People and Places of the Past - Named Entity Recognition in Swedish Labour Movement Documents from Historical Sources

Crina Tudor
Stockholm University, Sweden
first.last@ling.su.se

Eva Pettersson
Uppsala University, Sweden
first.last@lingfil.uu.se

Abstract

Named Entity Recognition (NER) is an important step in many Natural Language Processing tasks. The existing state-of-the-art NER systems are however typically developed based on contemporary data, and not very well suited for analyzing historical text. In this paper, we present a comparative analysis of the performance of several language models when applied to Named Entity Recognition for historical Swedish text. The source texts we work with are documents from Swedish labour unions from the 19th and 20th century. We experiment with three off-the-shelf models for contemporary Swedish text, and one language model built on historical Swedish text that we fine-tune with labelled data for adaptation to the NER task. Lastly, we propose a hybrid approach by combining the results of two models in order to maximize usability. We show that, even though historical Swedish is a low-resource language with data sparsity issues affecting overall performance, historical language models still show very promising results. Further contributions of our paper are the release of our newly trained model for NER of historical Swedish text, along with a manually annotated corpus of over 650 named entities.

1 Introduction

The Swedish labour movement is strong by tradition and has played a crucial role in the development of the Swedish welfare society and in shaping the structure of the labour market. The Swedish trade union federations play an important role internationally, and their archives offer a unique possibility to study the development of the trade unions and their key topics over time, and thereby also the social development nationally and internationally.

In the project *Labour's Memory. Digitization of annual and financial reports of blue-collar worker unions 1880-2020* , we aim to collect and digitize annual and financial reports from local, regional, national and international trade union organisations from 1880 onwards. The collection is to be stored in a database, and will be made searchable for people with an interest in diachronic labour movement documents through a user portal. This is achieved in collaboration between labour history experts, archivists, computational linguists and image analysis specialists.

In this paper, we aim to investigate to which extent current state-of-the-art models for Swedish can be used to extract named entities from historical sources, a key topic for enhanced searchability in the trade union documents. Secondly, we focus on maximizing usability for the intended end product by combining the strengths of different models. Last but not least, we evaluate the performance of these models in terms of accuracy, as well as F1 score. On a more practical level, we also release a new model that is fine-tuned for NER and trained on historical Swedish text, as well as a manually annotated gold-corpus of named entities extracted from Swedish labour union documents dated between 1892 and 1974.

The choices that we made in order to optimize the results and usability of our system were made in consultations with a group of experts from the Labour's Memory project, whose competence overlaps with that of the intended user.

2 Background

Named Entity Recognition (NER) is the process of automatically identifying and classifying name-like entities in text, such as names of persons, organizations and locations (Nadeau and Sekine, 2007). NER is an important subtask in many Natural Language Processing (NLP) applications, e.g., in information extraction/retrieval (see for example Brandsen et al. (2022)) and for anonymisation/pseudonymisation of sensitive personal data in a text (e.g. Bridal (2021) or Papadopoulou et al. (2022)).

For Swedish, researchers have recently worked

Proceedings of LaTeCH-CLfL 2024, pages 185–195

with developing a gold standard for Swedish named entity recognition (Ahrenberg et al., 2020), trying to merge and accommodate previous NER annotation schemes used for Swedish. There are also initiatives to adapt this standard to the task of annotating named entities in historical Swedish text, where the needs and features to consider differ slightly (Borin et al., 2007).

Outside the topic of NER itself, it is important to acknowledge that Swedish is still a low resource language, which does not have the same large-scale NLP infrastructure as other high-resource languages such as English, Spanish or Chinese. This is evident in terms of data sets, language models and tools, and even more so in the case of historical Swedish text. While there are efforts currently being made to build large language models for Swedish by organizations such as AI Sweden,[1] or historical resources from the side of SWE-CLARIN (e.g. Pettersson and Borin (2022)), it is still a tough undertaking to achieve high benchmark scores for NLP tasks on Swedish.

3 Method

The aim of our work is to perform Named Entity Recognition (NER) for trade union documents from the late 1800s and onwards, with the goal of enhancing searchability in the documents by automatically extracting metadata on persons, locations, organisations, events etc. An important subtask is therefore to adapt our tools to handling historical text, which is further elaborated on in Section 3.1. We move on by describing the different language models we use for the NER task in Section 3.2. In Section 3.3, we introduce the evaluation method that we use, and the gold standard created for this purpose.

