking. In addition, the list can be filtered to contain only the most frequent terms. A stop word list can be used to remove terms that the user is not interested in. Such terms could be terms in documents that do not describe the content but give structure to it such as form labels like name, address, and country.

#	Configuration name	Values	Description
1	**Name**	Text	Name of the ARPA query or ontology
2	**URL**	URL	URL to the ARPA query
3	**Map Property Type**	URL	Map results to a given property.
4	**Map Graph**	URL	Map results to a given graph.
5	**Frequency Limit**	Number	Lowest accepted term frequency
6	**Stop Word List**	File name	File that contains a list of used stop words.
7	**Ranked**	True / False	Set ranking on / off
8	**Top N**	Number	If ranking is on, this setting can be used to create a top n listing of keywords per document.
9	**Primary**	URL	Disambiguation strategy: the best target property for linking
10	**Second**	URL	Disambiguation strategy: 2nd best target property for linking
11	**Third**	URL	Disambiguation strategy: 3rd best target property for linking

Table 1. ARPA specific configuration options for AATOS.

In the last phase, the application has acquired the linked data and the keyword candidates. The data and the text are analyzed to determine which keyword candidates are useful in describing the content and to function as keywords for the given document. For this purpose, term relevance and different weighting schemes are required. Using weights, the extracted candidates can be ranked and the top candidates picked. In addition, the application has support for keyword density calibrations and it allows the user to pick top n candidates. The user can define whether to limit or a range for the keywords in the ontology configurations.

Finally, the application needs to produce an output after the candidate selection. The output contains the results in a user specified format such as RDF or CSV that can be defined in the tool configuration.

3 Use Case 1: Kansa Taisteli Magazine Articles

The first case for automatic annotation is the Kansa Taisteli magazine articles. The magazine articles are publicly available in PDF format via a website of The Association for Military History in Finland[8] in collaboration with Bonnier Publishing.

[8] `http://kansataisteli.sshs.fi`

The magazine articles were accompanied with manually collected metadata for 3,385 articles [24]. The metadata contains information regarding the article (author, title, issue, volume, and pages) in addition to annotations describing the content (war, arms of service, a military unit, place, and comments). The articles can be browsed according to their metadata via a faceted search demo application[9]. The metadata is available at the WarSampo data service[10].

3.1 Extraction of Text

In order to perform automatic annotation for the Kansa Taisteli magazine articles, AATOS requires the articles in text format. For extracting texts from the PDF files, two tools were used: ABBYY FineReader[11] and Tesseract[12].

During the evaluation of the OCR tools, it was noted that Tesseract consistently produced solid results that contained a few errors. ABBYY FineReader, on the other hand, seemed to fare better with the Finnish texts as the error rates were much lower than with Tesseract. However, during testing it was noted that, unlike with Tesseract, ABBYY seems to mix up paragraphs for unidentified reasons. Therefore, it was decided that both tools needed to be used to get the best results from the OCR process. Both tools would be used to extract text and the results would be combined.

The process of combining the results was semiautomatic, using both, comparing the results, and merging them into one result in the end. In addition, occasionally some errors (such as problems with paragraphs) needed to be fixed manually. The preprocessing and post-processing of the articles was laborious and could not eliminate all the errors. Therefore only a small sample of 433 articles was annotated for evaluation. From each decade a year was selected randomly and all magazine issues of that year were selected for processing.

3.2 Automatic Annotation Process

In order to annotate the articles, AATOS requires a set of ontologies and their configuration in the ARPA annotation service. The chosen ontologies come from the WarSampo project: people, military units, Karelian places, and municipalities. External ontologies, such as KOKO ontology and DBpedia, were also used to enrich the annotations with more general concepts. The order of ontologies impacts the annotations and aids in disambiguation; the first ontologies match most of the terms from their vocabularies and this can impact the ability to match terms into other ontologies. In this case the order of ontologies was the following: people, military units, Karelian places and municipalities, DBpedia, and lastly the KOKO[13] ontology.

The ARPA tool configurations[14] for military unit, DBpedia, and KOKO ontologies included the filtering of forenames and surnames. In place ontologies the filtering of

[9] http://www.ldf.fi/dataset/kata/faceted-search/
[10] http://www.ldf.fi/dataset/warsa
[11] http://www.abbyy.com
[12] https://github.com/tesseract-ocr
[13] https://finto.fi/koko/en/
[14] https://github.com/SemanticComputing/aatos-arpa-configurations

forenames and surnames cannot be used because Finnish names for places and villages are similar to surnames [17], such as ”Kestilä” which can be a name of a place or to a person's surname. The military personnel (people) ontology had the highest N-Gram (5) in order to include a full name and a title whereas others had a lower n-gram length (2 to 3) to target words and open compound words.

