Proceedings of the ESSLLI 2022 Student Session, 38–49
doi.org/10.21942/uva.20367948

Bootstrapping Icelandic Knowledge Graph Data

Steinunn Rut Friðriksdóttir and Hafsteinn Einarsson

University of Iceland, Reykjavík, Iceland
{srf2,hafsteinne}@hi.is
https://english.hi.is/computer_science_phd

Abstract. A knowledge graph is a semantic network of named entities, e.g. people, objects and organizations, that can be used to uniquely identify mentions in text. In order to create such a graph, it is crucial to possess plenty of specifically annotated data that includes not only the entities themselves but the relations that hold between them. Traditionally, such data has only been available for high-resource languages. In this paper, we present our approach to bootstrap training data using machine translation and open relation extraction methods. We hypothesize that by automatically translating our data to English, we can perform relation extraction using SOTA language models before translating the entities back to the source language, significantly reducing startup costs when developing such models for a given language. Our results show that this approach has promise for lower-resource languages such as Icelandic. However, it is currently limited due to the quality of translation and open relation extraction models.

Keywords: Relation Extraction · Knowledge Graphs · Low-to-mid-resource languages · Information Extraction

1 Introduction

The task of Information Extraction (IE) is to extract specific information from an unstructured source document, thus facilitating the retrieval, classification and storage of Named Entities (NEs). Relation Extraction (RE) is an important part of this process where the goal is to retrieve Head - Relation - Tail triplets, thus enabling finer NE distinction by extracting their internal connections. In other words, RE methods make use of the contextual information that enables systems to distinguish *Anne Hathaway* who **acts in** various *movies* from the *Anne Hathaway* who was the **spouse of** *William Shakespeare*.

One can think of a successful RE system as comprising of two steps: Named Entity Retrieval (NER), in which proper nouns are identified as entity mentions in the textual data, and relation extraction where relations between NEs are extracted and the type of each relation is determined. This is a non-trivial task. Relation types can vary greatly between domains, as can the linguistic properties of various languages. For instance, a RE system built for Subject - Verb - Object

(SVO) languages might not work properly for RE in VSO languages. When successful, however, the RE task can be greatly beneficial in building a knowledge graph (KG), which in turn can be used to disambiguate NEs - a task known as Entity Linking (EL).

Most work in RE is on data in English but it remains unclear how such results translate to other languages. What usually stands in the way of successful RE systems is the lack of training data, particularly for low-to-mid-resource languages. The idea behind this paper is therefore to investigate if training data for low-to-mid-resource languages can be bootstrapped to reduce the startup cost required to develop RE methods in a given language.

In this paper, we present our general approach to this task, using Icelandic as an example. In order to avoid the problem of limited training data, we start by using SOTA machine learning models on our corpus [1], translating it into English and then taking advantage of a high-performing English RE system, REBEL, to extract relations from the translated data. Using this approach, we hope to generate valuable data that can either be utilised as is or used as training data for an Icelandic version of the RE system.

2 Literature review

The concept of RE made its debut at the turn of the century but at that time it was only possible to create such systems for high-resource languages. In their 2010 paper, however, Kim et al. [8] utilised a parallel corpus of word alignments to project relation annotation from English (a high-resource language) to Korean (a then-low-resource language). While eventually successful, they noted that the biggest risk to this approach is noise brought on by bad projections. That issue continues on to this day, as cross-lingual approaches to RE can only be as good as the models used for linguistic transfer.

In their 2016 paper, Verga et al. [17] use the idea of a universal schema to build a knowledge base (KB) of NEs and relations, embedding relation types from the input KB as well as the textual patterns from raw text that are likely to contain relations. They embed English and Spanish corpora jointly using an ensemble of USchema and LSTM models, thus managing to train a Spanish RE model using no direct annotation for relations in the Spanish data. They test their method on a slot-filling task where the goal is to correctly fill as many slots as possible using entity-relation triples (e.g. *Barack Obama, per:spouse* should return *Michelle Obama*) without using hand-coded rules or additional annotations as had been done in previous approaches.

