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

Named entity linking in a complex domain

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

DOI:
10.1007/978-3-319-59888-8_10

Published: 01/01/2017

Document Version
Peer reviewed version

Named Entity Linking in a Complex Domain: Case Second World War History

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

Semantic Computing Research Group (SeCo), Aalto University 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 discusses the challenges of applying named entity linking in a rich, complex domain – specifically, the linking of 1) military units, 2) places and 3) people in the context of interlinked Second World War data. Multiple sub-scenarios are discussed in detail through concrete evaluations, analyzing the problems faced, and the solutions developed. A key contribution of this work is to highlight the heterogeneity of problems and approaches needed even inside a single domain, depending on both the source data as well as the target authority.

1 Introduction

This paper addresses entity linking [8,1] in a rich, complex, but focused environment. We extract links from textual data related to Second World War history against richly described linked data datasets on the places, people, events and army units mentioned therein. Here, problems arise due to the rapid and wide-ranging turbulence caused by the war. Structures of army units as well as the existence and roles of the people associated with them change rapidly, while geographical entities also change administrative jurisdiction. In addition, the people touched by war are a multitude from all ranks of society, causing further problems in for example disambiguation of homonymous people.

The primary contribution of this paper is in showing the individual deliberations and decisions taken to increase recall and precision in such a focused environment. On the other hand, while the individual tweaks are bound to the data, they do take place inside a pipeline that orients these considerations to a more general framework.

The context of this work is the WarSampo aggregated linked open dataset[1], which aims to provide richly interlinked data into the Second World War in Finland [12]. In total, the WarSampo dataset contains data of more than a dozen different types (e.g. casualty data, photographs, events, war diaries, and historical maps) from an even larger pool of sources (e.g. the National Archives, the

[1] http://www.ldf.fi/dataset/warsa

Defense Forces, and scanned books). On top of the data, the WarSampo portal[2] serves multiple end user viewpoints. In the portal, persons, units, places, and events have homepages of their own, generated and linked to each other automatically based on the underlying Linked Open Data cloud. This rich interlinking allows one to e.g. move from the homepage of a war event to units and people participating in it, and to the photographs and articles depicting the persons, units, and places, or the event itself.

In order to create the links these functionalities are based on, different paths needed to be taken based on the type of source. This paper focuses on the datasets where the links to the actors, places and military units involved were described as free text, thus needing the application of Named Entity Linking (NEL) techniques [8,1]. These were as follows (examples translated from Finnish):

Events Short descriptions of about 1000 events gathered from numerous sources such as timelines printed in books. Example: "Defense battles near Viipuri continued. Lieutenant general Öhquist gave colonel Kaila an order to occupy the main defense line of the 3rd Division at Patterimäki.", 1940-03-12

Photographs Metadata, including captions, for a collection of some 160 000 wartime photographs from the Finnish Defense Forces. Example: "Field Marshal Mannerheim with his entourage at the headquarters of the 4th Division, negotiating with colonel Autti.", "Savujärvi" (place), 1943-06-22

Articles Over 3300 articles from the Kansa Taisteli war remembrance magazine published by Sanoma Ltd and the Sotamuisto association between 1957 and 1986 [21], each generally 2-5 pages of prose.

In the following, we will first present the reference datasets of places, people and military units against which these sources were linked. After that, the general pipeline used for linking will be shortly discussed, followed by detailed experiences and evaluations of linking each source against each type of data.

2 The WarSampo Reference Datasets

In creating a geospatial reference for WarSampo, the main source for trouble was the fact that at the end of the Second World War, Finland was forced to cede large areas of land to the Soviet Union. Fortunately, these changes happened only at the end of the war, so actual temporal reasoning over places [11] in this timeframe wasn't needed.

On the other hand, merely relying on a modern gazetteer wouldn't work. After the end of the war, the ceded areas have naturally not been included in any Finnish place registries, while in more general registries, they are referred to by their modern Russian names. Unrectified, this would cause major problems to named entity linking of Finnish wartime material, as these areas, particularly the Karelia region, also happened to be the major arenas of action.

