Effective Entity Disambiguation in Low-Resource Languages: A Study of Icelandic

Valdimar Ágúst Eggertsson[§]
Snjallgögn ehf.
Reykjavik, Iceland
valdegg@gmail.com

Benedikt Geir Jóhannesson[§]
Dept. of Computer Science
Reykjavik University
Reykjavik, Iceland
bennigeir@gmail.com

Hafsteinn Einarsson
Dept. of Computer Science
University of Iceland
Reykjavik, Iceland
0000-0001-5072-3678

Hrafn Loftsson
Dept. of Computer Science
Reykjavik University
Reykjavik, Iceland
0000-0002-9298-4830

***Abstract*—Entity disambiguation (ED) is integral to the task of entity linking (EL), the task of identifying named entities in text and linking them to their corresponding entries in a knowledge base (KB). In this paper, we present an effective and efficient ED system for Icelandic, using the Icelandic Wikipedia as a KB. We focus on disambiguation, the linking aspect of EL, assuming that an entity mention has already been located. We perform candidate generation using an alias table and Wikipedia search, and achieve candidate ranking through fine-grained entity typing and the use of an entity-aware Icelandic language model, IceLUKE. We study and compare the effects of different variations of candidate generation and candidate ranking, with the best approach reaching an accuracy of 95.2%. Our results highlight the importance of using an entity-aware language model in the candidate ranking step and show a minor improvement in using fine-grained entity typing to decrease the size of the candidate set before ranking.**

***Index Terms*—entity disambiguation, entity linking, knowledge base, language model, candidate ranking**

I. Introduction

Entity linking (EL) is the task of identifying mentions of named entities (NEs) in text and linking them to their corresponding entries in a knowledge base (KB). The latter part is usually referred to as entity disambiguation (ED). EL is important for a language because it allows for the disambiguation of NEs and the integration of external knowledge into text, which can improve understanding and processing of the language. EL is particularly relevant in the era of big data, as the proliferation of online information has led to an increase in the mention of NEs and the need for accurate and efficient methods of linking them to KBs. EL remains a challenging task, particularly for low-resource languages that have limited data available.

The focus of this paper is on the task of developing an effective and efficient ED system for Icelandic, using the Icelandic Wikipedia as the KB. Once an entity mention has been located in text, it can be linked to a record in a KB in two steps. First, the candidates are generated and, second, they are ranked. Earlier work on named entity recognition (NER) has led to good models for locating NEs in Icelandic [1], [2], but we are not aware of previously published work that explicitly focuses on evaluating EL for Icelandic. Therefore, our work focuses on the latter part, i.e., we assume that an entity mention has been located but needs to be linked. We note that candidate generation can be performed in a variety of ways but, in this work, the candidates are generated using an alias table and Wikipedia search through an API. We further apply a filtering step through entity typing (ET)[1] by only ranking candidates that match the inferred type of the entity mention.

Icelandic is a morphologically rich language with relatively few speakers ($< 400k$) and low resources in the domain of EL, which makes EL more challenging than for high-resource languages. NER4EL, a recent data efficient EL approach, reached state-of-the-art (SOTA) performance on English EL with 18k training examples, whereas previous approaches required 9000k examples for the same performance [3]. Such results are encouraging for low-resource languages, such as Icelandic. Furthermore, for candidate ranking, improvements have been seen with models such as LUKE [4].

Due to recent encouraging results for English, we aim to address the challenge of EL in Icelandic by exploring combinations of advancements that have showed promise on their own. We explore the incorporation of fine-grained ET in the candidate generation step of an EL pipeline. This is based on a key insight from the NER4EL method, which uses fine-grained ET to filter away irrelevant candidates before the candidate ranking step. We use 18 entity types instead of the standard number of 4–8 types (listed in Section II-A). Furthermore, we explore several approaches for candidate ranking, such as one based on LUKE. In particular, we present IceLUKE, an entity-aware Icelandic language model (LM). We use an IceBERT-igc model [2] as a starting point for IceLUKE, and then perform entity-aware pretraining and fine-tune the model on the MIM-GOLD-EL corpus [5].

We study the effects of several variations of candidate generation and candidate ranking. For candidate generation, we study two approaches to look up entities, and we compare fine and coarse entity categories in our filtering step. For ranking, we study the effect of IceLUKE by replacing it with several baselines such as a popularity heuristic and other

[§]Equal contribution.

[1]For the sake of clarity, we note that NER can be thought of as a two step process where the aim is to locate an entity mention and then infer its type. ET corresponds to the second step, i.e., mapping a given entity mention to its semantic class.