Ni and Florian (2019) [10] propose a cross-lingual approach where they project word embeddings linearly from a target language to a high-resource source language, using only a bilingual dictionary of aligned word pairs and thus eliminating the need for parallel corpora of aligned sentences. They then use the

[1] The corpus we use is MIM-GOLD, described in Section 4.

source language (English) four-layer neural RE model directly on the target language without retraining. This method works relatively well on the seven target languages they experiment on. However, the SOV languages score significantly higher than the VSO languages, indicating that RE methods are much more easily transferred between languages that share linguistic properties.

In their 2021 paper, Faruqui and Kumar [4] present a projection method where data from a source language is translated into English using Google Translate. The translated sentence is then used as input for the English RE system OLLIE, and finally translated back to the source language. The authors repeat this process for 61 different languages available in Wikipedia, from various resource-tiers and typologies, Icelandic among them. However, only three of those languages are checked manually, and based on our limited observations, the Icelandic data did not seem to be of high quality. Additionally, their work is based solely on Wikipedia, which does not guarantee good performance on NE and relations outside of Wikipedia's coverage, especially in the case of languages where the subjects covered by Wikipedia are limited. Inspired by their idea, we therefore decided to run our own corpus translation experiment, described in Section 5.

It is evident that RE is a hard task for any model. Pavanelli et al. [11] propose a multilingual BERT-based system for joint entity recognition and relation extraction in text from the healthcare and news domains. While their overall performance is very promising, reaching a 70.6% F1-score on the entity recognition task, they reach only a 26.85% F1-score for the relation extraction task. Despite this, their overall method placed first in the eHealth-KD Challenge 2021 [12] and fourth for the RE task alone. As multilingual approaches have been gaining popularity recently in various machine learning tasks, we suspect multilingual RE will become more commonplace in the near future. We, however, leave their use for Icelandic to future work.

3 Models used

Our work is based on the assumption that there exists a good translation model which can translate back and forth between English and a given target language. By making use of the models presently ready and available in English, we eliminate the need for generating massive amounts of training data from scratch in Icelandic but a sufficiently large monolingual corpus is fundamental for training such a model. The quality of our translations is discussed in Section 6.1. Additionally, it is our assumption that by using predictions generated by a SOTA RE model, we lessen the cognitive workload for our annotators, who only have to accept or reject the model's predictions rather than define their own relations from the source text.

3.1 Translation model

At the Sixth Conference on Machine Translation (WMT21), Facebook released a multilingual model that covers translations of 7 languages to and from English, including Icelandic [16]. The model is trained on several sources of bitext data, including ccMatrix [13], ccAligned [3] and OPUS [15], as well as monolingual data from Common Crawl [2]. The model uses a modified transformer architecture which reaches a BLEU-score of 39.4-40.5 for Icelandic, rendering the model SOTA in machine translations.

The model is fine-tuned on news-domain-specific data, although in the case of Icelandic, this data was mined and filtered according to what the authors judged to be "most likely news domain". Given that they are not native speakers of Icelandic, there is a risk of the fine-tuning data being slightly off. Nonetheless, they reach SOTA results for all language directions and therefore we opted to use their model for our translations. While the model is fine-tuned on news-domain data specifically, the input data can be prefixed to indicate that it comes from another domain.

3.2 Relation extraction model

In many cases, RE models have been treated as a two-part pipeline with the first step involving NER, where NEs are extracted and categorised directly from a text source. The second step is checking whether or not the extracted NEs share a pairwise relation, as well as determining the type of each relation, in a process known as Relation Classification (RC). However, identifying which NEs truly share a relation can be cumbersome, and therefore end-to-end approaches have been developed in order to tackle both problems simultaneously. In 2021, Cabot and Navigli published a paper on their sequence-to-sequence RE model, REBEL [2]. This end-to-end model is based on the Encoder-Decoder Transformer BART and covers over 200 different types of relations. The model reaches SOTA performance for a variety of RE benchmarks, including 75.4% Micro-F1 on the CONLL04 dataset, and is easily adaptable to new domains, making it ideal for our experiment.