Thus, historical sources were used to create a snapshot of Finnish places covering the years 1939-44. Four sources were used: 1) the National Archives of

[2] `http://sotasampo.fi/en/`

Finland's map application data of 612 wartime municipalities,[3] 2) the Finnish Spatio-Temporal Ontology [11] describing Finnish municipalities in different times, 3) a dataset of Karelian map names (35 000 map names with coordinates and place types from the years 1922–44), and 4) the current Finnish Geographic Names Registry[4] (800 000 places) for places that had no reference source for the years 1939-44. In addition, some 450 historical map sheets from two atlases were rectified on modern maps, which makes it possible to examine the places on both modern and historical maps, without having to create explicit links between the place names and historical map sheets.

In contrast to the places, the actors participating in the war did change their status constantly. Thus, there was no way around a model that takes into account temporal changes. Accordingly, in WarSampo, the actor data model is an actor-event-model based on CIDOC CRM[5] [3]. According to CIDOC CRM, an event represents any change of status that divides the timeline into a period before and after the event, allowing for reconstructing the status of an actor at a certain moment by following these events through time.

Currently the actor data in WarSampo contains information on 99 000 people, collected semiautomatically from various sources: lists of generals and commanders, lists of recipients of honorary medals, the Finnish National Archives casualties database [14], the Finnish National Biography[6], Wikidata, and Wikipedia. Besides military personnel, 580 civil persons with political or cultural significance were included in WarSampo from the aforementioned sources due to their connections to other WarSampo data. The military unit data on the other hand consists of over 3000 Finnish army units, sourced mainly from War Diaries and Organization Cards.

Examples of events extracted from these sources into WarSampo are listed in Table 1. Following through such events, one can for example identify the rank and unit of a person at a particular date, as well as track their geographic position. However, due to gaps in the source data, exact dates are not available for all events. For example, a promotion event is created for all ranks mentioned in the sources, even if no specific date of attaining that rank is known.

For ease of use, all unit names as well as their abbreviations and nicknames are also repeated as alternate labels for the unit resources themselves. For people, their first and family names are given in separate properties, with known nicknames given as alternative labels.

3 Modular Architecture for Named Entity Linking

Named entity linking (NEL) [8,1] 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

[3] `http://kronos.narc.fi/kartta/kartta.html`
[4] `http://www.maanmittauslaitos.fi/en/digituotteet/geographic-names`
[5] `http://cidoc-crm.org`
[6] `http://www.kansallisbiografia.fi/english/`

Event type	Example
Unit formation	Troop founded as 24th Squadron (abbr. LLv 24)
Unit joining	Being part of Flying Regiment 2
Unit naming	Changing the name to 24th Fighter Squadron (abbr. HLeLv 24)
Troop movement	Troop Movement to Vesivehmaa and Selänpää
Unit dissolution	32th Squadron was dissoluted in December 1944
Birth	Born at Pyhäjärvi, 1913
Person joining	Serving as commander in the 24th Fighter Squadron, 1939
Promotion	Promotion to the rank of captain, 1941
Medal awarding	Awarded with the Mannerheim Cross of Liberty, 1942
Wounding	Simo Häyhä was wounded by an enemy sniper, 6th of March 1940
Disappearing	Disappearing of Onni Aaltonen at Äyräpää
Death	Died at Tampere, 2002
Battle	Aerial victory in Tainionkoski: enemy SB-2 shot down, 1939

Table 1. Event types and examples

consists of *named entity recognition (NER)*, followed by *named entity disambiguation (NED)* [16,8]. NER [19,6] recognizes the occurrence or mention of a named entity (e.g., names of persons, organizations, locations) in text, and NED [1,22,2] identifies which specific entity it is. A further refinement to this formulation is suggested by Hachey et al. [8], which divides NEL into *extraction*, *searching* and *disambiguation* steps.

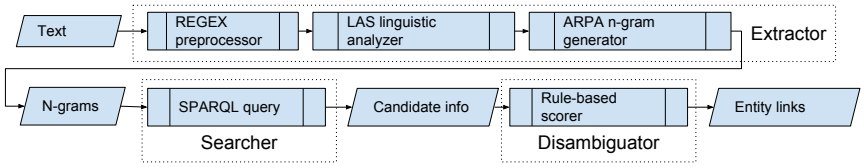

Fig. 1. The Named Entity Linking architecture used

The system used for NEL linking in this paper is based on the ARPA modular configurable annotation architecture[7] [17], and can also be described using the framework presented above, as shown in Figure 1.

