Northern Sámi: A Case Study in Low Resource NER

Erik Andersen {erikandersen,
Julia Cathcart juliacathcart,
Kuan-Yu Chen chenky}@brandeis.edu

Abstract

In this paper, we compare a low resource Named Entity Recognition system trained on a small dataset of news articles in the Northern Sámi language with a system of the same size trained on English data. Both of these systems use the same four main entity types found in the CONLL dataset: ORG, PER, MISC, and LOC. We found that the system trained on Northern Sámi data performed comparably to the English data when augmented with pre-trained Sámi-Finnish word embeddings and a name list for Sámi given names.

1 Introduction

This paper aims to present a Named Entity Recognition system in a low resource language in order to examine the challenges associated with performing entity recognition on languages with only a small amount of available data. In other words, we attempted to find out if it is possible to create a state-of-the-art entity recognition system when we only have limited data available. If it were possible to create a low resource system that is comparable to a Named Entity Recognition system trained on the same amount of English data, this would be revolutionary for entity recognition on languages with only a small amount of data available for feature extraction, as we would be able to extend the applications for Named Entity Recognition across many more languages without having to create extra data. For the low resource language, we decided on Northern Sámi, an agglutinating language and member of the Finno-Ugric subbranch of the Uralic language family, which is spoken in Northern Norway, Finland, and Sweden, by around 20,000 people. Its most famous close relatives include Finnish, Estonian, and more distantly, Hungarian. Though not the native language of many people, Northern Sámi has a vibrant community of speakers as well as a surprisingly large number of online articles. Furthermore, there exist some sets of open source pre-trained word embeddings for Northern Sámi, which when added, became an important element of our low resource system.

2 Related Work

Much of the research on named entity research for languages with little or no annotated data utilizes multilingual transfer which involves training neural networks using a small amount of data in the target low-resource language in conjunction with a data set in a more resource-rich language (Ikhwantri, 2019; Chen et al., 2018; Feng et al., 2018; Rahimi et al., 2019; Murthy et al., 2018; Cotterell and Duh, 2017). This method is what was used to train the Northern Sámi word embeddings that we used in our model (Lim et al., 2018a,b).

There is also some research involving building corpora in low-resource languages for NER using automated (silver-standard) annotations (Jie Zhou and CHEN, 2015; Das et al., 2017). These annotations come from Wikipedia pages which have somewhat marked entities through links as well as categories.

3 Method

For the annotation section of our project, we annotated 50 news articles in Northern Sámi from the news website Yle Sápmi(ann). For the entity types used, we simply kept the four types that we have had available to us in previous assignments: PER, ORG, MISC, and LOC. If we restricted the named entity types to only these four, we could then easily compare with use of the CONLL data for the English part of our experiments, and we would not have to worry as much about ambiguity across named entity types. Even though we were most familiar with these four named entity types,

Type	N. Sámi Example	Description
PER	Ailu Valle	Sámi rap artist
ORG	Ovttastuvvan našuvnnat	United Nations
LOC	Ohcejoga gieldda	municipality
MISC	Viiddon sieiddit	album title

Table 1: Examples of Each Named Entity Type

annotating in another language that was not our mother tongue proved difficult. As Northern Sámi is an agglutinative language, we wanted to use a more morphologically rich language than English as a helper when creating our annotation guidelines. For this project, we found a set of annotation guidelines for Slovenian named entities (Katja Zupan, 2017), which mostly fit with these same four named entities. The Slovenian guidelines add a fifth named entity type, DERIV-PER, but this does not apply to Northern Sámi, so we left this off. A brief overview of our annotation guidelines is found in the bulleted list below.

- PER: personal names, pet names, nicknames, fictional characters
- ORG: schools, companies, businesses, theaters, restaurants, museums, buildings with "organizational structure", bands, coalitions
- LOC: country names, city and town names, regions, street names, church names
- MISC: song titles, movie titles, album titles, names of events, most named entities not easily fitting in any of the above three categories

An example of each type can be found in Table 1. Note that three of the examples exhibit capitalization only on their first element. This is a common theme for Northern Sámi named entities. In many cases, it is straightforward to identify where a named entity begins. However, there are cases where multiple interpretations are possible. Take *Ohcejoga gieldda* in Table 1, for example. *Ohcejoga* is a place in the genitive case, so we could potentially annotate it *[Ohcejoga]LOC gieldda* as well. However, *Ohcejoga gieldda* is the official name of that municipality, so we decided to annotate the full official name of an entity, if it appears that way.