3.1 Handling historical text

With the aim of improving the performance of our chosen language models on historical text, we apply several pre-processing steps in order to modernize the original text and bring it closer to the kind of text the readily-available models were originally trained on, as illustrated in Figure 1 and further described below.

The first step in our pipeline concerns abbreviations. The abbreviations used in historical times do not always follow the same standards as present-day abbreviations, meaning that NLP tools trained on contemporary sources may be confused by these. Therefore, we use a dictionary of abbreviations taken from Swedish family history sources[2] to automatically expand as many abbreviations as possible. Since our data contains a large amount of names in the form of *initial* + *surname* (e.g. *F. Linden*), we remove one-letter abbreviations from the list of abbreviations, in order to avoid confusion. For example, the list contains abbreviations for Swedish counties using one upper-case letter, such as 'F' for *Jönköping*, which would coincide with the 'F' in *F. Linden*, so expanding these would be counterproductive.

In the second step, we use the dictionary *Ordbok Öfver Svenska Språket* by Dalin to map historical Swedish spellings to their contemporary counterpart. This dictionary is available in digital format, with over 62,000 entries of words with their historical spelling mapped to its modern version (Borin et al., 2011).

Lastly, we perform spelling normalization of the words not covered by the Dalin dictionary. Spelling normalization is the process of automatically transforming historical spelling to a more modern, standard spelling. This can be done in several ways. In this paper, we choose to use a rule-based approach, with rules based on the Swedish spelling reform in 1906 and previously implemented in Pettersson (2016). The motivation for using this approach, is that since the documents are not several hundreds of years old, the spelling differences are rather modest and assumingly pretty well covered by a rule-based approach. Furthermore, the rules are already defined and described, and thereby easy to include in our pipeline.

Throughout the rest of the paper, when we mention normalization, we refer to the inclusion of all the pre-processing steps described above with the purpose of normalizing the original text.

3.2 NER modelling

After the historical texts have been pre-processed, the actual NER process takes place, through the use of language models. We try three off-the-shelf models for contemporary Swedish, and one model trained on historical Swedish text, as further described below.

As our point of departure, we choose an off-the-shelf model for contemporary Swedish developed

[1] https://www.ai.se/en

[2] https://www.familysearch.org/en/wiki/Sweden_Abbreviations_in_Family_History_Sources

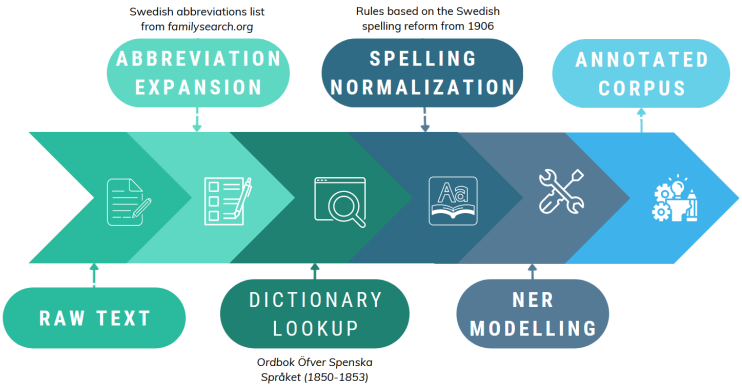

Figure 1: Handling historical texts among the labour movement documents.

by spaCy.[3] SpaCy is an open-source NLP software library which provides language models for over 65 languages, for a wide array of practical applications. In our case, we had three different pipelines to choose between which all can perform NER for Swedish. We ended up going for *sv_core_news_lg*, as that one had slightly better reported F1 score for the task at hand in comparison to the other two. This model is trained on data from news and media sources, as well as the Stockholm-Umeå corpus, v3.0 (SUC3), a balanced corpus of texts from different genres (Språkbanken, 2023).

The second model we try is also built on the spaCy infrastructure, but produced by the Swedish National Library.[4] The training data for this model largely overlaps with that of the previously mentioned model, but the reported F1 score is 4 percentage points higher.

The third model was selected after further investigating the work done by the NLP lab (i.e. KB lab) at the Swedish National Library. It is an updated version of the aforementioned model that makes use of Hugging Face[5] and their transformer architecture. For this model, they are experimenting with Hyper Parameter Optimization (HPO) leading to additional increases in F1 score for NER, reaching up to 91%.

