Towards the First Named Entity Recognition of Inuktitut for an Improved Machine Translation

Ngoc Tan Le and **Ikram Kasdi** and **Fatiha Sadat**
Université du Québec à Montréal
le.ngoc_tan@uqam.ca and ikramkasdi@gmail.com and sadat.fatiha@uqam.ca

Abstract

Named Entity Recognition is a crucial step to ensure good quality performance of several Natural Language Processing applications and tools, including machine translation and information retrieval. Moreover, it is considered as a fundamental module of many Natural Language Understanding tasks such as question-answering systems. This paper presents a first study on NER for an under-represented Indigenous Inuit language of Canada, Inuktitut, which lacks linguistic resources and large labeled data. Our proposed NER model for Inuktitut is built by transferring linguistic characteristics from English to Inuktitut, based on either rules or bilingual word embeddings. We provide an empirical study based on a comparison with the state of the art models and as well as intrinsic and extrinsic evaluations. In terms of Recall, Precision and F-score, the obtained results show the effectiveness of the proposed NER methods. Furthermore, it improved the performance of Inuktitut-English Neural Machine Translation.

1 Introduction

In recent years, Artificial Intelligence has recently gained much attention in research and development, particularly when applied to the field of Natural Language Processing (NLP) and Human Language Technologies. This paper focuses on Named Entities Recognition (NER), one of the crucial tasks in several NLP applications and resources. The latter consists in identifying and classifying the names of the specified categories according to predefined semantic types, such as, the names of people, the place, the organization and the numerical expressions, in particular, the currency, the date and the percentage (Nadeau and Sekine, 2007). NER being among the most important tasks of NLP; however, the success of such models is highly dependent on the amount of available annotated data, which is scarce and difficult to obtain. Furthermore, because of the unavailability of annotated data, it is more difficult to apply these NLP methods to low resourced languages and domains, such as Inuktitut, one of the main Indigenous languages in North America and the Canadian Arctic, and part of a larger Inuit language family, stretching from Alaska to Greenland[1].

According to UNESCO, 75% of Indigenous languages are threatened with extinction, and language loss is currently occurring at an accelerating rate due to globalization. Therefore, the revitalization of endangered languages has become an important task for the preservation of cultural diversity on our planet (Bird, 2020).

In our research, we are interested in Inuktitut. Our main objective in this framework is to address the linguistic challenges and to detect named entities for this language through the following contributions:

- Explore the NER task for the Inuktitut language. To our knowledge, works on this task in related with Indigenous languages such as Inuktitut are rare, or non-existent. Therefore, our study will be the first to be carried out for this task.

- Perform a comparative study between two methods, using: (i) rule-based projection based on a morphological analyzer and a word aligner; and (ii) bilingual word embeddings based on semantic similarity in a bilingual vector space.

- Build an annotated corpus in Inuktitut for the NER task. This corpus will contribute to future work for various subfields of NLP, namely information retrieval, neural machine translation, and conversational agents (chatbots).

[1] https://www.thecanadianencyclopedia.ca/en/article/inuktitut

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 84–93

Also, this work would contribute to the preservation and revitalization of the Inuktitut as well as other (related) Indigenous languages.

- Improve the performance of a Neural Machine Translation (NMT) system by including a NER module.

The current paper is organised as follows: Section 2 introduces Indigenous knowledge including research on several domains such as language, culture, and identity, as well as the relevant works in NER domain. Section 3 presents our methodology via several methods to deal with NER task, and an empirical case study of the Machine Translation task including the NER. Experiments and evaluations are presented in Sections 4 and 5. Finally, Section 6 gives some conclusions and future research directions.

2 A dive into Indigenous Research and NLP

Since 2020, new directions for Indigenous research were put in place by Canada research coordinating committee[2], to help Indigenous peoples and communities partner with research fields, to support and to encourage them to conduct their own research[3]. As with any culture, language is an essential part of Indigenous knowledge, it is also one of the important disciplines of Indigenous research in Canada.