[7] `https://github.com/jiemakel/arpa/`

Here, the *extractor component* of the system further divides into a preprocessor, a linguistic analyzer and an n-gram generator. Of these, the preprocessor applies transformations based on a configurable battery of regular expressions in an attempt to regularize textual mentions. For example, unit and military rank abbreviations are standardized, and lists of people normalized to common form. In the case of the Kansa Taisteli magazines, this phase is also responsible for correcting recurring OCR errors found in the automatically extracted text.

As Finnish is a highly inflected language, where e.g. the noun for shop, "kauppa", can appear in a total of 2,253 different forms[8], candidate extraction heavily relies on linguistic analysis [15], whereby each word is annotated with morphological information, including the base or inflected form by which it appears in the reference datasets [18]. As the final part of this candidate extraction phase, the n-gram generator makes use of the linguistic information produced to craft n-grams [17]. Here, rules are used for example to filter candidates by part of speech, and inflectional information is applied to correctly format n-grams comprised of multiple parts or compound words (e.g. in Finnish the base form of "Helsingin varuskunnassa" ["in the garrison of Helsinki"] is "Helsingin varuskunta", where only the latter word is baseformed, and not "Helsinki varuskunta").

After candidate extraction, the *search component* is responsible for searching the reference datasets for potential strong identifiers for the candidates. As opposed to many systems relying on general-purpose knowledge bases or lexical resources such as Wikipedia [20,8] or WordNet [5], the ARPA architecture is tuned for utilizing configurable, domain specific vocabularies. For the WarSampo NEL task, this is important, as for example the Finnish Wikipedia lists only a limited amount of military units and personnel. Thus, in our pipeline, the searcher refers to custom SPARQL endpoints and queries defined for each entity type. These are then used both to retrieve candidates, as well as additional information such as ranks in the case of people.

Finally, given candidate identifiers, it is the job of the *disambiguator component* to rank and select from them. The disambiguator makes use of the information retrieved by the searcher, the original text, and any other data available regarding the text, such as dates, to feed a set of configured rules that rank the candidates and choose the most likely for linking.

4 Named Entity Linking of Military Units

Military unit mentions are generally quite unambiguous, so the main problem in their case was that they can be referred to by their full name, an abbreviation, or a nickname. An example of a photograph caption mentioning military units would be: "Kuvia Peiposten kylästä, jonka II/JR 8. ja 8./JR 8. ankaran taistelun jälkeen saarrostushyökkäyksellä valtasivat" ("Photographs from Peiposten village which II/JR 8 and 8./JR 8 took with an encircling attack after intense

[8] `http://www.ling.helsinki.fi/~fkarlsso/genkau2.html`

battles"). Here, "II/JR 8" refers to the second battalion of the eight infantry regiment, and "8./JR 8" refers to the eight company of the same regiment.

As the actor ontology already contains the aliases for the units, only normalization of the mentions was needed. In addition, because of the unambiguous nature of military unit mentions, no disambiguation was needed when linking them – all candidates returned by the searcher were accepted.

The quality of the linking accomplished was evaluated by taking a random sample of the resources and checking the links that had been produced for those resources. For photographs the sample size was 100, and for events and magazine articles it was 50. The evaluation for each of the different links – i.e. military units, places, and people – was done by a single person, and the same sample sets were used for all link types. The evaluation method could be improved by having multiple people and/or a domain expert check the results.

For military units, the results of the evaluation are presented in Table 2. In the table, Target refers to the target data, N is the the number of mentions, TP is the number of true positive matches and FP is the number of false positives, with P being the calculated precision. Because, in our approach, the recall of the method is maximally limited by the presence or absence of the entity in the reference data, two sets of numbers are reported for both false negatives as well as recall and the F_1 score. $FN_{ont.}$ is the number of entities that would have been available in the reference data but which the method missed, while $FN_{out.}$ is the number of entities outside the ontology. Accordingly, $R_{ont.}$ and $F_{1ont.}$ report recall and the F_1 score with regard to the reference data, while R_{all} and F_{1all} is overall recall and F_1 score.