Due to the fact that language models for historical text are in short supply, even more so when it comes to languages such as Swedish, there is no readily available model that can perform NER for historical Swedish text, to the best of our knowledge. In order to overcome this, we try a recently developed BERT model for historical Swedish text and fine-tune it using the same SUCX 3.0 corpus as the model developed by KBLab, so that we can run the same NER experiments and compare accordingly. The original model is released by the National Archives of Sweden and is using data from the 15th up to the 19th century, as well as the Hugging Face ecosystem. Our fine-tuned NER model is freely available to the public on Hugging Face.[6] The self-reported training statistics for our model are available in Appendix A.

For the sake of readability, we will refer to these models by an acronym for the rest of the paper, as follows:

DEF the default spaCy model for Swedish
BIB the spaCy model built by the Swedish National Library (i.e. Kungliga Biblioteket in Swedish)
KB the Hugging Face model developed by the KBLab group at the Swedish National Library
RA the model developed by the Swedish National Archives (i.e. Riksarkivet in Swedish) and fine-tuned for NER by us

We partially summarize the attributes of all the models that we experiment with in Table 1.

Model	Platform	Time	NER corpus
DEF	spaCy	Contemporary	SUC3.0
BIB	spaCy	Contemporary	SUC3.0
KB	Hugging Face	Contemporary	SUCX 3.0
RA	Hugging Face	15th-19th century	SUCX 3.0

Table 1: Summary of the models we use for NER.

[3] https://spacy.io/
[4] https://github.com/Kungbib/swedish-spacy
[5] https://huggingface.co/
[6] https://huggingface.co/crina-t/histbert-finetuned-ner

3.3 Evaluation

To be able to compare the performance of language models on equal grounds, we create a gold standard dataset that is manually labelled by a human annotator, and validated by a second annotator in order to settle eventual uncertainties. For the annotation, we use the same standard as presented by Borin et al. (2007). We use a total of 8 labels, as follows:

PRS "person" names of people
LOC "location" names of locations and other types of geographical entities
TME "time" temporal expressions
EVN "event" well-known events and celebrations
ORG "organisation" names of corporations and other kinds of organisations
WRK "work of art" names of movies, sculptures, periodicals etc.
MSR "measure" numerical expressions, such as monetary expressions or distances
OBJ "artifact" names of food/wine products, vehicles etc.

For the gold corpus, we select a total of 50 pages of sample text. In order to account for the shift in spelling conventions introduced through the 1906 Swedish spelling reform (Jansson, 2023), we select 25 pages which are dated before 1906, and the remaining 25 pages from years dated post-reform. To the best of our ability, we attempt to make sure that these pages are equally spaced out in terms of time elapsed between the dating of each one, as well as that they contain a reasonable body of text (i.e. at least half a page), and not just a few lines.

It can be noted that quite general phrases, not referring to a well-defined point in time, such as *under året* 'during the year' or *de senaste åren* 'in recent years' are labelled as time expressions ('TME') by most models. However, we choose to omit them from our analysis since our group of experts deem them irrelevant. We therefore only keep those that contain numerical expressions and/or names of months or their respective acronyms (e.g. *December* or *dec*). We take a similar approach when it comes to other kinds of named entities as well in the cases where they are too vague and do not point to a specific, individual entity (e.g. *förbund* 'trade union' is too vague, but *Typografförbundet* 'Typographers' Union' would be included in the manual annotation).

Other than the aforementioned 50 pages, we annotate an additional 10 pages from the same time span as the original gold standard (i.e. 1892–1974). We do this in order to be able to evaluate our final hybrid approach on unseen data so that we can more accurately assess its performance, following the same principles for data selection and annotation as the gold corpus.

In total, our gold corpus contains 570 manually annotated entities, plus an additional 85 entities in the test set, which we summarize by label in Table 2. Both of these are freely available to the public on Hugging Face.[7] We mention here that a total of 35 entities representing names of people were replaced with a placeholder in the released version of the corpus at the request of the archive in order to comply with their privacy policy. The documents containing placeholders are clearly pointed out in the description of the dataset.

	Gold set	Test set
EVN	19	6
LOC	86	19
MSR	97	7
ORG	71	13
PRS	162	17
TME	134	22
WRK	1	1

Table 2: Label count for manually annotated entities.