Indigenous languages in Canada have changed and evolved over time and over generations. Like all languages, they carry literary, cultural, traditional, but also historical values (Dorais, 1995). One of the particularities of the Indigenous languages of Canada is that, for some, they are not spoken elsewhere in the world and are specific to Canada[4]. As a result, these languages must be preserved because they represent one of the linguistic and therefore cultural riches of Canada. It is mentioned in the Canadian statistics[5], that the 2016 census recorded more than 70 Indigenous languages divided into 12 language families. The Inuit languages are considered the second Indigenous language family with the largest number of speakers after the Algonquian languages. The most used language in this linguistic family is Inuktitut, mainly spoken in Nunavut and Quebec. In our research, we are particularly interested in this language, rich and at the same time morphologicall complex, as presented in the following section.

2.1 Linguistic challenges in Inuktitut

Indigenous languages in Canada are considered as endangered languages, that reflect the richness of cultures, the history of a people and the diversity of knowledge. Inuktitut is one of the four major sets of dialects of Inuit languages in Canada, from Alaska to Greenland. Mainly spoken in Nunavut and Quebec, it is also spoken in areas of Newfoundland and Labrador as well as in the Northwest Territories. In 2016, the census counted 39,770 speakers, with 65% living in Nunavut and 30.8% living in Quebec.

The preservation of Inuit languages is valued by Indigenous peoples because they are languages that are not spoken elsewhere in the world and their transmission to future generations is not easy. Indeed, Statistics Canada reports that in 2006, 21.4% of the Indigenous population reported being able to carry on a conversation in an Indigenous language. Nevertheless, this percentage decreased to 15.6% in 2016.

Inuktitut is written with a syllabic system, that said, it also has an orthography of the Roman alphabet and the orientation of the writing of the sentences is done, as for French or English, from left to right. The Inuktitut syllabary has differences between dialects. This is because certain sounds exist in one dialect and not in the other. This feature is also found in the spelling system or the spelling of the Roman alphabet of the Inuktitut language, these differences are represented by additional symbols.

The Inuktitut spelling, based on the letters of the Roman alphabet, aims to be more faithful to the pronunciations and specificities of the language in order to be standardized and made more systematic (Compton, 2021).

The Inuktitut language has a particular grammar and fairly complex word compositions that differ-

[2]Canada Research Coordinating Committee: https://www.canada.ca/fr/comite-coordination-recherche/priorites/recherche-autochtone/plan-strategique-2019-2022.html

[3]Indigenous Peoples and Communities: https://www.rcaanc-cirnac.gc.ca/fra/1100100013785/1529102490303

[4]Indigenous languages of First Nations, Métis and Inuit: https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016022/98-200-x2016022-fra.cfm

[5]Census record: https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016022/98-200-x2016022-eng.cfm

entiate it from other languages.

Example[6] :

Tusaatsiarunnanngittualuujunga, that means *I don't hear very well*

That sentence word could be segmented as follows: The root Tusaa- (to hear) is followed by 5 suffixes: tsiag- (well), -junnag- (to be able), -nngit- (negation), -tualuu- (much), -junga (first person singular and present tesnse).

2.2 NER for Indigenous languages

In the NER task for Indigenous languages, we classify mainly two types of methods: (1) the one based on rules, and (2) the other ones based on transfer learning, a method that uses the knowledge acquired from one task to be transferred to a second task, recently relying heavily on deep learning. In the first category, sets of rules are manually made for each entity type based on context and morphological features (Fong et al., 2011). In the second category, the transfer approach, such as in the NER model, from rich-resource languages, is an attractive achievement, due to the large amounts of annotated data available (Collobert et al., 2011; Huang et al., 2015; Peters et al., 2018). In this research, we propose methods that use parallel corpora or word embeddings to project the annotation across languages.