In place ontology configurations, places such as water formations and buildings were ignored and only villages, towns, and municipalities were targeted for linking. In most cases most of the smaller places are never mentioned in the Kansa Taisteli articles. Often times a village may be carrying the same name as a building or a lake. Therefore, it was seen as useful to rule out all but municipalities, towns, and villages to minimize confusion. All ontologies include extraction of terms that have been POS tagged as nouns or proper nouns, base forming of the words and setting the default language to Finnish. Nouns and proper nouns were selected as keyword candidates because nouns are preferred parts of speech for terms [1]. In addition, selected ontologies mainly have the terms in the form of nouns (e.g., the KOKO ontology) [22] and proper nouns (e.g., the ontologies of the WarSampo project).

For the Kansa Taisteli magazine articles, the application was configured to produce the output in Turtle format and add the annotations into their corresponding properties. These properties are defined in the configuration along with the output format and target file. In the case Kansa Taisteli magazine articles, all found and linked annotations were added into the dataset without candidate filtering based on term relevancy.

3.3 Evaluation

The automatic annotation results were evaluated by calculating precision, recall, and F-measure for 50 randomly selected articles. The evaluation is laborious and therefore not all 433 articles could be used. In contrast to original manual annotations, the results were richer. This also became visible when calculating and inspecting the precision and recall results.

	METHOD 1		METHOD 2		METHOD 3	
	M. Units	**Places**	**M. Units**	**Places**	**M. Units**	**Places**
P	26.14 %	6.78 %	30.26 %	10.47 %	82.02 %	61.69 %
R	69.70 %	38.46 %	67.65 %	51.92 %	54.89 %	44.28 %
F	38.02 %	11.53 %	41.82 %	17.42 %	65.77 %	51.56 %

Table 2. Evaluation of the annotations produced from the unfixed Tesseract output of Kansa Taisteli magazine articles. P is the precision, R is the recall, and F is the F-measure.

The measures were calculated by comparing the automatic annotations with the original manual annotations. In addition, three different executions of the application for three different sets of inputs that have been produced by the OCR process: untouched OCR output text from the Tesseract OCR tool, automatically fixed text using regular expression patterns, and semi-automatically fixed text. The automatically fixed

	METHOD 1		METHOD 2		METHOD 3	
	M. units	**Places**	**M. units**	**Places**	**M. units**	**Places**
P	25.26 %	6.78 %	30.38 %	10.47 %	79.17 %	61.69 %
R	72.73 %	38.46 %	72.73 %	51.92 %	57.14 %	44.28 %
F	37.50 %	11.53 %	42.86 %	17.42 %	66.38 %	51.56 %

Table 3. Evaluation of the annotations produced from the automatically fixed Kansa Taisteli magazine articles. P is the precision, R is the recall, and F is the F-measure.

text utilizes the regular expression created while combining the results of two OCR tools. The regular expression patterns were created based on systematic and frequently occurring OCR errors. For example, a military unit name in inflected form *JR 35:n* contains a colon that was often transformed into i or z in the OCR output.

In addition to comparing these different versions of the articles, different annotation methods were used to calculate the precision and recall: exact matches (method 1), accepting also direct meronyms (method 2), and all correctly linked terms (method 3). Method 1 accepts only the exact matches of terms. In method 2, exact matches and meronyms are also counted as positive matches because original annotations sometimes use municipalities instead of villages that are part of the municipality. For example, sometimes in the manual annotations the articles have been annotated with specific municipalities. For example, the text itself may mention the villages of that municipality and they were counted as positive matches for the municipality in method 2. In method 1, the villages are negative matches and only the municipality is a positive match. In addition, a third method was also used to calculate the measures in comparison to what is found from the article texts. It interprets all correctly extracted and linked matches as true positives.

	METHOD 1		METHOD 2		METHOD 3	
	M. units	**Places**	**M. units**	**Places**	**M. units**	**Places**
P	25.77 %	6.80 %	30.77 %	10.55 %	80.61 %	61.82 %
R	75.76 %	38.46 %	75.00 %	51.92 %	59.40 %	44.42 %
F	38.46 %	11.56 %	43.64 %	17.53 %	68.40 %	51.69 %

Table 4. Evaluation of the annotations produced from the semi-automatically fixed Kansa Taisteli magazine articles.

