A New Resource for Swedish Named-Entity Recognition

Lars Ahrenberg
Computer and Information Science
Linköping University
lars.ahrenberg@liu.se

Johan Frid
Humanities Lab
Lund University
johan.frid@humlab.lu.se

Leif-Jöran Olsson
Department of Swedish
University of Gothenburg
leif-jöran.olsson@svenska.gu.se

Abstract

We present a new gold standard resource for Swedish Named-Entity Recognition and Classification. Existing free resources are either domain-specific or represent Swedish as it was some 30 years ago. This resource includes texts from different genres including social media. It has been developed by a special working group in the Swe-Clarin consortium.

1 Introduction

Named-Entity Recognition and Classification (NERC) is a standard task in natural language processing needed for many applications. While new methods for NERC are being developed, there is a lack of Swedish data for training and evaluation. For Swedish, the most recent publicly available gold standard resource is the Stockholm-Umeå Corpus (SUC) which reflects the language of the late 20th century. Thus, SUC contains no data from social media.

In this paper we present a new NERC resource, with text from eight different genres sampled from documents published in and around the year 2010. The work has been done as part of activities in the Swedish CLARIN consortium[1], *Swe-Clarin*, by a special project group. The invitation to participate was open, and a couple of open meetings were held in the beginning of the project. After this initial phase the three authors of this report have been the main actors. The decisions reported here have been made by us in meetings and discussions, mostly over the Internet.

In this report we describe the work done so far and the guidelines developed as part of the effort. We also report some evaluation results to be used as benchmarks.

1.1 Aims

Currently available gold standards for Swedish NERC either represent language from the 1990ies or a single genre. With this resource we wished to include more recent language, in particular as it is used in social media. As NERC has gained increased interest in medical application over the years, we also included two semantic types from the medical domain. Still, our aim is far from producing an all-embracing resource; rather, the scope is limited to eight different categories that provide different challenges for automatic named-entity recognition.

The aims have been:

- to provide a free resource for research and development
- to provide at least 1000 instances for each selected category
- to select categories that are relevant and at the same time provide challenges of different kinds for developers
- to develop detailed criteria and guidelines for the categories that can be distributed with the resource
- to base the resource on annotations from three different annotators with known inter-annotator agreement
- to provide benchmarks on the basis of state-of-the-art software

2 The Data

To guarantee that the resource can be distributed freely the data have been taken from resources that are already available from Språkbanken Text[2]. This has meant that the sentences are often scrambled,

[1] https://sweclarin.se/

[2] SIC, the Stockholm Internet Corpus, is an exception. It is created by Robert Östling at Stockholm University and can be found at https://www.ling.su.se/english/nlp/corpora-and-resources/sic/stockholm-internet-corpus-sic-1.99019

but the resource also contains data from unscrambled documents. The majority of the texts included were produced in or around 2010.

The corpus covers the following genres:

- Bloggmix, blog texts on the life of a young person, (12 samples; 29,052 tokens; 1,079 instances; scrambled)
- Familjeliv-barnhälsa, a social forum on children's health, (11; 25,867; 768; scrambled)
- Flashback-fordon, a social forum on cars and other vehicles, (10; 23,824; 940; scrambled)
- SIC, blog texts on everyday activities, (16; 9938; 397; unscrambled)
- Göteborgsposten, news text, (10; 20,971; 1549; scrambled)
- Smittskydd, a medical journal on health protection, (7; 16,927; 928; scrambled)
- Wikipedia-krig, Wikipedia texts on war history, (6; 15,855; 1815; unscrambled)

The aim to include categories that provide challenges of different kinds to automatic systems resulted in a set of eight categories. The general, ubiquitous categories Person, Location, and Organisation are included, but the set also includes categories with more variation in their expressions, such as Events, Times, and Works of Art, and domain specific categories, such as Symptom and (medical) Treatment. An overview of the categories with frequncy information is shown in Table 1. There is also additional data which has been annotated automatically but not manually checked.