Recently, the state of the art of NER for low-resource languages relies on multi-parallel corpora or word embeddings as proposed by Ehrmann et al. (2011), with a goal to annotate corpora in several languages such as French, Spanish, and German. In their research, they used the IBM model to extract word-for-word alignments and therefore aligned entities that represented a group of words.

Other methods used Machine Translation (MT) to project the annotation between languages. Tiedemann et al. (2014) aimed to rule out noisy annotation as to the source language of a parallel corpus. They relied on manual annotation through the UD tree bank (Universal Dependencies) combined with MT. This combination made it possible to train a fully lexicalized analyzer. On the other hand, Mayhew et al. (2017) performed word-to-word or sentence-to-sentence translation using lexicons to translate available annotated data in rich-resource languages.

Stengel-Eskin et al. (2019) introduced an alignment model based on an encoder-decoder architecture, which was integrated into a MT model based on Transformers. They evaluated the performance of their system on the projection of NER data from English to Chinese and outperformed the fast-align based model in terms of F-measure.

Jain et al. (2019) proposed a system that improved through three methods of entity projection: (a) to exploit machine translation systems twice: first, sentence translation; next, entity translation; (b) to match entities based on spelling and phonetic similarity; and (c) to identify matches based on distributional statistics drawn from the parallel data set. Their approach achieved improvements on the cross-lingual NER task and achieved state-of-the-art F1 score for the Armenian language.

In addition, more relevant research to the NER task on Nordic languages, are presented as under-represented or Indigenous languages, such as Icelandic (Ingólfsdóttir et al., 2019), Finnish (Hou et al., 2019; Luoma et al., 2021), Nynorsk (Johansen, 2019), Danish(Plank, 2019).

Other works, such as Azmat et al. (2020), introduced a named entity annotation transfer method also based on NMT. Their approach consists in pre-training an NMT system, from a parallel Uyghur-Chinese corpus. Then, the boundary information that marks the named entities is added to the source language sentences to re-train the previously trained model so that it can learn to align the named entities. The results show that their system obtains a considerable improvement over the base model in terms of F_measure.

Hatami et al. (2021) used the fast-align tool to extract word matchings. Then, two heuristics were applied to obtain alignments in both directions for parallel English-Brazilian Portuguese data. The latter being a low-resource language.

Xie et al. (2018) proposed a method which trains the monolingual word embeddings, projects the two spaces of embeddings of the words of the two languages in the same space, translates each word into the source language by finding the nearest neighbor, uses MT to translate named entities.

Adelani et al. (2020, 2022) considered that the incorporation of word embeddings represents a key element for NER. First, they used a rule-based method to identify named entities in addition to entity lists obtained from dictionaries. Second, they used a noise elimination technique based on the (Hedderich and Klakow, 2018) method in order to clean the annotated corpora automatically by the

[6] https://www.mustgo.com/worldlanguages/inuit/

rule-based method. The performances shown that their method was successful for the two Indigenous languages of Africa: Hausa and Yoruba.

Among the methods that deal with low-resource languages, Yohannes and Amagasa (2022) introduced TigRoBERTa which was trained on corpora in Tigrinya, an Ethiopian Semitic language. Then they performed fine-tuning on downstream tasks such as NER.

3 Methodology

A promising solution for NER task in low-resource languages, without annotated data, is from rich-resource languages using unsupervised transfer models. Given the unavailable annotated data for the Inuktitut and the availability of the latter in English, the main idea of our approach is to transfer the linguistic features of English to Inuktitut. However, the main challenge of this method is the mapping of lexical items between languages. Indeed, this is due to differences in words and word order across languages.

We present, here, two approaches. The first approach consists of transferring the NER annotation from English to Inuktitut by combining rules using a morphological analyzer with word alignment; while the second approach is based on the bilingual word embeddings using a bilingual dictionary (English-Inuktitut) that we built.