The difference between the results of unfixed, automatically fixed, and semi-automatically fixed results, shown in Tables 2, 3, and 4, are notable. Depending on the method the results vary. The precision is poor for all but method 3. The precision for methods 1 and 2 depends on the interpretation of original annotations and their correctness. Whereas the method 3 measures how well the mentioned military units and places were found and linked correctly from the article texts.

The difference between precision and recall for places and military units is notable. The precision is lower for the places mainly because of the regular expression fixes con-

centrating on military units. In comparison to a study by Kettunen et al. [14], AATOS produces similar results. It performed somewhat better in finding correct matches. OCR post-processing had a positive impact on the results and it is visible that the recall was impacted by the amount of OCR post-processing, especially in the case of military units. However, the military unit results are weighted down by a few remaining irregular OCR errors whereas the issue. The issues that impacted the linking of places are presented in Table 5.

Error type	**Amount**	**Percentage**
1 Wrong place	32	12.12 %
2 Ambiguous	14	5.30 %
3 Confusion between places and people's names	16	6.06 %
4 Noise from other articles	9	3.41 %
5 Clutter (for example advertisements)	7	2.65 %
6 ARPA / LAS error	1	0.38 %
7 Misidentified POS	9	3.41 %
# TOTAL	88	36.07 %

Table 5. The breakdown of the error types found when the place annotations for semi-automatically texts were analyzed.

The errors encountered can be divided into three groups: firstly, the most numerous category is that of disambiguation errors, further divided based on if they arise from ambiguity in the place data itself, or from the extractor confusing the people's names' with place names. The second category contains errors arising from the faulty article segmentation in the magazine data. Finally, there are errors relating to the tool itself, arising for example from faulty inflection handling or incorrect part of speech filtering. From these results it is apparent that more robust disambiguation of the places would be needed. Luckily, this is a well-researched area, so ready choices for this are available for future work, e.g. [21,10,11].

3.4 Application: Semantic Search and Recommending

The purpose of the faceted search application[15] is to help a user to find Kansa Taisteli articles and to provide context to the found articles by showing links to related WarSampo data. Contextual Reader (CORE) [19] was integrated into the application, to highlight found concepts and offering additional information about them, when viewing the PDF format article.

The updated Kansa Taisteli magazine article perspective is shown in Figure 2. In the perspective, the user can find articles by using the author, magazine, a related place, army units, or mentioned terms facets. The facet will show a list of mentioned terms and names that can be used to filter the article list. The mentioned terms facet adds diversity

[15] `http://sotasampo.fi/articles`

into the article search. By adding the mentions of terms and names as a facet into the web application, the user can find articles that contain certain terms, army units, people, or places. For example, a user can search for articles that mention a person or the term *lice*.

Fig. 2. The faceted search browser targeting the Kansa Taisteli magazine articles.

4 Use Case 2: Semantic Finlex

Semantic Finlex is a service that offers the Finnish legislation and case law as Linked Open Data. The purpose of automatic annotation in the Semantic Finlex project was to make it easier to read, find, and browse statutes and case laws. To achieve this, the metadata had to be enriched by linking it to ontologies. [7] The goal of automatic annotation is to describe the contents of each document accurately and plentifully using keywords.

4.1 Automatic Annotation Process

The ontologies used for the law documents were: Combined Legal Concept Ontology, Original Finlex Vocabulary (FinlexVoc), EuroVoc[16] ontology, KOKO ontology, and Finnish DBpedia. The general ARPA configurations for all cases included the filtering out all but nouns and proper nouns and base forming of terms, and the default language is set to Finnish. The typical n-gram length for these ontologies is set to 3. The SPARQL query is set to exclude numbers and the length of the terms is calculated to enable the selecting of the longest match for the terms. For example, when linking the text *European Union* the ARPA can find, depending on the ontology, matches such as *Europe* and *European Union*. It is important that the tool picks the longest option out of the two as it is the correct one and therefore it was implemented into the application.

[16] `http://eurovoc.europa.eu/`

In the EuroVoc ontology, the used SPARQL query was set to match strings into synonyms to maximize the amount of found links. In addition, the Combined Legal Concept Ontology, EuroVoc, FinlexVoc, DBpedia, and KOKO ontologies are set to target only Finnish terminology. Also, DBpedia is restricted to law terminology in SPARQL and the matches to category names or properties are ignored whereas KOKO ontology targets general-purpose terminology.