In relation to the Kansa Taisteli magazines, multiple versions of the data were tested. Here, the score for $Magazines_{orig.}$ relates to the data as it was received, which in this case means text automatically extracted from scanned images of the magazine using optical character recognition (OCR). $Magazines_{auto.}$ on the other hand reports results for a version in which the *extractor component* utilizes regular expressions to correct for commonly occurring OCR errors. Particularly, it was found that the OCR software often misread unit name abbreviations, rendering for example a 1 in them as an l, I or |, and a : as z. Given the particular context in which this happened, these were easy to correct using regex rules. Finally, $Magazines_{clean}$ reports results for a manually cleaned up version.

Target	N	TP	FP	P	$FN_{ont.}$	$FN_{out.}$	$R_{ont.}$	R_{all}	$F_{1ont.}$	F_{1all}
Photographs	9	7	0	1.00	0	2	1.00	0.78	1.00	0.88
Events	11	8	0	1.00	1	3	1.00	0.73	1.00	0.84
$Magazines_{orig.}$	133	73	16	0.82	24	36	0.75	0.55	0.78	0.59
$Magazines_{auto.}$	133	76	20	0.79	21	36	0.78	0.57	0.79	0.66
$Magazines_{clean}$	133	79	19	0.81	18	36	0.81	0.59	0.81	0.68

Table 2. Precision and recall in linking of military units.

As can be seen from the table, precision was perfect for both events and photographs, and satisfactory for the magazines at around 80%. Here, their precision is hindered by problems in the original data which contained only a page level segmentation of the articles. Due to this, for example advertisements were not filtered away from the automatically extracted text, and articles not terminating on page boundaries caused entities to spill over from one article to the other, both increasing the number of false positives. When comparing the different versions of the magazine articles, one can see that OCR errors in the original data did cause significant problems for military unit recall. Of these, the automated cleanup using regular expressions in the extractor component was able to counter some, but not all.

The number of mentions in both the photographs and the events is quite low (9 and 11, respectively). This is because the samples taken were random, and only a portion of the somewhat short descriptions of events and photographs mention military units (or any named entities).

The low overall recall for units for the magazine articles (less than 60%) is mostly due to foreign units and units of the Finnish Civil War which are currently not available in the actor ontology.

5 Named Entity Linking of Places

Regarding the linking of places to the material, the most problematic part of the task was disambiguation. As stated before, the project had at its disposal both a temporally restricted snapshot of places relevant to the war period, as well as a general Finnish gazetteer to fall back upon. Together, these promised good recall, but as places are often very homonymous, both with regard to each other as well as family names, good precision could be difficult to attain. To ascertain the scope of the possible issues, a survey was done on the assembled place registries. The results of this are presented in Table 3, which shows the shares of unique place names by place type, both inside the type as well as overall.

Based on the table, it was decided that a simple priority list approach would be taken. In essence, the disambiguation process was as follows:

1. First, match against historic names of populated places in decreasing order of size inside the focus area of the war, as well as all Finnish municipalities regardless of location.
2. Then, match against other historic geographic names inside the focus area of the war.
3. Finally, match against contemporary modern place names in any source.

Additionally, a list of ca. 100 place names that were consistently confused with other words, such as “Pohjoinen” (north) and “Suomalainen” (Finnish), were excluded. In this simple approach, each place mention was disambiguated independently, and no other information was used.

Place type	Total count	Portion unique	Portion unique inside type
municipality	625	76 %	99 %
town	50	54 %	100 %
village	1544	59 %	88 %
hypsographic feature	10 864	66 %	71 %
body of water	5553	63 %	66 %
man-made feature	14 362	40 %	45%

Table 3. Unique place name portions in the place ontology

Table 4 shows the evaluation results of the approach. In this evaluation, in cases where there is a village in a municipality with the same name, a link to the municipality was considered correct, even if this might have caused a loss of geographic precision.

Target	N	TP	FP	P	$FN_{ont.}$	$FN_{out.}$	$R_{ont.}$	R_{all}	$F_{1ont.}$	F_{1all}
Photographs	92	54	16	0.77	17	21	0.76	0.59	0.77	0.67
Events	67	30	4	0.88	4	33	0.88	0.45	0.88	0.59
Magazines	411	182+1	113	0.62	107	122	0.63	0.44	0.62	0.52

Table 4. Precision and recall in place linking.