When comparing the gold standard with the automatically extracted entities, we identified some consistent differences between the different kinds of matches we encountered. For this reason, we create 8 distinct categories to differentiate between them in the evaluation phase. While this is a more fine-grained evaluation when compared to what is more widely used in the field (e.g. Chinchor and Sundheim (1993), Tjong Kim Sang and De Meulder (2003) or Segura-Bedmar et al. (2013)), we believe that our evaluation schema can greatly help in easily identifying the source of prediction errors generated by the model. We define and exemplify these categories below:

- **EXM** – exact match
  Both the entity and the label overlap exactly between the model and the annotator.
- **PAM** – partial match

[7] https://huggingface.co/datasets/crina-t/UnioNER

A substring of the gold standard entity is identified by the system, with the same label - e.g. *J.E Blomkvist* (PRS) in the gold standard, while the system outputs *Blomkvist* (PRS).

- **ENM** – entity match
  The exact same string is annotated by both the annotator and the system, with different labels (e.g. manually annotated *Harg* (LOC) is output as *Harg* (PRS) by the system).
- **VAM** – vague match
  The system predicts part of the gold standard entity, but with a different label - e.g. the annotator would label *E Lund* (PRS), while the system outputs *Lund* (LOC).
- **COM** – compound match
  The system merges several entities that the annotator identified as being separate units. E.g. *Hilmer Johansson* (PRS) and *Ernst Hörngren* (PRS) in the gold standard are predicted as *Hilmer Johansson Ernst Hörngren* (PRS).
- **SPM** – split match
  The system splits an entity from the gold standard into separate entities (e.g. gold standard *Uppsala Typografiska Förening* (ORG) vs. the automatically annotated *Uppsala* (LOC) and *Typografiska Förening* (ORG), referring to the Uppsala branch of the Typographers' Union).
- **FP** – false positive
  The system labels a unit that is not in fact a named entity.
- **FN** – false negative
  The system fails to identify an entity that the annotator has identified.

4 Results

In order to identify which approach gives us the best results, we calculate the accuracy of each model, and in doing so we also look at the count for each type of match per individual model. By accuracy we mean the percentage of entities from the gold standard which have a counterpart in the output predicted by the model, regardless of the type of match. The matches we look at are detailed in Section 3.3, and we do not include the counts for FP and FN in calculating accuracy. We also calculate the (accuracy) count for all the different kinds of matches, as well as NER labels, which we present in Figures 2 through 10. We calculate both these metrics first on the original text, then on the normalized version of the text in the gold standard. Our goal is to reach as many exact matches as possible (at best) and to minimize the number of cases where the model returns no match for a given entity in the gold standard (FN).

As a secondary step, we provide calculations for F1 score as applied to our text, since F1 score is a widely used metric for NER, and which also accounts for the FP and FN tags in our case. We present our results for F1 score in Section 4.4

4.1 Baseline

After applying the default spaCy model (DEF) to historical text, we obtain an accuracy of only 57.54%, due to the fact that the number of entities that are left unannotated (i.e. FN) by the DEF model exceeds the number of exact matches. In this case, normalization is detrimental to the model's performance, bringing accuracy down to 51.57%. Given the fact that more than half of our entities are not recognized by the system, we opt out of using this model further.

Moving onto the next spaCy model developed by the Swedish National Library, we immediately notice an improvement in performance. The BIB model predicts significantly more exact and partial matches than its predecessor, while the occurrence of false negatives is also decreased by almost 30% when compared to the DEF model, leading to an overall accuracy of 69.64%.

Normalization does help in this case, but it is barely enough to get the model over the 70% mark. We suspect that this jump in performance between the DEF and the BIB model is due to the fact that the DEF model was trained mainly on news and media text, while the BIB model was trained on a more balanced corpus of text from different sources. While this is clearly an improvement, it is not satisfactory enough to motivate using this model for our purposes.

The Swedish National Library KB model (Kurtz and Öhman, 2022) is an updated version of the previous BIB model. In comparison with its predecessor, the KB model, while not much better at extracting exact matches, does help decrease the number of FNs, as demonstrated in Figure 2. This happens as a result of the increase we see when it comes to partial matches, and at a smaller level in entity matches, vague matches, compound matches and split matches. The overall accuracy for the model lands at 77.34%, with normalization ebbing this number by only 0.29%.