3.1 Rules-based approach

In this approach, we used word alignment information with a morphological rule set. The main steps consist of:

- Extracting named entities from the English corpus. For instance, *Ms. Perkison*, *first Legislative Assembly of Nunavut*.
- Performing a morphological analysis of Inuktitut sentences. Example: the morphological analysis of the word *Titiraqsimaningit* which means in English *First* is: {titiraq:titiraq/1v}{sima:sima/1vv} {ni:niq/2vn}{ngit:ngit/**tn**-nom-p-4s}. The word ending is a tn, which means it's a noun ending.
- Identifying nominal groups of Inuktitut text. For instance, in Inuktitut text, *mis puukisan*, *sivulliqpaami nunavuup maligaliurvinganni*.
- Filtering out nominal groups that do not represent named entities, by using word alignment.
- Building a dictionary of bilingual named entities (English-Inuktitut). For instance:
  - Ms. Perkison - mis puukisan - PER
  - Legislative Assembly of Nunavut - nunavuup maligaliurvinganni - ORG
  - Assembly - maligaliurviup - ORG
- Building a knowledge base in the Indigenous language (Inuktitut), which will help in carrying out NLP tasks downstream and in preserving Indigenous culture.

Figure 1 illustrates the pipeline of our rule-based method.

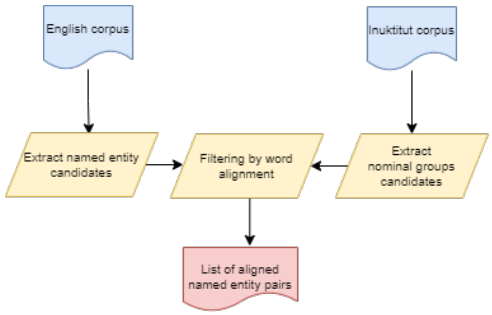

Figure 1: Architecture of our framework: rule-based approach.

3.2 Bilingual word embedding-based approach

Cross-lingual named entities is the transfer of knowledge from a rich-resource language supporting many named entity tags to a low-resource language (Ehrmann et al., 2011). In this approach, we adopt the unsupervised transfer method based on the bilingual word embeddings. This approach addresses the two major challenges: how to solve the word order problem between the languages and effectively to perform the lexical mapping between the two languages. The main steps consist of:

- Building a bilingual English-Inuktitut dictionary.
- Recognizing named entity in English source.
- Training monolingual word embedding on each corpus (English and Inuktitut).
- Translingual projection by performing a linear mapping between the two monolingual word

embeddings in the same space and using a bilingual dictionary.

- Calculating the distance between vectors of bilingual named entities.

- Selecting the nearest neighbor as the translation entity.

Figure 2 shows the pipeline of our word embeddings-based approach.

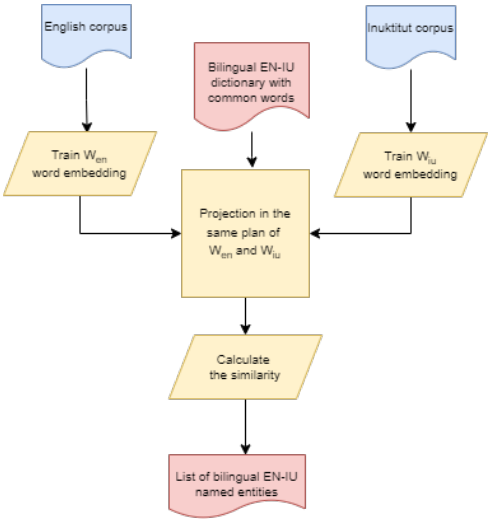

Figure 2: Architecture of our framework: word embedding-based approach.

3.3 Machine Translation Downstream Task

Inspired by (Font and Costa-Jussa, 2019), we built an NMT framework (English-Inuktitut) by taking advantage of pretrained word embeddings, and also source-target alignment information as additional feature.