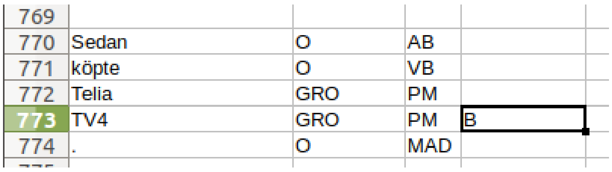
The corpus has actually been developed incrementally, as we found that certain categories were not well represented in the first selection of data sources. This motivated the addition of the Smittskydd- and Wikipedia-krig corpora.

Documents have been sampled with a size of 2000-2500 tokens and formatted in a spreadsheet. The spreadsheets have three columns, one for tokens, one for the named-entity tag, and one for a part-of-speech. An example is shown in Figure 1.

The fourth column is used for keeping track of instances that follow after one another as in *Sedan köpte Telia TV4.*. In those cases a 'B' is added in the fourth column for the token that begins a new instance.

Category	**Tokens**	**Instances**
Event (EVN)	564	275
Organisation (GRO)	1910	1439
Location (LOC)	1205	1031
Treatment (MNT)	267	197
Person (PRS)	2186	1323
Symptom (SMP)	1248	792
Time entity (TME)	2806	1377
WorkOfArt / Product (WRK)	2081	1042
Sums	12267	7476
Other (O)	130239	-

Table 1: The eight entity types of the resource with their frequencies.



769				
770	Sedan	O	AB	
771	köpte	O	VB	
772	Telia	GRO	PM	
773	TV4	GRO	PM	B
774	.	O	MAD	

Figure 1: Data as presented to annotators. Necessary changes are made in the second column.

2.1 Benchmarking

Three types of systems have been trained on the data, a dictionary-based system, a system using Stanford's CRF-NER system (Finkel et al., 2005), and a system using contextualized representations. We have evaluated at the level of both tokens and instances. The best results were produced by the system employing contextualized representations, so those results are the ones we present here.

For this system we used the pre-trained BERT language model for Swedish developed at KB-Lab (Malmsten et al., 2020) and plugged this into a NER system developed by Kamal Raj[3]. The model was fine-tuned on different subsets of our NER data with default parameters.

For cross-validation, the data set was split into training and test sets in six different ways. Each split contained one sample from each sub-corpus with all the remaining samples used as training data. No sample was used as test data more than once. Table 2 shows the results for NE instances.

Performance is uneven across categories and genres. The Person category gives the best results with F1 at 0.949 while WRK (Products and Works of Art) has the lowest results with F1 at 0.525. In another round of testing we used all documents from

[3] https://github.com/kamalkraj/BERT-NER

Split	Precision	Recall	F1
Split 1	0.8255	0.8009	0.8099
Split 2	0.8127	0.8121	0.8110
Split 3	0.8201	0.8070	0.8119
Split 4	0.8312	0.8186	0.8242
Split 5	0.8219	0.8168	0.8182
Split 6	0.8266	0.8408	0.8327
Averages	0.8230	0.8160	0.8180

Table 2: A summary of results for the BERT NER-system on different splits of the data.

one sub-corpus as test data and trained on the rest. Results decreased by some 10 points on precision and even more on recall. Detailed results will be published with the data.

We observe that the results are about 10 points below the results obtained by (Malmsten et al., 2020) for their NER system. Results cannot be directly compared, however, as they used SUC 3.0 for training, a corpus which is larger than ours, manually tokenized and does not categorize for Symptoms and Treatments. For the common categories of Person and Temporal entities the differences are smaller: 0.961 vs 0.949 for Person, and 0.906 vs. 0.898 for Times.

2.2 Availability

The resource will be available as *Swe-NERC Version 1* under a CC-BY license from the resources page of Språkbanken Text and their CLARIN repository. The resource will also contain additional data from the same sources with automatic annotations from our BERT-NER model.