REBEL autoregressively generates sets of linearised triplets representing each *entity - relation - entity* triplet present in the input text. The linearisation is accomplished using special tokens as markers, where <triplet> indicates the start of a new triplet, <subj> marks the end of the head entity and <obj> the end of the tail entity, indicating the start of the relation between the head and the tail in its surface form. REBEL requires much less training than previous models which is beneficial in the study of RE for low-to-mid-resource languages. Additionally, the authors provide an efficient way to generate RE data sets from a Wikipedia dump, which in turn can be beneficial for training an Icelandic version of the model later down the line.

[2] https://data.statmt.org/cc-100/

4 Data used

4.1 The Tagged Icelandic Corpus (MIM)

The Tagged Icelandic Corpus (MIM) [9] is a morphosyntactically tagged corpus of about 25 million tokens from diverse sources including media, adjudications and blogs, and includes information on each token's POS and morphosyntactic elements (such as case, number and gender for nominals) as well as their lemmas. It was compiled and tagged during the years 2006–2010. In 2009, work began on a gold standard subcorpus, MIM-GOLD [14], where approximately 1 million tokens were sampled from the original MIM corpus and errors corrected manually. The accuracy of MIM-GOLD is estimated to be 99.6%. As such, MIM-GOLD makes for great data on which to perform NLP experiments.

4.2 MIM-GOLD-NER

In 2018, a team of researchers at Reykjavík University compiled a new annotated version of the corpus, MIM-GOLD-NER [6,7], which includes over 48 thousand NEs of eight types (Person, Location, Organization, Misc, Date, Time, Money and Percent). A semi-automatic annotation process was used to extract NEs and subsequent errors were corrected manually. The corpus is in the CoNLL format and the position of each token is marked using the BIO tagging format.

4.3 MIM-GOLD-EL

In 2021, MIM-GOLD-NER became the basis for another annotated version of the MIM-GOLD corpus. In MIM-GOLD-EL [5] approximately 21 thousand mentions have been linked to their corresponding NEs in Wikidata. The multilingual EL model mGENRE was used, as well as a query run on the Wikipedia API for English and Icelandic, to suggest Wikidata records which were then manually accepted or rejected by the research team.

5 Experiments

5.1 Setup

Our aim in this paper is to create a manually corrected RE version of MIM-GOLD, based directly on the work presented in MIM-GOLD-EL. By doing so, all three aspects of building a knowledge base (KB) or a knowledge graph (KG) will be covered, opening the door for further expansion of IE technology in Icelandic. As shown in Figure 4.3, our process is as follows:

1. Automatically translating the data using Facebook's WMT21 model.
2. Running the translated data through REBEL, thereby generating several relation predictions for each sentence of the corpus.