In the photograph dataset place information for each photograph was available separately as a textual representation. Therefore, the problems of precision in that dataset reflect purely the ambiguity of place names. In the other datasets, place mentions had to be extracted from text, bringing about problems in e.g. interpreting family names of people as places.

As can be seen from the table, the simple priority list approach yielded acceptable, if not thrilling results in precision and recall within the place ontologies. Overall recall, however, was somewhat low because the place ontologies are still missing some extremely relevant places, such as the Karelian Isthmus. The better performance in linking events is due to these often referring to the major important places where coverage was better, in contrast to the photographs and articles that often also refer to much smaller places, where both ontology coverage is poorer, as well as homonymy problems more numerous. With regard to the different versions of the magazine articles, performance was nearly identical in this task – the only difference was the identification of one additional true positive for the manually cleaned up version of the data (yielding the 182+1 in the TP column of Table 4).

Examining the precision and recall more closely, one gets the notion that these were determined mostly by the ability of the disambiguator component to choose the correct place instance from available options – by making a bad choice, false positives rise by one diminishing precision, while false negatives also rise by one, diminishing recall. To further examine this hypothesis, a separate analysis was done on the magazine article corpus to identify the sources of false positive matches. The general breakdown of this analysis is presented in Table 5.

Error type	**Amount**
Wrong place chosen with correct also available	32
Wrong place chosen, correct not in ontology	28
Person name misidentifed as a place name	18
Other word misidentified as a place name	14
Noise from other articles	11
Noise from advertisements and other non-content	9
TOTAL	113

Table 5. A breakdown of place annotation false positives for semi-automatically cleaned magazine article texts.

As can be seen, indeed the most numerous type of error is where the system has not chosen the correct place even though it was available. This points towards needing more robust geographical disambiguation than the simple, local approach taken here. Luckily, this is a well-researched area, so ready choices for this are available for future work, e.g. [7,10,4,9].

On the other hand, in almost an equal number of cases, a wrong place has been chosen in a situation where the right place didn't exist in our domain ontology at all. For example, in the place ontologies the Karelia region itself is not identified as a geographic location, but instead there is a village, and a historical municipality in western Finland carrying the same name. This points towards the need for simply improving our gazetteer coverage.

The third category of false positives covers situations where a personal name was misidentified as a place name. This is a particularly hard problem for Finnish, because Finnish last names often originate from place names. Finally, also some other words were misidentified as places. For example, sometimes a sentence starts with a capitalized adjective such as Uupunut (tired in English) that could be confused with a village that has the same name.

To an extent, these two last categories could be further optimized for by tuning the configuration. For example, more aggressive filtering based on part of

speech could be added, or further rules defined to try to guess whether a name refers to a person or a place [13].

In addition to these errors, a separate, significant source of false positives for the magazine articles arose from the fact that the articles were automatically segmented from raw OCR results. This in turn caused names appearing outside the article text, such as in other articles or advertisements, to sometimes be erroneously associated with the text under analysis.

6 Named Entity Linking of People

The person ontology of WarSampo contains a total of 99 483 people. As with places, the scope of disambiguation issues here was approximated by counting how many of them could be uniquely identified by various combinations of their names, as shown in Table 6.

Part of name	**Unique instances**	**Portion of unique person names**
family name	10 185	10.2 %
family name and any first name	50 553	50.8 %
full name	92 098	92.6 %

Table 6. Examples of unique name portions in the person ontology

As can be seen, basing linking on just family names for example would create huge problems for disambiguation. On one hand, requiring the full name including all first names would result in low recall as people are not often mentioned by their full name in the material. On the other hand, as discussed in the section on the actor ontology, the linking tool has rich contextual information at its disposal relating to the people – information on e.g. their ranks, awards, units, and deaths, often with attached dates. Thus, here a true effort was made to use as much of this information as possible in identifying the correct people from the material.

Accordingly, for people, the extractor has three tasks, designed to provide the further components with as much contextual information as possible. These are: rank normalization, list standardization, and handling pre-defined cases. First, ranks are normalized by replacing abbreviations and aliases by their proper names. Second, especially the photograph dataset contains lists of people in the form of "majors Jones, Smith, and Davis". In order to have the mentions in the format the searcher expects, lists like these were expanded to include the rank for each individual person: "major Jones, major Smith, major Davis". The expansions were done automatically using regular expressions. Lastly, identified

important corner cases are handled: the spelling of specific names are adjusted to correspond to the person ontology, and other mentions which have been identified as resulting in an incorrect link or a missing link are amended to an unambiguous form.