Transformer [6] models, e.g. model architectures that are very similar to the one in the original NER4EL pipeline.

Our results show a strong benefit in using an entity-aware LM like IceLUKE in the candidate ranking step. For fine-grained filtering of entities in the candidate generation step, we observe a minor improvement when compared to no or coarse filtering. Our best approach reaches an accuracy of 95.2%, using all additions, i.e. fine-grained ET, IceLUKE, Wikipedia search, and an alias table. However, we do not see a general trend that an alias table or fine-grained filtering improves performance, in addition to using Wikipedia search.

II. Background

The first end-to-end EL system [7], addressed the three main tasks of EL, entity mention extraction, candidate generation, and candidate ranking, simultaneously. Recent approaches to EL have focused on using pre-trained LMs, such as BERT [8], and fine-tuning them for the task. These models can achieve strong performance on EL benchmarks [9] that are usually based on Wikipedia and news datasets, but may not always generalize well to other domains or languages [10]. They are usually trained on texts with a specific structure, such as news or encyclopedic articles, and cannot deal with differently structured texts, e.g. from social media or discussion boards. A system trained on the AIDA evaluation set [11] achieved 94% F1 score on it, but only 66% on an ED evaluation set constructed from Reddit posts [12].

Multilingual entity linking is the task of performing EL in some language using a KB that contains entity information in one or more languages. EL has recently been reformulated as a multilingual task and used a bi-encoder to encode entities and contextual mentions [13]. The mGENRE model [14] is a multilingual version of the GENRE (GENerative RETrieval) model [15], covering approximately 730 million Wikipedia hyperlinks in 105 languages (including Icelandic).

A. NER4EL

The NER4EL method [3] utilizes ET to reduce the need for extensive training data when developing an EL system. This approach introduces a finer-grained set of entity types, expanding on the standard NER types from the traditional 4–8 (`Person`, `Organization`, `Location`, and `Misc` + `Date`, `Time`, `Percentage`, and `Money`) to 18 types. The types can be used to filter the candidate set in a standard EL system to match the inferred type of an entity mention, resulting in high accuracy with only thousands of training examples rather than the millions typically required.

Reference [3] compared the accuracy and training data size for NER4EL vs. the SOTA method GENRE on the standard evaluation set AIDA, with and without utilizing finer-grained ET in the pipeline. GENRE trained on 18k examples achieved 88.6% accuracy, but 93.3% when trained on 9000k examples. In comparison, their baseline EL system, that used BERT to rank entity candidates based on cosine similarity between the input text and entity descriptions, obtained 88.8% accuracy when trained on 18k examples, but with a finer-grained ET in the pipeline the same system obtained 92.5% accuracy when trained on the same examples.

The main components of NER4EL are the following: 1) **Finer-grained ET**, which we refer to as fiNE mapping. It maps entities in the KB to their finer-grained type; 2) **The fiNE typing model**, which was trained to assign a type label to a text containing a delimited entity mention; and 3) **The ED model**, which ranks the candidate entities and selects the one that matches the context best.

B. LUKE

LUKE [4] is a new contextualised representation that is specifically designed to tackle entity-related tasks. LUKE is trained to predict randomly masked entities, as well as words, in an entity-annotated corpus from Wikipedia. LUKE introduces a new self-attention mechanism that is entity-aware, which is an extension of the mechanism found in the original Transformer [6]. When the attention score is determined, the type of token is taken into account, and for each type of token-type pair, a special query matrix is used. In addition to treating words as independent tokens, LUKE treats entities as independent tokens. By doing this, relationships between entities can be modelled. LUKE achieves SOTA results in five entity-related tasks: ET, relation classification, NER, Cloze-style question answering, and extractive question answering. This performance record and the publicly available code base[2] are the reasons why we chose to explore LUKE in this work.

Following the publication of the original LUKE paper, the authors proposed a model based on LUKE that addresses ED specifically [16]. Their model accomplishes this using two orderings to disambiguate entities. In the default approach, the entities are disambiguated one by one in the order determined by the text. In a confidence-order approach, the mentions are disambiguated greedily where the order is determined by the confidence of the model over all unresolved entity mentions in the specified context.

III. Data

In this section, we describe the data generated and used in our work.

A. Finer-grained ET data

The 18 entity types used in this study are the same as in the original NER4EL paper [3]. In addition to the standard types of `Person`, `Organization` and `Location`, additional types such as `Event`, `Animal`, and `Food`, were included. These types were all based on the `Misc` type that has traditionally been used in NER applications.