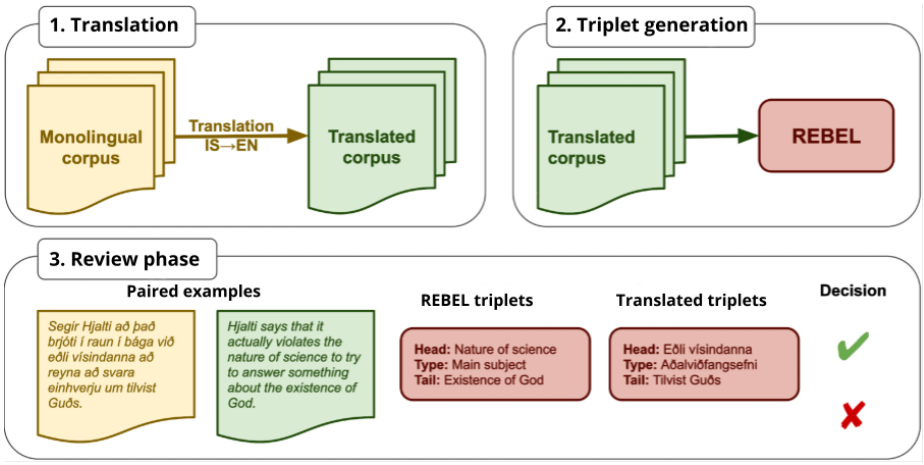

Fig. 1. Our process includes translating the Icelandic corpus into English, followed by running the data through REBEL before manually annotating it.

3. Back-translating the NEs within the triplets generated by REBEL into Icelandic.
4. Manually accepting or rejecting the triplets in order to output trustworthy data to be used either directly as input for a KB/KG or as training material for an Icelandic RE model.

Using automatic methods to generate predictions only requires human annotators to accept or reject the proposed relations. This significantly reduces their cognitive load, making the process quicker, easier and consequently cheaper.

In order to evaluate the practicality of our method, we conducted an experiment using the first 200 sentences[3] from each of the 13 text categories in MIM-GOLD. This sample of 2600 sentences is sufficiently large to infer general performance of the method with a reasonable workload and time frame. As previously stated, the corpus is composed of a variety of sources, e.g. news articles, adjudications and laws, blogs and emails from the University of Iceland staff mailing list. We then additionally gain an understanding of the quality of WMT21's translations, particularly how much or little they are affected by text that falls outside of the news domain.

5.2 Results

Table 5.2 depicts the overall precision of each category from MIM-GOLD. Furthermore, we calculated the proportion of relation triplets where both the head

[3] Exempting the emails category, which contains a total of 60 sentences, and the webmedia category, which contains a total of 195 sentences in MIM-GOLD.

and tail NEs have been retrieved in MIM-GOLD-EL and thus linked to their corresponding entries in the Wikidata knowledge base. We refer to this proportion as entity precision and it reflects how many of the relations we find are actually connecting entities in an existing knowledge graph. The final column indicates how many total predictions were made by REBEL for each category. It should be noted that we deliberately did not accept predictions made by REBEL that, while correct, have nothing to do with the sentence being evaluated. For instance, when REBEL is not able to make any predictions for a given sentence, it tends to fall back on its training data and makes predictions on common subjects such as John F. Kennedy or popular sport events regardless of the subject of the sentence in hand. It is clear that the precision depicted in Table 5.2 would be quite a lot higher if these examples were included, but we wanted to focus specifically on the NEs that appear in our corpus and the information contained explicitly within it.

On average, approximately three predictions are made for every sentence. The overall performance of our method is quite poor, averaging at 26.9% for any relations marked as correct and 2.0% for relations where both the head and the tail are established NEs that have been labelled as such in MIM-GOLD-EL. This poor performance can partially be explained by translation issues, discussed further in Section 6.1. Examining the individual prevision for each category, we can see that the blogs category is the most problematic, followed by the adjudications and laws that include very domain-specific language. The highest scoring categories are those that have undoubtedly been proofread, as well as having a higher number of NEs being presented and defined for the first time (as opposed to more informal text such as the blogs where the writer assumes the reader's knowledge of certain subjects). The emails category has the highest precision out of all the categories but considering its small size it's hard to conclude anything about emails in general.

In order to gain an understanding of the recall of our method, we manually annotated 100 sentences from the books category for explicit relations that include established NEs as both head and tail. By explicit relations we mean that the head and tail entity must be present in the sentence example and they must in fact be NEs (see Section 6.2 for discussion on open versus closed RE). This manual annotation resulted in 47 explicit relation triplets. Out of those, 25 were retrieved by our methods, of which 9 were perturbed by the translation process (see Section 6.1). Although these calculations are based on little data, they suggest that the recall score of around 53% is much higher than the precision reported at 26.9%. This indicates that if the translation process was avoided, e.g. by training REBEL from scratch on Icelandic data, or simply by restricting REBEL to consider only relations between established NEs, we could significantly improve the overall scores. However, that would come at the cost of missing out on relations between entities in the text that are not registered in the knowledge base.