The searcher then retrieves candidates based on the names and rank in mentions. As just the family name is considered too ambiguous for linking in general, a first name, initial, or rank is required in order for a mention to yield candidates. In addition to the candidates' names, the searcher retrieves their dates of birth and death, ranks, units they served in, and the sources where the information about the candidate originates. The positions of the candidates' ranks in the rank hierarchy are also fetched, as well as promotion dates.

The disambiguator then takes into account the names, ranks, lifespans, military units, and decorations of the candidates. As opposed to military unit and place linking where at least one of the candidates retrieved by the searcher was always selected for linking, the person disambiguator selects the top ranked candidate only if its score is above a specified threshold.

As a longer match is generally more specific, person candidates matching the longest piece of text are scored higher than those matching only a part of that text. The lifespan of the candidate is then compared to the date of the text. If the candidate has died before that date, the score is heavily reduced. On the other hand, the score of candidates that have died only a short while before the date are not reduced, as there are e.g. photographs depicting funerals.

Then, the ranks of the candidate are taken into account by comparing the rank mentioned in the surrounding text to the ranks of the candidate. As there are substantially more enlisted service personnel than there are high ranking officers, different ranks have different degrees of disambiguation power. For example, a reference to a general using their rank and family name is usually not ambiguous whereas a similar reference to a private is highly ambiguous. For this reason candidates were scored based on their rank: the higher the rank the better the score. The ranks of the candidates were also compared to the rank mentioned in the text, if any. If a rank was mentioned in the text but the candidate could not have had that rank at the time, the candidate's score is lowered. This includes soldiers who did not have such a rank, and those who were either only later promoted to that rank or already had a higher rank at the time. A reference to a person by their rank and family name was generally considered unambiguous enough to warrant a match in terms of the minimum threshold for scoring. However, as references to enlisted service personnel by their rank and family name are highly ambiguous without further information, a candidate matching, for example, "private Davis" would not receive enough points to result in a match unless the candidate scored highly in other aspects.

As initials are often used instead of the full first names of people in the texts, the format of the first names in a mention affects the candidates score. A full first name is better for disambiguation than an initial, and multiple initials is better than a single one.

A slight nod in the scoring is given to knights of the Mannerheim Cross (i.e. bearers of the Mannerheim Cross of Liberty), and people that were extracted from Wikipedia with the assumption that these people are well-known and are therefore more likely to be the correct match when the disambiguation is otherwise inconclusive. In case the surrounding text mentions the knighthood of the Mannerheim Cross, any candidates who are knights also receive a boost to their score.

The linked military units are also used when disambiguating people. If a unit has previously been identified in the text containing the person mention, candidates who have served in said unit receive additional points to their score. This was deemed possible based on the good recall and precision of the unit extraction itself.

In order to maximize recall, if multiple candidates received the same score that was also over the minimum threshold, all of them were selected rather than choosing one arbitrarily. This caused a dip in precision in some cases, where a mention of a kind generally considered unambiguous turned out to be ambiguous. For example, colonels are generally unambiguous but there are some with the same family name; in cases where there is no information by which to choose either, both are chosen. In the evaluated configuration, also an intermediate rank plus a family name were themselves deemed unique enough. However, this led to further false positives. Based on the evaluation, in later runs it could be beneficial to raise the bar in this regard to exclude linking based on just intermediate rank plus family name as well.

After all this, table 7 shows the precision and recall attained. The results are shown only for photographs and events, as the magazine article evaluations are currently pending.

Target	N	TP	FP	P	$FN_{ont.}$	$FN_{out.}$	$R_{ont.}$	R_{all}	$F_{1ont.}$	F_{1all}
Photographs	42	22	8	0.73	3	17	0.88	0.52	0.80	0.61
Events	34	26	0	1.00	7	1	0.79	0.76	0.88	0.87

Table 7. Precision and recall in person linking.