First, the pretrained word embeddings are used to initialize the embedding layers of the NMT model, both in the encoder and the decoder. We deal with the morphology complexity by applying the morpheme segmentation for Inuktitut (Le and Sadat, 2020).

Second, source-target alignment information is incorporated in the training step. We apply an unsupervised word aligner (Dyer et al., 2013) to generate symmetrical source-target alignments.

Third, we inject, in the decoding, the source-target morphological information, such as bilingual lexicon. We apply a lexicon extractor from Moses (Koehn et al., 2007) to prepare a bilingual lexical shortlist which is passed to the decoder.

4 Experiments

4.1 Data preparation

This corpus includes the proceedings of the 687 days of debates with 8,068,977 words in Inuktitut and 17,330,271 words in English, which gives approximately 1,3 million sentence pairs. This corpus has been used in several research works, particularly in the shared task. The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 (Joanis et al., 2020) is used to train and to evaluate our proposed models (Table 1).

Dictionary

Using the UQAILAUT[7] project database, we were able to build a bilingual dictionary of 1,560 words. This constitutes root word meanings as well as suffix meanings. We used the Microsoft Bing translator[8] to translate the most frequent English words in the parallel corpora into Inuktitut.

Dataset	Train set	Dev set	Test set
Inuktitut (iu)	1,293,348	5,433	6,139
English (en)	1,293,348	5,433	6,139

Table 1: Statistics of Nunavut Hansard for Inuktitut-English (Joanis et al., 2020).

4.2 Settings for embeddings pretraining

We setup an experimental environment in Table 2. To pretrain word embeddings, the hyper-parameters are configured in Table 2. The *fastText* toolkit (Bojanowski et al., 2017) is used to pretrain them.

Hyper-parameters
Epochs = *50*
Dimension size = *300*
Window size = *2*
Alpha value = *0.03*
Loss function = *softmax*

Table 2: Settings of the hyper-parameters for embedding pretraining.

4.3 Settings for Neural Machine Translation

Regarding the NMT task, we used the *fairseq* tool (Ott et al., 2019) to train the Transformer-based models with the parameters mentioned in Table

[7]UQAILAUT project database: https://www.inuktitutcomputing.ca/Uqailaut/

[8]Bing translator: https://www.bing.com/translator (accessed: March 2023)

3. As pre-processing, we used Moses tool (Koehn et al., 2007) to tokenize. Additionally, we applied Byte-Pair Encoding (BPE) subword segmentation with *subword-nmt* tool (Sennrich et al., 2015) to create a 20k vocabulary. In this paper, we performed only two specific experimental models as follows:

- Baseline: standard Transformer-based model
- Model 1: Transformer-based model with word alignment information
- Model 2: Transformer-based model with bilingual word embedding information

The relevant hyperparameters of NMT models are shown in Table 3.

Hyper-parameter	Value
Maximum sentence length	128
Batch size	32
Dropout rate	0.3
Transformer layers	12
Transformer hidden layers	768
Learning rate	0.0005
Epoch	40
Optimizer	adam

Table 3: Settings of hyper-parameters for NMT models

5 Evaluations

To evaluate our proposed method, we used automatic evaluation metrics such as, Recall, Precision, F1, alignment error rate (AER), BLEU score for BiLingual Evaluation Understudy (Papineni et al., 2002) with SacreBLEU (Post, 2018), chrF++ (Popović, 2015) for calculating character n-gram F-score, and translation error rate (TER).

5.1 Evaluations on word alignment

Table 4 represents the word alignment results of words tested using several alignment tools. We also compare our results with those of the Shared Task (Koehn et al., 2005) obtained by (Langlais et al., 2005) namely NUKTI and JAPA in the Table 5.