Category	Overall precision	Entity precision	Predictions examined
Adjudications	16.1	2.6	740
Blogs	12.8	1.3	688
Books	**34.0**	**6.2**	777
Emails	***36.1***	4.5	202
Laws	18.5	0.9	710
Newspaper 1	22.5	0.4	711
Newspaper 2	33.4	1.3	779
Radio/TV news	28.9	0.6	720
School essays	27.6	2.9	761
Scienceweb	30.9	0.6	713
Webmedia	27.6	2.2	671
Websites	33.9	2.1	714
Written to be spoken	32.1	2.2	713
lightgray **Overall**	**26.9**	**2.0**	**8899**

Table 1. Precision per category. The overall precision shows the percentage of all relations marked as correct. Entity precision shows the percentage of relations marked as correct if both head and tail have previously been identified as NEs and linked to their corresponding Wikidata entries.

6 Limitations and considerations

There are several limitations to our methodology that must be taken into consideration. In this section we discuss the main ones.

6.1 Machine translation

It is evident that running the data through WMT21 not only once but twice causes several problems. In fact, our estimation is that 20–30% of all predictions made by REBEL are perturbed by the translation, making them invalid in the labelling process. Some of these errors are results of direct translations, where an Icelandic name has been translated directly to English but loses its meaning along the way. Examples of this include the masculine name *Erlendur*, which gets translated as *foreign*. That translation is not incorrect, as this is literally what the name means, but when it gets translated back to Icelandic, we no

longer have a name of a person but rather an adjective that doesn't fit inside the original context.

On the other hand, WMT21 often seems to perturb people's names in a strange manner. We suspect this has two main reasons. Firstly, WMT21 is a multilingual translation model and some of the translation mishaps could be the result of the model's training in other languages getting in the way. For instance, the masculine name *Alfreð* gets changed to *Alfredo* in the English translation, *Rafn* becomes *Rafael* and *Sveinn* becomes *Sweene*. Secondly, some of the mishaps might be the model's unsuccessful attempt to adapt to the Icelandic noun cases. For instance, the feminine name *Hildigunnur* has the genitive form *Hildigunnar*. In the hands of WMT21, we get the incorrect form *Hildiganna*. Another example is the feminine name *Sólrún* which is *Sólrúnu* in both the accusative and dative cases but *Sólrúnar* in the genitive. When translated by WMT21, it incorrectly becomes *Sólrúna*.

There are several other translation mishaps that can be reasonably explained by the model's attempts to use its training data with unfortunate results. The novel *Dauðarósir* by Arnaldur Indriðason should, ideally, not get translated at all but a direct translation of the title would be *Death roses*. WMT21 changes the title completely, making it *Deathly hallows*, which is a real novel by a different author. Translating it back to Icelandic therefore results in a different title that has nothing to do with the original sentence. Another example of this kind of mishap is when *Kauphöllin* (short for *Kauphöll Íslands*, the Icelandic stock exchange) gets translated as *Exchange*. This is not particularly problematic in the English translation but when translated back to Icelandic, the literal translation *skipta* (meaning *to exchange something*) is used.

Some of the reasonable, yet strange, translation mishaps are less relevant to our attempt to capture relations that include established NEs, but still worth mentioning. For instance, *bleikja* means *river trout*, a fish commonly eaten in Iceland. In the English translation, it instead becomes *bleach*. At first glance, this seems very strange, but in fact, *bleikja* can also be an old-fashioned version of the verb *to bleach something*. Other examples include *dópsali* (*drug dealer*) becoming *pharmacist*, *hnakka* (the oblique case form of *the nape of the neck*) becoming *saddle* (*hnakkur* in Icelandic) and *Samfylking* (an Icelandic centrist political party) becoming *Social Democrats*. A particularly strange, yet somewhat sensible translation mishap, is when *Þjóðviljinn* (an old newspaper title, meaning *the will of the nation*) gets translated to *German newspaper*. The *German* insertion can potentially be explained by the fact that Germans are *Þjóðverjar* in Icelandic, yet WMT21 still somehow realizes that *Þjóðviljinn* is a newspaper.