The results show that in the case of people, attaining even comparable results to the unit and place linking required a much more complicated process of disambiguation. Recall dropped sharply in the case of photographs when people outside the ontology were taken into account, whereas the recall in events stayed almost the same. This is because the event descriptions mention mostly well-known people or high-ranking officers, whereas there are photographs of all kinds of people from Russian prisoners of war to small children. As further work, it would be interesting to investigate in more depth also here where exactly the disambiguation errors arise from. It would also be interesting to see if disambiguation could be improved by handling the entity linking as a group,

and weighting more strongly those candidates who are known from background information to have links to each other.

7 Discussion

This paper discussed challenges encountered in NEL when applied to texts with mentions of military units, historical places, and person names. A key lesson learned during our work was that depending on the case, text, and data available, different approaches and methods are needed. For example, quite specific knowledge-based heuristics were needed in the case of the disambiguating military person names. Without resorting to such domain specific heuristics, the precision and recall would have remained unsatisfactorily low.

At the same time, there is an effort to encase such heuristics inside a common, configurable, modular pipeline. Of this pipeline, the language analysis, n-gram generation and search components[9] are currently the most well-developed [17], but also the preprocessing and disambiguation components have recently been made available as open source on GitHub[10].

Acknowledgements Our work is funded by the Open Science and Research Initiative[11] of the Finnish Ministry of Education and Culture, the Finnish Cultural Foundation, the Media Industry Research Foundation of Finland, and the Academy of Finland.

References

1. Bunescu, R.C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: EACL. vol. 6, pp. 9–16 (2006)
2. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: EMNLP-CoNLL. vol. 7, pp. 708–716 (2007)
3. Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)
4. Godoy, J., Atkinson, J., Rodriguez, A.: Geo-referencing with semi-automatic gazetteer expansion using lexico-syntactical patterns and co-reference analysis. Int. J. Geogr. Inf. Sci. 25(1), 149–170 (Feb 2011), http://dx.doi.org/10.1080/13658816.2010.513981
5. Gracia, J., Mena, E.: Multiontology semantic disambiguation in unstructured web contexts. Proceedings of the 2009 K-CAP Workshop on Collective Knowledge Capturing and Representation pp. 1–9 (2009)
6. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: Coling. vol. 96, pp. 466–471 (1996)

[9] https://github.com/jiemakel/arpa/

[10] https://github.com/SemanticComputing/python-arpa-linker, with the Warsampo configurations at https://github.com/SemanticComputing/warsa-linkers

[11] http://openscience.fi/

7. Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., Ball, J.: Use of the edinburgh geoparser for georeferencing digitized historical collections. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 368(1925), 3875–3889 (2010), http://rsta.royalsocietypublishing.org/content/368/1925/3875
8. 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
9. 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
10. 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
11. Hyvönen, E., Tuominen, J., Kauppinen, T., Väätäinen, J.: Representing and utilizing changing historical places as an ontology time series. In: Ashish, N., Sheth, A. (eds.) Geospatial Semantics and Semantic Web: Foundations, Algorithms, and Applications. Springer–Verlag (2011)
12. 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). pp. 758–773. Springer-Verlag (2016)
13. Kettunen, K., Mäkelä, E., Kuokkala, J., Ruokolainen, T., Niemi, J.: Modern tools for old content - in search of named entities in a finnish ocred historical newspaper collection 1771-1910. In: Proceedings of LWDA 2016 (September 2016)
14. Koho, M., Hyvönen, E., Heino, E., Tuominen, J., Leskinen, P., Mäkelä, E.: Linked Death - Representing, Publishing, and Using Second World War Death Records as Linked Open Data. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C. (eds.) The Semantic Web: ESWC 2016 Satellite Events. Springer-Verlag (June 2016)
15. Löfberg, L., Archer, D., Piao, S., Rayson, P., McEnery, T., Varantola, K., Juntunen, J.P.: Porting an english semantic tagger to the finnish language. In: Proceedings of the Corpus Linguistics 2003 conference. pp. 457–464 (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. 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)
18. Mäkelä, E.: LAS: an integrated language analysis tool for multiple languages. The Journal of Open Source Software 1(6) (oct 2016), http://dx.doi.org/10.21105/joss.00035
19. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
20. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering 27(2), 443–460 (2015)

21. The Association for Military History in Finland: Kansa taisteli lehdet 1957 - 1986 (2014), http://www.sshs.fi/sitenews/view/-/nid/92/ngid/1
22. 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/