The word alignment tools were trained on the Nunavut Hansard Inuktitut–English parallel corpora (Joanis et al., 2020), as our same training dataset, and were evaluated on a gold alignment set used in the Shared Task. The performances obtained with the *Eflomal* tool (HMM + fertility) shown a significant improvement in the alignment error rate compared to the others. This is explained by the iteration sampling method that this model uses.

	AER	P	R	F1
Fast align	0.643	0.25	0.623	0.25
GIZA	0.669	0.32	0.33	0.33
Eflomal (ours)	0.474	0.367	0.930	0.367
Eflomal (IBM + HMM)	0.499	0.351	0.874	0.351
Eflomal (IBM1)	0.596	0.281	0.721	0.281

Table 4: Performance of the word alignment tools.

The word alignment results obtained a higher alignment error rate compared to the shared task aligners. Our results are close to the results obtained by the NUKTI model combined with the JAPA model but still remain less efficient than the NUKTI model.

5.2 Evaluations on rule-based method

In order to evaluate the named entities projection performance, we built a small annotated dataset of named entities in Inuktitut. This dataset contains 4 types of named entities: 45 LOC (location) entities, 38 ORG (organization) entities, 111 PER (person) entities and 11 MISC (miscellaneous) entities which do not belong to any type. Table 6 presents the evaluation results of our rule-based method.

The results of this approach could be interesting, especially for the PER entity in proportion to all named entities. Due to the morphology of Inuktitut which is very different from that of English, the word alignment tool could be misled.

Unlike languages admitting the same morphological typology, the alignment error rate is much lower. Moreover, the parts of the text which represent a PER entity consisting of n words generally admit a translation of n words (word-for-word translation). For instance, the translation of the

	AER	P	R	F1
Eflomal (ours)	0.474	0.367	0.930	0.367
NUKTI	0.306	0.631	0.659	0.645
NUKTI+JAPA	0.465	0.513	0.536	0.524
JAPA	0.713	0.262	0.745	0.387

Table 5: Comparison about performance of several word aligners.

	P	R	F1
PER	0.84	0.73	0.78
ORG	0.81	0.54	0.65
LOC	0.95	0.59	0.73
MISC	0.90	0.20	0.33

Table 6: Performance of our proposed rule-based NER model, with 4 classes such as Person, Organization, Location and Miscellaneous.

en2iu	BLEU	ChrF++	TER
Baseline	31.31	42.02	53.83
Model 1	**32.84**	**44.07**	**56.46**
Model 2	31.70	42.54	54.49

Table 8: Performances on NMT in terms of lowercase word BLEU score in the direction English to Inuktitut. BLEU signature: "nrefs:1| case:mixed| eff:no| tok:13a| smooth:exp| version:2.0.0".

PER entity "Glenn McLean" is "gilin maklain". On the other hand, the translation of the LOC entity "Whale Cove" is "tikirarjuaq".

5.3 Evaluations on bilingual word embedding-based method

In order to evaluate the translation performance in the common word embedding space, we constructed a bilingual evaluation dictionary consisting of 30 word pairs.

The evaluation was done by calculating the accuracy of the translation of the words in the neighborhood of k = 1, 5, 10. We took into account the similarity between the word to be translated and the neighboring words.

k	Precision
1	0.367
5	0.400
10	0.433

Table 7: Results of the word-to-word translation by our proposed bilingual word embedding-based method, in terms of precision.

We notice that the performance for the neighborhood of k = 10 is the best, with 0.433 in terms of precision (Table 7). This is explained by the fact that the probability of finding the correct word translation is high when the number of neighbors is large.

5.4 Results on Neural Machine Translation downstream task

For the NMT downstream task, we observed a gain in the performance. The model 1 obtained the best performance than the baseline and the model 2 in terms of BLEU, ChrF++ and TER. The reason is that model 1 succeeds in aligning the entities in the parallel corpus despite the alignment error rate.

Contrary to the model 2 which performed the translation of named entities word by word in the space of bilingual word embeddings by selecting the nearest neighbor. This sometimes distorts the translation of named entities, particularly Inuktitut words representing sentences.