Lastly, we evaluate the performance of our fine-

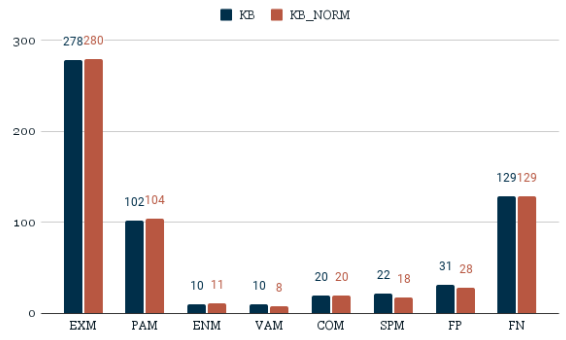

Figure 2: Accuracy count for the Hugging Face model from the Swedish National Library. KB = original text, KB_NORM = normalized text.

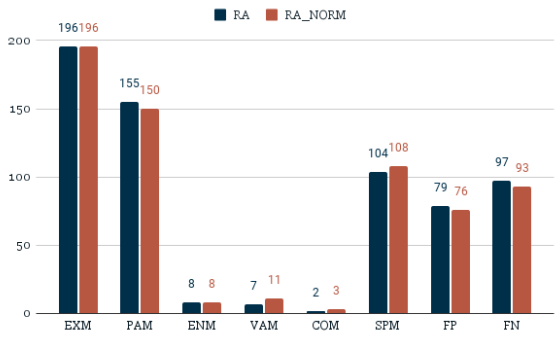

Figure 3: Accuracy count for the model from the Swedish National Archives. RA = original text, RA_NORM = normalized text.

tuned RA model, which is based on a model created by the National Archives of Sweden and trained exclusively on historical data, but fine-tuned for NER on the same SUCX3.0 corpus as the previously evaluated KB model. In this case, we do expect an increase in performance due to the fact that the model is trained on data from the same time period as our gold standard corpus, as opposed to the aforementioned models which are all trained on contemporary text.

When conducting the analysis of this model, we were surprised to see that the number of exact matches dropped significantly. Even more surprising is the fact that despite this, the model shows the least amount of false negatives among all the models we evaluated, which means that it manages to capture (to some extent) about 80% of the entities from the gold corpus. After normalization, the overall accuracy of the model reaches up to 83.41%, which is the highest we were able to reach in our experiments. However, it is worth noting that this high accuracy does not account for the doubled amount of false positives compared to previous models, or the staggering increase in split matches. We investigate this further in Section 4.2.

We summarize the accuracy of each model in Table 3.

Model	Original	Normalized
DEF	57.54%	51.57%
BIB	69.64%	70.35%
KB	77.34%	77.05%
RA	82.60%	83.41%

Table 3: Overall accuracy for each NER model.

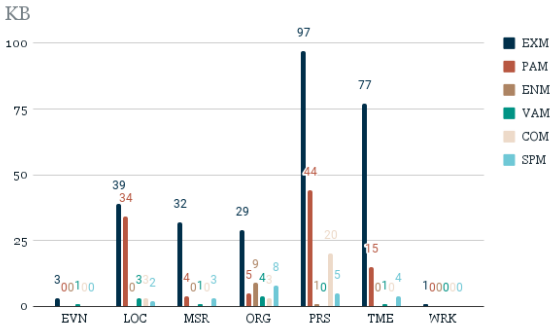

Figure 4: Accuracy count per label for the Hugging Face model from the Swedish National Library, applied to the original text.

4.2 A closer look

After the evaluation described in Section 4.1, it is clear that our two front runners are the KB and the RA models. Were it not for the notable increases in split matches and false positives, the RA model would take precedence, but as it stands, it is worth investigating more in-depth what could be causing these fluctuations. We therefore take a closer look at the way it performs for each individual label, while also extracting the same statistics for the KB model as a point of comparison.

Figure 4 shows how the KB model consistently extracts more exact matches for all different label types, with partial matches being a close second. Same pattern is visible after normalization as well, which can be observed in Figure 5, with slight improvements in exact matches for the person (PRS) and organisation (ORG) labels.