On the other hand, we have quite a few examples of translation mishaps for which we do not have a good explanation. We can divide these problems into two main categories. Firstly, we have translations that change the meaning of a word without any reasonable explanation. Examples of this include when *Samtök atvinnulífsins* (*Confederation of Icelandic Employers*) somehow

becomes *Confederation of German Industry* and when *kynsystrum* (this word does not have a proper English translation but the literal translation would be *gender sisters*, meaning something similar to *people who have it in common to be women*) ends up being translated as *sex workers*. Secondly, we have translations where the words themselves have been perturbed. *Barðastrandarsýslu* (a location in Iceland) becomes *Bðastrandarsýslu*, *Hjólabœjar* (a company name) becomes *Hjóllabœjar* and *Steingrímur* (a masculine name) becomes *Stingrimur*.

6.2 Open relation extraction

Another issue to consider when annotating RE data is whether or not to exclusively consider relations that contain established NEs. Open RE models such as REBEL extract a diverse set of relations without the need for relation-specific human input [1]. The biggest advantage to this methodology is that it does not require prior annotation and thus can be used directly on any unstructured data. This should, theoretically, be beneficial for lower-resource languages. However, this turns out to be a disadvantage to our goal. While REBEL returns approximately 3 predictions per sentence in our data, only a small portion of those predictions actually refer to established named NEs or, frankly, NEs at all. For instance, the relation triplet *clouds - part of - sky* is correct and should be labelled as such but it does not include any NEs and is therefore not usable for our purposes. This high number of non-NE triplets introduces noise that is necessary to address in order to make this approach practical. The noise could be avoided using other means, discussed further in Section 7.

6.3 Domain specificity

The variety of sources available in MIM-GOLD provides a great opportunity to showcase the models' performance on text types that will certainly appear in a real-life downstream context. As was previously discussed in Section 5.2, REBEL performs better on more formal texts that have most likely been proofread. The poorest performance is unsurprisingly in the blog category but the adjudications and laws categories also receive low scores. This could partially be explained by the domain-specific language of these categories. As an example, one of the adjudications in our data discusses the liability of an educational institution for a student's allergic reactions to adhesives. Specific details are discussed on the material used which requires the annotator to have in-domain knowledge. Similarly, the laws category discusses administrative organisation and legislation in detail which can be problematic in the annotation process. Additionally, the adjudications category includes NEs that have been de-identified (a person might be referred to simply as *A*) which makes it impractical for NE work in general.

7 Conclusion and future work

In this paper, we explored a bootstrapping method where we utilised a SOTA relation extraction model trained on English, a high-resource language, on data in Icelandic, a low-to-mid-resource language, by translating the data from the source to the target language using an automatic translation model. We utilised Facebook's SOTA WMT21 model for our translations and REBEL for our relation extractions. Our results indicate that the translation process poses severe problems that interrupt the RE model from reaching its optimal performance. Additionally, the open RE methodology used by REBEL is not ideal for our goals to extract relations between NEs that have previously been linked to their corresponding entries in a knowledge base since a high fraction of the knowledge triplets does not involve NEs.

It is difficult to evaluate the recall of our methodology as correctly labeled RE data for Icelandic does not exist. We have already started working on a crowdsourcing platform where we intend to use manual labour to extract relations between NEs in our data. The resulting work can then be used both directly and as evaluation data for automatic methods such as the one we have proposed here. An Icelandic adaptation of the REBEL model could be trained and applied to other data sources, paving the way for future research and development in Icelandic IE technology.