The results were set to be filtered based on the relevancy of the concept to the text. The application was set to produce the results in RDF format and to add selected annotations (based on the linked ontology) into their corresponding properties. The unidentified textual entities are filtered out respectively.

The initial results, however, were not satisfactory as there were problems with word recognition and ambiguity. A stopword list was required to filter out the most common terms such as *article*, *Finland* or *law*. The need to add term relevancy analysis or weighing schemes arose, as the purpose of the task is to identify the relevant concepts and not all named entities like in the Kansa Taisteli case. In the annotation process, a simple TF-IDF measure was used to rank each term found in the text.

4.2 Evaluation

The annotation process was executed for 2,803 law documents. The evaluation of AATOS was done by using the R-precision measure. R-precision expresses the precision for the top n keywords where n is the number of keywords in the original annotations. In order to measure the R-precision of the annotation for the law documents, AATOS was configured to use the same controlled vocabulary FinlexVoc with the same keyword density that was used in the original material. After the automatic annotation, 30 documents were selected randomly and their keywords compared with the original annotations.

The calculations for R-precision were done by selecting the same amount of keywords from the automatically produced keywords as in the original keywords and comparing them. The keywords from the annotation tool result set were selected by picking the keywords that were evaluated by the weighting scheme as the most relevant to the document. The R-precision result is equal to the precision and recall measures when the amount of keywords for both sets used in the calculations is the same. The result of the R-precision calculation is 45.45 % for this result set.

The low amount of keywords in the original annotations has impacted the result of the R-precision calculations. For example, sometimes a keyword was found by the annotation tool but it was evaluated not relevant enough for the document. If the amount of keywords for a document would have been 5 instead of 1 in the original annotations, the keyword would have been included in the list of generated keywords. The results are, however, similar but not fully comparable to the results of Sinkkilä et al. [23] for different Finnish texts. AATOS performed well in contrast to the tools and strategies used in the study. The precision and recall are higher than the produced precision 27.00 % and recall 24.40 % using TF-IDF, FDG[17] and other tools by the earlier study by Sinkkilä et al.

[17] `http://www.connexor.com`

#	Error type	Amount	Percentage
1	Keyword found but evaluated not relevant enough	20	29.85 %
2	Keywords not found in the document	14	20.90 %
3	Configuration error (language detection)	1	1.49 %
4	Source material error	1	1.49 %
5	Tool error	1	1.49 %
	TOTAL	37	55.22 %

Table 6. Error types found from the result set of the Semantic Finlex.

The evaluation results are presented in Table 6. The encountered errors can be divided into three groups: firstly, the most numerous category is that of low keyword relevancy. The second category contains configuration errors, tool errors, or errors related to the source materials. From these results it is apparent that more robust method of evaluation for the keywords would be needed.

5 Conclusions, Discussion, and Future Work

This paper presents a new highly configurable and generic tool for annotating and subject indexing documents. It can be configured in multiple ways to produce semantic annotations for different Finnish texts. It links textual entities to matching concepts in controlled vocabularies of the user's choice and produces output in RDF and CSV formats. For subject indexing, the application supports adding different evaluation methods such as TF-IDF that was added into the application during the project. It also supports multiple ways to define keyword density.

This paper presented two use cases for AATOS: Kansa Taisteli magazine articles and Semantic Finlex. In both cases the success of the tool depended on the interpretation of the results. Compared with a human annotator the tool provides a richer amount of annotations.

Disambiguation of the annotations proved to be a challenging task. The selection and the order of ontologies can be used to remove ambiguity. For example, in Kansa Taisteli magazine articles the issue was approached by prioritizing the context specific ontologies. In addition, there are ontology specific configurations for determining if some concepts are better than others and need to be prioritized. These actions helped to minimize the amount of issues regarding the ambiguity of terms. In case Kansa Taisteli, there remain challenges such as differentiating between places with the same names, last names, and place names.

The OCR quality impacted the results for Kansa Taisteli magazine articles. A semi-automatic handling of the results was required and as a byproduct a list of regular expressions was constructed to aid in the correction of the errors. During the evaluation it was noticed that the post-processing of the OCR output improved the annotations and prevented erroneous annotations. However, there is still a need for improvement and further developing an automatic set of rules could speed up the process of post-processing of OCR output.