In the case of the RA model, there is a clear decrease in the number of exact matches across labels, and significantly more split matches, as shown in Figure 6. Among all the labels, time expression

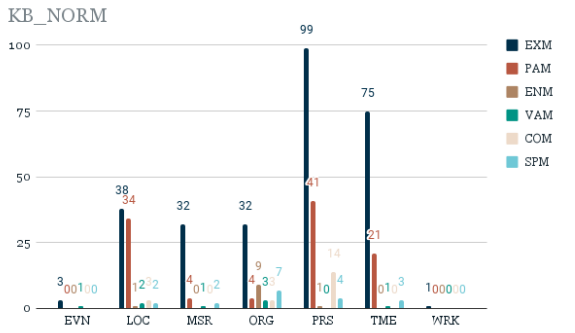

Figure 5: Accuracy count per label for the Hugging Face model from the Swedish National Library, applied to the normalized version of the text.

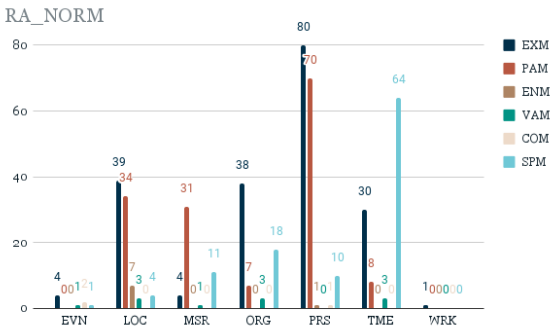

Figure 7: Accuracy count per label for the model from the Swedish National Archives, applied to the normalized version of the text.

(TME) is the one that is most affected by the high number of split matches, which we suspect is due to a tokenization issue. There is a high chance that, since the original model is trained on a smaller corpus, there were not enough numerical values for the model to know how to handle years, dates etc. From Figure 7, we can see that the trend remains the same even after normalization, which reinforces the theory that the errors are stemming from the tokenization process.

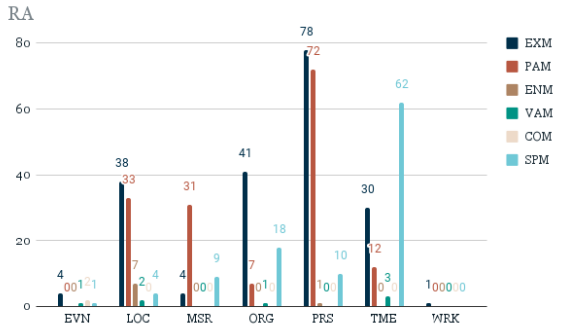

Figure 6: Accuracy per label for the model from the Swedish National Archives, applied to the original text.

4.3 A hybrid approach

After looking in-depth at the strengths and weaknesses of the KB and RA models, we want to investigate to which extent combining their outputs could benefit the end results. More specifically, we aim to avoid overgenerating split matches in the RA model and to try to increase the accuracy of the KB model. For this reason, we prioritize high counts of exact and partial matches, and take specific labels from the RA model (PRS, ORG), merging them with the rest of the labels from the KB model.

By doing this hybrid approach (HYB), we reduce the number of false negatives that were initially present in the KB model, and we also manage to drop the number of split matches that were problematic for the RA model, as shown in Figure 8.

From Figures 9 and 10, we can clearly see that this approach is beneficial in reducing the number of split matches for those categories that are prone to having numerical expressions, such as MSR and TME, and which the RA model could not handle very well.

For accuracy, we obtain an increase from the KB model, reaching 79.82% on the original text, which drops slightly after normalization - to 79.47%.

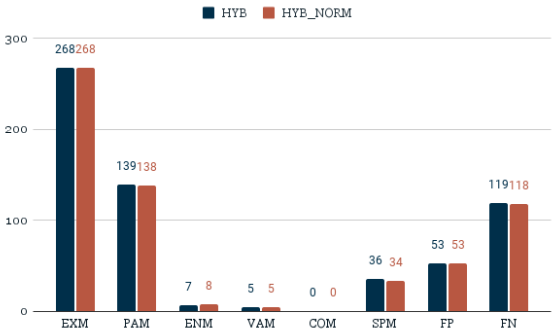

Figure 8: Accuracy count for the hybrid approach. HYB = original text, HYB_NORM = normalized version text.