5.5 Error analysis and discussion

Regarding the method based on rules and word alignment, the performance is higher for PER(son) and LOC(ation) entities. This is explained by the morphology complexity. However, proper nouns are usually translated verbatim, while other entities such as ORG(anization) and MISC(ellaneous) represent sentences whose the translation in Inuktitut is just a single word.

Example: the translation of the PER entity "Hunter Tootoo" is "Hanta tutu", the translation of the ORG entity "Legislative Assembly" is "Maligaliurvik".

The morphological difference between the two languages caused misalignments of words, which resulted in the erroneous projection of named entities.

The evaluation results of the three models show that the model 1 which is based on the words alignment is the most efficient, then the model 2 which is based on the bilingual word embeddings. The reason is that the model 1, apart from alignment errors, is still able to align named entities in both languages.

On the other hand, the model 2 performed word-to-word entity translations. However, as previously explained, the Inuktitut language, being a polysynthetic language, a sentence can be represented by a single word.

We noticed the main error types as follows:

(1) *Projection errors due to word alignment errors*, as illustrated with the following examples:

(iu) Uqausiksait jain sutuuatmut, maligaliuqti, inulirijituqakkunnut ministarijaujuq.

(en) Presentation by the Hon. Jane Stewart, MP, Minister of Indian Affairs and Northern Development.

Here, the PER entity "Jane Stewart" is aligned with "sutuuatmut", instead of "jain sutuuatmut".

(2) *Errors in the identification of nominal groups.* Sometimes, a noun, that follows or precedes an entity named in Inuktitut, is considered part of the entity, since sequences of names have been considered named entities, as illustrated in these examples:

(iu) Nuqqausirutigilugu, uqausirikkannirumavakka katimajiuqatima ukausirisimajangit taivit alakannuap, sinnattuumajunnaiqpugut.

(en) In closing, I would like to echo comments by my colleague Ovide Alakannuark, we are no longer dreaming.

The PER entity Ovide Alakannuark has been aligned with the whole nominal group ukausirisimajangit taivit alakannuap instead of taivit alakannuap.

(3) *Translation errors due to out-of-vocabulary words and restricted data domain.*

This is due to the data source which concerns the legislative assembly. Unlike the dictionary built from the UQAILAUT project database, the word pairs come from the general domain, as well as the out-of-vocabulary words. Examples:

(en) Legislative Assembly Of Nunavut.

(iu) maligaliurvia Ralaa Jumaar Nunavut, instead of nunavut maligaliurvia.

(en) South Baffin

(iu) Nginni baffin, instead of qikiqtaaluup niggiani.

Through the conducted error analysis, we found shortcomings in our models. However, we have found that the method based on word embeddings is less efficient than the method based on rules because of the change it brings to the translation of named entities.

It is interesting to carry out a hybridization involving the two methods based on rules and word embeddings.

6 Conclusion and perspective

In this paper, we have built a named entity recognition system for Inuktitut, an Inuit language of Canada. Counted among the four major dialectal groups of Inuit languages, Inuktitut is written using the Native Canadian syllabary. Indeed, it is a low-resource Indigenous language that has no labeled data for NER; which presents a great challenge to the construction of the first NER system. Also, the Inuktitut language, being a polysynthetic language, has a particular grammar and fairly complex word compositions that differentiate it from other languages. To overcome these problems, the main idea of our approach is to use English, given that it is a language rich in resources and that has labeled data for NER and a parallel Inuktitut-English corpus is available. Thus, in this paper, we built a model capable of detecting named entities in Inuktitut, by transferring linguistic characteristics from English to Inuktitut.

In addition to being the first research on named entities recognition for Inuktitut Indigenous language, this project contributes to the preservation of this language and its culture. Furthermore, by building a knowledge base in the Inuktitut language involving named entities, this will contribute to the realization of future works that affects other NLP sub-tasks , such as Information Retrieval, Machine Translation or question0answering systems.