The fine-grained ET (fiNE mapping) data consists of pairs where the first entry corresponds to text with a delimited entity mention, and the second entry corresponds to the entity category. Entities from the Icelandic Wikipedia were labelled (assigned one of the fiNE types) and texts in Wikipedia that link to the entities were used as training data. For example [3]:

[2]https://github.com/studio-ousia/luke

[3]English translation: While he stayed in England, Holberg started to write an academic book ...

```
{
  "left context": "Á meðan dvöl hans í ",
  "mention": "Englandi",
  "right context": " stóð byrjaði
   Holberg að skrifa fræðirit ...
  "output": "LOCATION"
}
```

For the pages that have a corresponding page in the English Wikipedia, their English counterparts were found and then the original English NER4EL mapping was used to to find the category.

For pages without an English counterpart, we used a heuristic approach to label the entity mentions. Each of the 18 NER categories was located as one or more inner nodes in the Wikipedia Category Tree[4] of the Icelandic Wikipedia. All leaf descendants of such nodes correspond to Wikipedia pages and were thus assigned the given category[5]. The categories are not mutually exclusive so the resulting entities were manually reviewed for correctness. Entities with more than a single label and entities in wrong categories were discarded.

Using this process, around 75% of the pages in the Icelandic Wikipedia were labelled with the fiNE categories used in [3]. The finer-grained ET dataset contains $105,388$ examples, made from the Icelandic Wikipedia by using paragraphs from articles as the text context and using hyperlinks in the paragraphs, along with the category of the page which is linked to, as labels.

B. EL data

For EL, we made use of MIM-GOLD-EL, a corpus developed as a resource for Icelandic EL [5]. MIM-GOLD-EL is based on the MIM-GOLD-NER corpus [1], which, in turn, is based on the part-of-speech tagged MIM-GOLD corpus of 1 million tokens [17].

MIM-GOLD-EL was created by linking four types of NEs (PERSON, LOCATION, ORGANISATION, and MISC) to their corresponding entries in Wikipedia. The mGENRE model was applied to the entity mentions in the MIM-GOLD-NER corpus (see Section II) covering 46.1% of the entities. Each prediction by mGENRE was reviewed by a human annotator who accepted or rejected the prediction to create gold annotations. Wikipedia API search was used for the leftover entities that mGENRE could not identify, and each search with at least one result was reviewed by a human annotator to create further gold annotations. 53.9% of the entities from MIM-GOLD-NER were assigned to a Wikipedia entry using this annotation approach and 46.1% of the entities were not identified in Wikipedia.

We transformed MIM-GOLD-EL to a format similar to AIDA-CoNLL [11]. This was carried out so that the source code of LUKE could be used without major changes. Furthermore, we believe that the AIDA-CoNLL format is simpler and more beneficial for future work. In addition to creating an AIDA-CoNLL version of MIM-GOLD-EL, we created necessary files for the fine-tuning of IceLUKE. A file containing all persons found in the Icelandic Wikipedia, a file containing redirects, and a file containing all titles found in the Icelandic Wikipedia obtained from Wikimedia[6].

MIM-GOLD-NER contains no information on which sentences within each section originate from the same document. Hence, MIM-GOLD-EL contains very limited information that may be used to explore the surrounding context of each entity mention. We could only work with the words and entity mentions appearing within each sentence. To be able to properly evaluate IceLUKE on more context, we retrieved the original context from the Icelandic Gigaword Corpus (IGC) [18]. To distinguish between the two versions, we refer to them as the *sentence-level* version and the *document-level* versions, respectively.

Using the sentence-level version of MIM-GOLD-EL, we were able to use only the context found in the sentence in which an entity mention appears, whereas in the document-level version, we were able to use the context from the entire document in which an entity mention appears.

C. Alias table

For each entity mention m in a set of pre-identified entity mentions M, there is a collection of possible entities E_m that may refer to m. A common way to find E_m from m is to have a table that associates each mention with the entities that it can be used as an alias for. We constructed such an alias table.

The alias table is a mapping from strings to entity ids. For every entity id, the table contains all words that refer to the entity and all declensions of the words, i.e. all ways that entity can appear in text." For example, 'Berlín', 'Berlínar' and 'Berlínarborg' all map to the id for the city of Berlin. In a similar way, "borgarastyrjöld" ('civil war') can refer to 8 different pages in the Icelandic Wikipedia for different civil wars.