4.4 F1

As a last evaluation step, we calculate F1 score for the KB, RA and HYB models using seqeval (Nakayama, 2018), which is presented in Table 4. The reason why we do this is two-fold - on the one hand, we want to assess performance on new, unseen data, while on the other hand we want to see

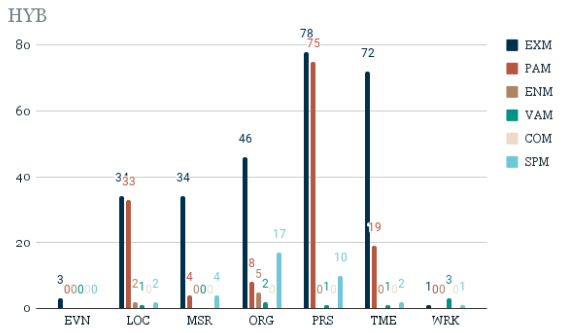

Figure 9: Accuracy count per label for the hybrid approach, applied to the original text.

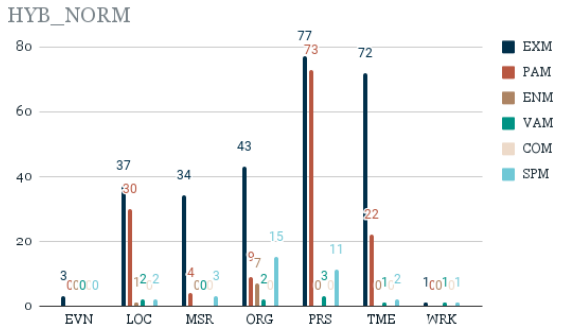

Figure 10: Accuracy count per label for hybrid approach, applied to the normalized version of the text.

how the score is affected by a metric that accounts for false negatives and false positives.

Model	Gold	Test
KB	75.68%	66.23%
KB_NORM	76.79%	71.14%
RA	71.23%	67.36%
RA_NORM	71.68%	66.66%
HYB	76.25%	68.23%
HYB_NORM	76.66%	71.85%

Table 4: F1 score for the KB, RA and HYB models, before and after normalization, from the gold corpus as well as the test set.

Not surprisingly, the RA model performs the worst in this case, due to the high volume of false positives that it predicts. Even though it has a lower false negative count, this is not enough to counterbalance the effect of PAM where the predicted entities did not match the gold standard entirely, or the split entities where the gold entity span was split into several different ones.

It is however interesting to see that the HYB model overtakes the KB model on almost all accounts, and while normalization did not help for accuracy, it does increase F1 score in this case. The HYB model with normalization manages to obtain the highest score on the test set at 71.85%.

5 Discussion

Our study focuses on the application of NER to historical Swedish text, specifically documents sourced from Swedish labor unions dating back to the 19th and 20th centuries. The primary challenge lies in adapting contemporary state-of-the-art NER systems to effectively process and extract entities from historical text, which often differs significantly in linguistic norms, vocabulary, spelling, and syntactic structures from contemporary Swedish.

Our research delves into a comparative analysis of multiple language models applied to NER for historical Swedish text. Three off-the-shelf models designed for contemporary Swedish text were experimented with, alongside a custom-built language model trained on historical Swedish text. This unique approach allowed us to explore the adaptability of existing models and assess the feasibility of fine-tuning historical language models for NER tasks. Through our experiments, we show that current off-the-shelf models have the capability to extract named entities from historical text, but at the same time they can benefit from training on historical data, as shown by the high accuracy of our RA model.

Moreover, we believe that the inconsistent effect of the normalization rules could be partly due to the rather small amount of normalization rules, as well as the nature of a rule-based approach to spelling normalization, where it is hard to write efficient rules without risking overgeneration. Another, more data-driven approach, might have given more consistent results.

It is also important to keep in mind that our evaluation metrics were customized according to the needs of our future users. Since our target groups are looking for as many named entities as possible, we attempt to adapt our approach in order to maximize the usability for the end product - which is the archival database of the Labour's Memory project, while at the same time maintaining a good level of quality for the automatically extracted named entities.

For future work, we would like to investigate the way different data augmentation methods can improve our results, since previous work done on

English text shows promising results when it comes to applications on pretrained transformer models (see, for example, Dai and Adel (2020)), such as the RA model we propose in this paper. Moreover, given the fact that our source material comes from labour union documents, it could also be interesting to look at a more fine-grained analysis of the PRS label in order to be able to identify potential biases in NER systems - similar to the work conducted by Lassen et al. (2023) for Danish text.