3 Guidelines

Every token of the texts carries a tag indicating its status as part of a phrase referring to a named-entity. Three letter abbreviations are used for the different NE-types, see Table 1, while tokens outside of a naming expression are marked O. A token must not carry more than one tag, cf. (Chinchor, 1997)). The definition of a NE-type is primarily semantic: what kind of entity it is referring to in the context where it occurs.

Manual annotation was performed by changing the proposed named-entity tag, when it is found to be erroneous. Abbreviations for the categories were chosen so that it would normally be sufficient to press the key for the first letter to change the tag. That's why the abbreviation GRO rather than ORG was used for organisations.

The following general principles have been followed: A naming expression is a syntactic phrase of some sort, that is an established standard reference for an entity, or including such a standard reference as its main part. A naming expression may thus include words that are not proper nouns but are rather referring to attributes of the referent. Pronouns, such as *han*, *hon*, deictic adverbs such as *då*, *här*, and verbs should as a rule be marked 'O'. Exceptions can be found with works of art, events, and time expressions.

Tokenization has been automatic and could not be changed by annotators. This means that the data often break rules of Swedish orthography.

When a longer NE-expression includes a shorter one, all tokens carry the NE-tag of the longer phrase. Thus both tokens of an organisation name such as *Uppsala universitet* are annotated GRO.

Genitive forms are marked in the same way as nominative forms.

In the course of the project the annotation guidelines have been revised several times.[4]

The guidelines for the categories Symptom and Treatment are modelled after the 2010 i2b2 / VA Challenge Evaluation (TranSMART Foundation, 2010).

Annotators	Kappa
1, 2, 3	0.88
1, 2, 4	0.87
1, 2, 3, 4	0.78

Table 3: Inter-rater agreements, measured by Fleiss' kappa before final decisions were made.

Inter-rater agreements have been checked on several occasions. Initially, they were made with tight intervals including discussions of problematic examples in between. Before producing the final annotations to be included in the resource, inter-rater agreements for all annotators were computed again. One annotator was found to be deviating greatly from the others and we decided to discard those annotations in the final phase. Inter-rater agreements, using Fleiss' kappa, for the different sets of annotators are shown in Table 3.

[4] We are indebted to Patricia Lindblad and Arne Jönsson for annotation work and discussion of the guidelines, to Elena Volodina for giving us the anonymisation guidelines of the SweLL project, and to Stian Rødven Eide for pointing out missing information in earlier versions of the guidelines.

The final annotations were produced using a spreadsheet where the available annotations, at least three for every token, were set side-by-side. One annotator was appointed for each sub-corpus to check disagreements against the guidelines one more time. In case a disagreement is a matter of interpretation, and not clearly specified in the guidelines, majority voting was applied. If that did not resolve the issue, the appointed annotator made the decision.

The guidelines are published in a Swe-Clarin report, (Ahrenberg et al., 2020).

4 Relation to previous work

The first larger gold standard for named entities in Swedish text was the Stockholm-Umeå corpus (SUC), which was supplied with named-entity annotation for its second version (Gustafsson-Capková and Hartmann, 2006). In the most recent version, SUC3.0, the annotation has been checked further. Formally, named entities are marked using the start tag <name> and its corresponding end tag </name> with the first carrying an attribute, *type*, to indicate the entity type. The types used are: person, animal, myth (ological entity), place, inst (itutional entity), product, work (of art), event, and other.

In addition, numbers are identified as a separate kind of entity-referring expression. The distribution is uneven over the categories with numbers having the most (18098), and events the fewest (245).

SUC2.0 was used by Salomonsson et al. (2012) to build a four-split system, where the categories animal, myth, inst, product, event and other were merged into a miscellaneous category.