References

1. Banko, M., Etzioni, O.: The tradeoffs between open and traditional relation extraction. In Proceedings of ACL-08: HLT, pp. 28-36. Association for Computational Linguistics (2008).
2. Cabot, P. L. H., Navigli, R.: REBEL: Relation Extraction By End-to-end Language generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2370-2381. Association for Computational Linguistics. (2021).
3. El-Kishky, A., Chaudhary, V., Guzman, F., Koehn, F.: CCAligned: A massive collection of cross-lingual web-document pairs. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5960–5969. Association for Computational Linguistics. (2020).
4. Faruqui, M., Kumar, S.: Multilingual open relation extraction using cross-lingual projection. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 1351–1356. Association for Computational Linguistics. (2015).
5. Friðriksdóttir, S.R., Daníelsson, H, Eggertsson, V., Jóhannesson, B.G., Loftsson, H., Einarsson, H.: MIM-GOLD-EL - entity linking corpus for Icelandic (22.01). CLARIN-IS. (2022). http://hdl.handle.net/20.500.12537/168
6. Ingólfsdóttir, S.L, Guðjónsson, Á.A, Loftsson, H.: MIM-GOLD-NER – named entity recognition corpus (20.06). CLARIN-IS. (2020). http://hdl.handle.net/20.500.12537/42

7. Ingólfsdóttir, S. L., Guðjónsson, Á. A., Loftsson, H.: Named Entity Recognition for Icelandic: Annotated Corpus and Models. In the 2020 International Conference on Statistical Language and Speech Processing (SLSP), pp. 46-57. Springer, Cham. (2020).
8. Kim, S., Jeong, M., Lee, J., Lee, G. G.: A cross-lingual annotation projection approach for relation detection. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling), pp. 564-571. Beijing, China. Coling 2010 Organizing Committee. (2010).
9. Loftsson, H., Yngvason, J. H., Helgadóttir, S., Rögnvaldsson, E.: Developing a PoS-tagged corpus using existing tools. In 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages LREC 2010, p. 53. Valletta, Malta. (2010).
10. Ni, J., Florian, R.: Neural Cross-Lingual Relation Extraction Based on Bilingual Word Embedding Mapping. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 399–409. Hong Kong, China. Association for Computational Linguistics. (2019).
11. Pavanelli, L., Schneider, E. T. R., Gumiel, Y. B., Ferreira, T. C., de Oliveira, L. F. A., de Souza, J. V. A., ..., Pagano, A.: PUCRJ-PUCPR-UFMG at eHealth-KD Challenge 2021: A Multilingual BERT-based System for Joint Entity Recognition and Relation Extraction. In IberLEF@ SEPLN. (2021).
12. Piad-Morffis, A., Gutiérrez, Y., Canizares-Diaz, H., Estevez-Velarde, S., Muñoz, R., Montoyo, A., Almeida-Cruz, Y.: Overview of the ehealth knowledge discovery challenge at IBERLEF 2020. CEUR. (2020)
13. Schwenk, H., Wenzek, G., Edunov S., Grave, E., Joulin, A., Fan, A.: CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 6490–6500. Association for Computational Linguistics. (2021).
14. Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E.: Analysing Inconsistencies and Errors in PoS Tagging in two Icelandic Gold Standards. In Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA), pp. 287-291. (2015).
15. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pages 2214–2218. (2012).
16. Tran, C., Bhosale, S., Cross, J., Koehn, P., Edunov, S., Fan, A.: Facebook AI WMT21 news translation task submission. In Proceedings of the Sixth Conference on Machine Translation, pp. 205–215. Association for Computational Linguistics. (2021).
17. Verga, P., Belanger, D., Strubell, E., Roth, B., McCallum, A.: Multilingual Relation Extraction using Compositional Universal Schema. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 886–896. San Diego, California. Association for Computational Linguistics. (2016).