A sample annotated sentence is as follows.

- *[Jikngon II]MISC lea vuosttas davvisámegillii dubbejuvvon [Disney]ORG-filbma.*

The above sentence can be translated as follows:

- *[Frozen 2]MISC is the first [Disney]ORG-film dubbed into Northern Sámi.*

Here, Northern Sámi (davvisámegillii), is not treated as a named entity because it appears entirely in lowercase.

Although most of the current research on low resource languages focuses on training neural networks using multilingual data, we did not have the resources to create a model on that scale. So, we used an implementation of Conditional Random Fields through CRFSuite and focused on annotation and feature tuning to find the best results, and utilized word embeddings that were pre-trained through cross-lingual methods. We started with some of the baseline features that we had used on larger data sets in English and found that from those features the ones that performed the best were a bias feature, a token feature, an uppercase feature (which fires if a word is all uppercase), and a word shape feature. The feature vectors were taken from a window of two tokens in either direction for each word in each sentence.

We added a suffix feature, which would add to the feature vector for a word its last few letters. We used several different constraints for this in terms of the number of letters considered a suffix, and the length a word needs to be in order for a suffix to be taken, ranging from taking only small suffixes from large words to taking much larger suffixes more liberally. We found that the best constraints were to take suffixes of size two from any token that is longer than five characters, and this improved overall accuracy.

Northern Sámi uses locative markers in some instances. With the inclination that it might help improve out results for the LOC type of entities, we also used a locative feature that would extract any instance of these morphemes at the end of a word. We tried this feature with and without the suffix feature and found that at best it changed nothing and at worst it mode the model less accurate, so we do not use it in our final model.

We also implemented a feature that would search name lists to see if a token is a name. We first used a list for Sámi given names (sam). We noticed that there seemed to be a lot of Finnish names in the data as well, so we added another list of Finnish given names (fin). We found that both improved the F1 scores a little, but the Sámi names performed a

Type	Count
MISC	179
LOC	504
ORG	517
PER	705
TOTAL	1,905

Table 2: Number of entities annotated per type

Type	Precision	Recall	F1
All	69.74	61.56	65.35
MISC	55.95	36.34	42.91
LOC	66.82	65.86	66.15
ORG	57.87	42.25	48.08
PER	78.02	78.58	78.04

Table 3: Results with Northern Sámi data

Type	Precision	Recall	F1
All	61.64	59.91	60.74
MISC	00.00	00.00	00.00
LOC	79.46	77.33	78.07
ORG	52.29	53.00	52.43
PER	75.78	73.14	74.09

Table 4: Results with English data

little better despite being a much smaller list. (And using them together did not work as well as using the Sámi names alone). So in our final model we use the Sámi names.

Our main focus in terms of features was on the word embeddings. We used open source multilingual word embeddings that were trained on Northern Sámi along with two languages (English and Finnish) that have more resources for larger annotated corpora (Lim et al., 2018a,b). In keeping with Lim et al., who use the embeddings in a Sámi parser, we found that the Sámi-Finnish word embeddings performed slightly better than the Sámi-English embeddings. We also tried different scaling factors with the embeddings, and with a baseline of 2.0, we found a noticeable improvement through using a scaling factor of 0.5 with the Sámi-Finnish embeddings.

We tried a few other avenues, such as searching for lists of Sámi organizations for another name list of organizations, trying to extract silver-standard annotations from Northern Sámi pages in Wikipedia, implementing prefix features, and trying to find other pre-trained word embeddings and other pre-annotated data. However, we were unsuccessful in finding anything else to improve our model.

4 Results

As mentioned in 3 above, we annotated a total of 50 documents taken from news articles in Northern Sámi (ann) using one annotator. The distribution of annotations is shown in Table 2. The most frequent type of entity that was annotated was PER at 705. The types LOC and ORG had roughly the same number of annotations (504 and 517, respectively). The least frequent named entity type that was labeled, which was much smaller in number than the other types, was MISC, which had 179 instances. The total number of entities that were labeled over all documents and types was 1,905.