6 Conclusion

In this paper, we show that current off-the-shelf models for Swedish can perform NER on historical text, but using a historical language model shows more promising results. However, data from historical sources could also be beneficial for training in order to achieve better F1 score and reduce errors. An alternative path we would like to explore in the future is training on multilingual data from other Scandinavian languages, given that multilingual models show great promise when it comes to cross-lingual transfer learning (see, for exasmple, Chi et al. (2020) or Katsarou (2021)), with the added bonus that Scandinavian languages have similar vocabulary and structure.

A significant contribution of this study lies in the release of the newly trained RA model tailored for NER of historical Swedish text. Additionally, we introduce a manually annotated corpus comprising over 650 named entities, offering a valuable resource for future research endeavors. We also show that combining the strengths of multiple models can be beneficial to our NER task.

In conclusion, while our study provides valuable insights and tools for NER in historical Swedish text, it also underscores the necessity for further advancements and novel methodologies to address the challenges posed by data sparsity in low-resource languages.

Acknowledgements

We would like to extend our heartfelt gratitude to our colleagues in the Labour's Memory project for their invaluable historical expertise and support.

This work has been supported by Riksbankens Jubileumsfond, grant IN20-0040, *Labour's Memory. Digitization of annual and financial reports of blue-collar worker unions 1880-2020*.

References

Lars Ahrenberg, Johan Frid, and Leif-Jöran Olsson. 2020. A new gold standard for Swedish named entity recognition: Version 1 contents. SWE-CLARIN Report Series SCR-01-2020.

Lars Borin, Markus Forsberg, and Christer Ahlberger. 2011. Semantic Search in Literature as an e-Humanities Research Tool: CONPLISIT – Consumption Patterns and Life-Style in 19th Century Swedish Literature. In *NEALT Proceedings Series (NODALIDA 2011 Conference Proceedings)*, volume 11, pages 58–65.

Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. 2007. Naming the past: Named entity and Animacy recognition in 19th century Swedish literature. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, pages 1–8, Prague, Czech Republic. Association for Computational Linguistics.

Alex Brandsen, Suzan Verberne, Karsten Lambers, and Milco Wansleeben. 2022. Can bert dig it? named entity recognition for information retrieval in the archaeology domain. *J. Comput. Cult. Herit.*, 15(3).

Olle Bridal. 2021. Named-entity recognition with BERT for anonymization of medical records.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Xiang Dai and Heike Adel. 2020. An Analysis of Simple Data Augmentation for Named Entity Recognition.

Martin Jansson. 2023. *Samtidens gränser : Om språkreformer och historisk tid runt sekelskiftet 1900*. Ph.D. thesis, Uppsala University, Department of History of Science and Ideas.

Styliani Katsarou. 2021. Improving Multilingual Models for the Swedish Language : Exploring CrossLingual Transferability and Stereotypical Biases. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS).

Robin Kurtz and Joey Öhman. 2022. The KBLab Blog: SUCX 3.0 - NER.

Ida Marie S. Lassen, Mina Almasi, Kenneth Enevoldsen, and Ross Deans Kristensen-McLachlan. 2023. Detecting intersectionality in NER models: A data-driven approach. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics*

for Cultural Heritage, Social Sciences, Humanities and Literature, pages 116–127, Dubrovnik, Croatia. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. Neural text sanitization with explicit measures of privacy risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 217–229, Online only. Association for Computational Linguistics.

Eva Pettersson. 2016. *Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction*. Ph.D. thesis, Uppsala University, Department of Linguistics and Philology.

Eva Pettersson and Lars Borin. 2022. *Swedish Diachronic Corpus*, pages 561–586. De Gruyter, Berlin, Boston.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Språkbanken. 2023. SUCX 3.0 - Stockholm-Umeå corpus 3.0, scrambled.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Appendix A. Self-reported raining results for the RA model

Training Loss	Epoch	Step	Validation Loss	Precision	Recall	F1	Accuracy
0.0403	1.0	5391	0.0316	0.8496	0.8866	0.8677	0.9903
0.0199	2.0	10782	0.0308	0.8814	0.9034	0.8923	0.9915
0.0173	3.0	16173	0.0372	0.8698	0.9197	0.8940	0.9913
0.0066	4.0	21564	0.0397	0.8783	0.9239	0.9005	0.9921
0.0029	5.0	26955	0.0454	0.8855	0.9181	0.9015	0.9923
0.0035	6.0	32346	0.0454	0.8834	0.9211	0.9019	0.9922
0.0009	7.0	37737	0.0495	0.8784	0.9261	0.9017	0.9922