As a future research, we aim to integrate knowledge bases such as those related to toponymy and data from Indigenous knowledge in training word embeddings and improving the performance of our systems (NER and NMT). In addition, we aim to emphasize a differentiation between named entities of Inuktitut origin (such as the names of people and places) and those borrowed. All with the aim of pursuing collaborations with an Indigenous community in Nunavut whose mother tongue is Inuktitut.

References

David Adelani, Graham Neubig, Sebastian Ruder, Tatiana Moteu, Dietrich Klakow, and et al. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Michael A. Hedderich, Dawei Zhu, Esther van den Berg, and Dietrich Klakow. 2020. Distant supervision and noisy label learning for low resource named entity recognition: A study on hausa and yorùbá.

Anwar Azmat, Li Xiao, Yang Yating, Dong Rui, and Osman Turghun. 2020. Constructing Uyghur name entity recognition system using neural machine translation tag projection. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 1006–1016, Haikou, China. Chinese Information Processing Society of China.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch.

Richard Compton. 2021. Inuktitut. In *The canadian encyclopedia*.

Louis-Jacques Dorais. 1995. Language, culture and identity: Some inuit examples. *Canadian Journal of Native Studies*, 15(2):293–308.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. pages 118–124.

Yong Soo Fong, Bali Ranaivo-Malançon, and Alvin W. Yeo. 2011. Nersil - the named-entity recognition system for iban language. In *PACLIC*.

Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

Ali Hatami, Ruslan Mitkov, and Gloria Corpas Pastor. 2021. Cross-lingual named entity recognition via FastAlign: a case study. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 85–92, Held Online. INCOMA Ltd.

Michael A. Hedderich and Dietrich Klakow. 2018. Training a neural network in a low-resource setting on automatically annotated noisy data.

Jue Hou, Maximilian Koppatz, José María Hoya Quecedo, and Roman Yangarber. 2019. Projecting named entity recognizers without annotated or parallel corpora. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 232–241, Turku, Finland. Linköping University Electronic Press.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Svanhvít Lilja Ingólfsdóttir, Sigurjón Þorsteinsson, and Hrafn Loftsson. 2019. Towards high accuracy named entity recognition for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 363–369, Turku, Finland. Linköping University Electronic Press.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual ner.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Bjarte Johansen. 2019. Named-entity recognition for Norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 222–231, Turku, Finland. Linköping University Electronic Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn, Joel Martin, Rada Mihalcea, Christof Monz, and Ted Pedersen, editors. 2005. *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Association for Computational Linguistics, Ann Arbor, Michigan.

Philippe Langlais, Fabrizio Gotti, and Guihong Cao. 2005. NUKTI: English-Inuktitut word alignment system description. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 75–78, Ann Arbor, Michigan. Association for Computational Linguistics.

Tan Ngoc Le and Fatiha Sadat. 2020. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666, Barcelona, Spain (Online). International Committee on Computational Linguistics (COLING 2020).

Jouni Luoma, Li-Hsin Chang, Filip Ginter, and Sampo Pyysalo. 2021. Fine-grained named entity annotation for Finnish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 135–144, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity

recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Barbara Plank. 2019. Neural cross-lingual transfer and limited annotated data for named entity recognition in Danish. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 370–375, Turku, Finland. Linköping University Electronic Press.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units.

Elias Stengel-Eskin, Tzu-Ray Su, Matt Post, and Benjamin Van Durme. 2019. A discriminative neural model for cross-lingual word alignment.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140, Ann Arbor, Michigan. Association for Computational Linguistics.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources.

Hailemariam Mehari Yohannes and Toshiyuki Amagasa. 2022. Named-entity recognition for a low-resource language using pre-trained language model. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 837–844, New York, NY, USA. Association for Computing Machinery.