The results on the Northern Sámi data are displayed in Table 3. We tested 50 annotated articles using cross-validation with a k-fold of 5.

The overall F1 score was 65.35. The lowest F1 score was for the MISC category at 41.91, followed by ORG at 48.08. The highest scoring type was PER which had an F1 score of 78.04. The score for LOC was closer to the overall score at 66.15. These results are from the model using Finnish-Sámi word embeddings (with a 0.5 scaling factor), and a name list of Sámi given names as well as the baseline features (bias, suffix, word shape, uppercase, and token).

This compared well to the English data displayed in Table 4. Testing the same amount of data on the baseline features with English word embeddings (with a scaling factor of 2.0), the overall F1 score was 60.74. Again, MISC was lowest but in this case it was 0.00 because there was no instances of MISC in the English data. ORG also followed MISC as the next lowest performing in the English model with an F1 score of 52.43. In the English model, the LOC and PER types were the two types with the highest F1 scores as in the Sámi model, but the scores in the English model were a lot closer at 78.07 and 74.09, respectively.

5 Conclusion

In conclusion, we found that our system trained on Northern Sámi data performed comparably with English data of the same size. As described above, the Sámi-Finnish word embeddings performed better than the Sámi-English embeddings, most likely due to the fact that Northern Sámi and Finnish are more closely related to each other. Also, we found that using a name list for Sámi names improved performance (if minimally), so we kept this as a feature in the final model.

However, using name lists on organizations did not seem to help, perhaps because the format of organizations can vary significantly depending on the writer or speaker. Furthermore, isolating the locative suffix -s did not help significantly, most likely because the locative suffix is not strictly a locative suffix, but has other functions. For example, in Northern Sámi, the locative suffix is also used to indicate possession.

One aspect that may have affected performance could be the quality of the annotations. Our annotator was not a native speaker of Northern Sámi. In the future, it might be beneficial to get annotations from a native speaker of the language, so that the rules learned by the model are more reflective of the true inner workings of the language.

In terms of name lists, it might also be beneficial to use a name list for common Sámi last names. Due to the fact that the Sámi are a tight-knit community, many surnames surfaced multiple times in the data. Using a name list for last names might be useful in this case, and it might be a tool to help with other tasks such as linking blood relatives together based on surname.

References

Appendix: Northern sami given names.

Category: Finnish given names.

Yle sápmi.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2018. Zero-resource multilingual model transfer: Learning what to share. *CoRR*, abs/1810.03552.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Arjun Das, Debasis Ganguly, and Utpal Garain. 2017. Named entity recognition with word embeddings and wikipedia categories for a low-resource language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 16(3):18:1–18:19.

Xiaocheng Feng, Xiachong Feng, Bing Qin, Zhangyin Feng, and Ting Liu. 2018. Improving low resource named entity recognition using cross-lingual knowledge transfer. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4071–4077. International Joint Conferences on Artificial Intelligence Organization.

Fariz Ikhwantri. 2019. Cross-lingual transfer for distantly supervised and low-resources indonesian NER. *Journal of the Association for Computing Machinery*, abs/1907.11158.

Bi-cheng Li Jie Zhou and Gang CHEN. 2015. Automatically building large-scale named entity recognition corpora from chinese wikipedia. *Zhejiang University Press*, 16.

Tomaž Erjavec Katja Zupan, Nikola Ljubešić. 2017. Annotation guidelines for slovenian named entities janes-ner.

KyungTae Lim, Cheoneum Park, Changki Lee, and Thierry Poibeau. 2018a. SEx BiST: A multi-source trainable parser with deep contextualized lexical representations. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 143–152. Association for Computational Linguistics.

KyungTae Lim, Niko Partanen, and Thierry Poibeau. 2018b. Multilingual Dependency Parsing for Low-Resource Languages: Case Studies on North Saami and Komi-Zyrian. In *Language Resource and Evaluation Conference*.

Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2018. Judicious selection of training data in assisting language for multilingual neural NER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–406, Melbourne, Australia. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.