In case Semantic Finlex, the challenge was the estimation of relevancy and keyword density. It would be interesting to try other strategies for selecting the keyword amount. In addition, a few new terms should be added to the stopword list to see how it would impact the results. All this is fine-tuning of the application configurations. In general, the application manages to produce satisfactory results.

In addition to improvements mentioned above, the application can benefit from future development. It requires more fine-tuning and optimization. In order to utilize the application more efficiently it needs to be possible to run as a compact command line tool. Also a graphical user interface could be useful for the users and for testing purposes. In addition to these improvements, large scale testing is needed.

Acknowledgements Our work was funded by the Media Industry Research Foundation of Finland, the Ministry of Education and Culture, the Finnish Cultural Foundation, and the Ministry of Justice. The Association for Military History in Finland and Bonnier Publications provided the project with resources and published the Kansa Taisteli magazine articles for public usage. Kasper Apajalahti originally converted the metadata into an RDF format. Timo Hakala provided the manual annotations for the Kansa Taisteli magazine articles.

References

1. Anderson, J.D.: Guidelines for indexes and related information retrieval devices. NISO Press Bethesda, MD, USA (1997)
2. Bunescu, R.C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: EACL. vol. 6, pp. 9–16 (2006)
3. Chung, Y.M., Pottenger, W.M., Schatz, B.R.: Automatic subject indexing using an associative neural network. In: Proceedings of the third ACM conference on Digital libraries. pp. 59–68. ACM (1998)
4. Committee on Cataloging: Task force on metadata. final report. Tech. rep. (June 2000), http://libraries.psu.edu/tas/jca/ccda/tf-meta6.html
5. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: EMNLP-CoNLL. vol. 7, pp. 708–716 (2007)
6. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics) (2013)
7. 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 (December 2014)
8. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: Coling. vol. 96, pp. 466–471 (1996)
9. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with Wikipedia. Artificial Intelligence 194, 130–150 (Jan 2013), http://dx.doi.org/10.1016/j.artint.2012.04.005
10. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 782–792.

EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145521
11. Hu, Y., Janowicz, K., Prasad, S.: Improving Wikipedia-based place name disambiguation in short texts using structured data from DBpedia. In: Proceedings of the 8th Workshop on Geographic Information Retrieval. pp. 8:1–8:8. GIR '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2675354.2675356
12. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 226–230. Springer–Verlag (May 2014)
13. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo data service and semantic portal for publishing linked open data about the second world war history. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). Springer-Verlag (May 2016)
14. Kettunen, K., Kunttu, T., Järvelin, K.: To stem or lemmatize a highly inflectional language in a probabilistic IR environment? Journal of Documentation 61(4), 476–496 (2005)
15. Lauser, B., Hotho, A.: Automatic multi-label subject indexing in a multilingual environment. In: Research and Advanced Technology for Digital Libraries, pp. 140–151. Springer (2003)
16. 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)
17. Mikkonen, P., Paikkala, S.: Sukunimet. Otavan kirjapaino Oy (2000)
18. Mäkelä, E.: Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 424–428. Springer–Verlag (May 2014)
19. Mäkelä, E., Lindquist, T., Hyvönen, E.: CORE - a contextual reader based on linked data. In: Proceedings of Digital Humanities 2016, long papers. pp. 267–269 (July 2016), http://dh2016.adho.org/abstracts/2580
20. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
21. Overell, S., Rüger, S.: Using co-occurrence models for placename disambiguation. Int. J. Geogr. Inf. Sci. 22(3), 265–287 (Jan 2008), http://dx.doi.org/10.1080/13658810701626236
22. SFS 5471: Guidelines for the establisment and maintenance of Finnish language thesauri. SFS standard, Finnish Standards Association (1988)
23. Sinkkilä, R., Suominen, O., Hyvönen, E.: Automatic semantic subject indexing of web documents in highly inflected languages. In: Extended Semantic Web Conference 2011. pp. 215–229. Springer (2011)
24. The Association for Military History in Finland: Kansa taisteli magazines 1957 - 1986 (2014), http://www.sshs.fi/sitenews/view/-/nid/92/ngid/1
25. Wentland, W., Knopp, J., Silberer, C., Hartung, M.: Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (May 2008), http://www.lrec-conf.org/proceedings/lrec2008/
26. Yimam, S.M., Biemann, C., Eckart de Castilho, R., Gurevych, I.: Automatic annotation suggestions and custom annotation layers in WebAnno. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 91–96. Association for Computational Linguistics, Baltimore, Maryland (June 2014), https://www.aclweb.org/anthology/P/P14/P14-5016.pdf