4.1 NomenNescio

A joint Nordic project developed a common framework for NERC on Scandinavian languages using six categories: PRS (Person), LOC (Location), ORG (Organization), EVT (Event), WRK (Work of Art), and OTH (Other) (Johannessen et al., 2005). The project compared and evaluated several methods, both manually and automatically on available gold standards. A conclusion of the project was the importance of gazetteers for achieving good performance.

4.2 SweNER

In the context of the NomenNescio project, Kokkinakis (2004) developed a NERC-system for a comprehensive taxonomy of types with eight top level types and altogether 47 subtypes. The top categories were: location, person, organisation, event, object, work and art, time, and measure. The object category covers products of various kinds but also prizes and, along with medical products, also names of diseases and genes. It is kept separate from the work and art category which, apart from works of creation also covers such products as newspapers.

The system was evaluated on a dataset of edited texts from different genres including newspaper texts of various kinds and excerpts from literature. The evaluation was performed on a token basis with an average precision of 0.9422 on all types. Surprisingly, including the subtypes in the evaluation decreased the results with only 0.7%.

The SweNER system of 2004 were largely based on rules and large lists of relevant names and multiword phrases. It has later been developed and reimplemented for different tasks (e.g., Borin and Kokkinakis, 2010). A major re-implementation is the HFST-SweNER which used the same eight categories as the previous system, but an enlarged set of subtypes (Kokkinakis et al., 2014). This time the system was evaluated on the SUC3.0 gold standard. However, due to the fact that the categories are not always one-to-one, some measures of harmonisation and re-mapping were needed. Although it could be shown that the output from HFST-SweNER overlapped with that of SweNER with only minor differences (1-2% of tokens), the performance this time was poorer with an average precision of 79.02% and average recall at 70.56%.

The web service for named-entity recognition (Sparv) provided by Språkbanken Text is based on SweNER providing the eight top level entity types and several subtypes.

It can be seen that our choice of categories is not an exact continuation of previous projects. However, there is overlap in the standard classes person, location, and organisation and also in the event class. For other classes, combining resources may require a closer look at definitions.

References

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

Lars Borin and Dimitrios Kokkinakis. 2010. Literary onomastics and language technology. In *Literary education and digital learning*, pages 53–78. Information Science Reference.

Nancy Chinchor. 1997. Muc-7 named entity task definition. https://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html.

Jenny R. Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.

Sofia Gustafsson-Capková and Britt Hartmann. 2006. Manual of the Stockholm Umeå Corpus version 2.0. `https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf`.

Janni Bondi Johannessen, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdottir, Anders Noklestad, Dimitris Kokkinakis, Paul Meurer, Eckhard Bick, and Dorte Haltrup. 2005. Named entity recognition for the mainland Scandinavian languages. *Literary and Linguistic Computing*, 20:1:91–102.

Dimitrios Kokkinakis. 2004. Reducing the effect of name explosion. In *Proceedings of the LREC Workshop: Beyond Named Entity Recognition, Semantic labelling for NLP tasks. ourth Language Resources and Evaluation Conference (LREC)*.

Dimitrios Kokkinakis, Jyrki Niemi, Sam Hardwick, Krister Lindén, and Lars Borin. 2014. HFST-SweNER — a new NER resource for Swedish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2537–2543, Reykjavik, Iceland. European Language Resources Association (ELRA).

Martin Malmsten, Love Börjesson, and Chris Haffenden. 2020. Playing with words at the National Library of Sweden – making a Swedish BERT. https://arxiv.org/abs/2007.01658.

Andreas Salomonsson, Svetoslav Marinov, and Pierre Nugues. 2012. Identification of entities in Swedish. In *SLTC 2012 : The Fourth Swedish Language Technology Conference*, pages 63–64. SLTC.

i2b2 TranSMART Foundation. 2010. 2010 i2b2/VA challenge evaluation: Concept annotation guidelines. `https://www.i2b2.org/NLP/Relations/assets/ConceptAnnotationGuideline.pdf`.