The different mentions that can refer to entities (the *aliases*) were found by crawling the Icelandic Wikipedia. From a list of all the titles of articles[7], all the pages were fetched and hyperlinks within the text, from an anchor (the alias) to another page (the entity), were put into a dictionary with the anchor as key and entity mention id as value.

Icelandic is a morphologically complex language and, therefore, The Database of Icelandic Morphology [19], encapsulated as a Python package[8], was used to enrich the dictionary with different forms that each word could take.

This resulted in a mapping between $176,026$ different aliases and $37,794$ entities. We note that around 75% of the examples in MIM-GOLD-EL map to a unique entity in the Icelandic Wikipedia.

[4]https://is.wikipedia.org/wiki/Kerfissíða:CategoryTree

[5]For example, in the Category Tree, an inner node called "Category: People" leads to the greek philosopher "Sextos Empeirikos" and this method would assign the label `Person` to mentions that refer to a page about him.

[6]https://dumps.wikimedia.org/iswiki/20220101/

[7]https://dumps.wikimedia.org/iswiki/latest/

[8]https://github.com/mideind/BinPackage

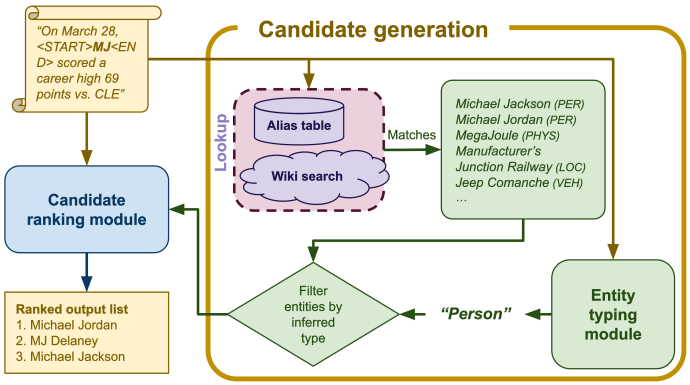

Fig. 1. Overview of the entity disambiguaton approach used in this work. The sequence of steps, starting from an input text that contains an entity mention and resulting in an ordered list of candidate entities, where the one that fits the context best is ranked highest. Candidates are generated (using an alias table and/or Wikipedia API search), filtered (using the ET module) and ranked (using IceLUKE or popularity baseline). Finally, the highest ranked candidate is linked to the entity mention.

IV. Methods

Figure 1 depicts the overall flow of our approach, where the input is a text with an annotated entity mention and the output is a ranked list of entities that match the text best. In between, ET is used to filter down the list of candidate entities to aid disambiguation. The ET module is in our case the fiNE typing model and the fiNE mapping is what assigns Michael Jackson the type `Person` and Manufacturer's Junction Railway the type `Location`.

The fiNE typing model is used to classify an input text that contains an annotated entity mention into one of the 18 categories for fine-grained ET, or 4 categories for coarse-grained ET. The candidates' entity types are looked up with the fiNE mapping and only the candidates who match the model's prediction are used in candidate ranking.

In NER4EL, BERT was used as the foundational model that was fine-tuned for the fiNE ET task. We used the multilingual XLM-RoBERTa model instead, which was recently fine-tuned for Icelandic and had the highest F1 score (92.5%) on the MIM-NER-GOLD evaluation set [2].

A. Performance evaluation

Performance of an EL model is typically measured using precision, recall, and F1 score. However, when entity mentions are part of the input, such as in the case of ED without end-to-end EL, accuracy is used as the evaluation metric. Accuracy is the ratio of correctly linked entities to the total number of entity mentions. In this paper, we report in-KB accuracy, which only considers entity mentions with valid KB entities for evaluation.

B. Candidate generation

For fine-tuning and evaluating ED, we created a set of candidates along with the correct one for each entity mention in MIM-GOLD-EL. To generate candidates, we used Wikipedia search and alias table lookup.

Wikipedia search: We performed a Wikipedia search on each entity mention to build a candidate set and we included up to 16 highest ranked articles. The search was performed programmatically using the Wikipedia search API and it was restricted to Icelandic articles.

Alias table lookup: Given a text with an entity mention, we searched for the mention in the alias table to find entities associated with it.

For the examples in the test set, 73% of the entity mentions only mapped to one entity in the alias table. 14% mapped to two and 13% mapped to more than two. This lack of ambiguity has to do with the fact that the Icelandic Wikipedia has only 55 thousand articles and, therefore, does not contain all possible entities a name could refer to.

We note that there is no general size requirement standard for an entity candidate set. The average size of a candidate set per entity mention in the TAC-KBP2010 dataset is 12.9 and the average size in the TAC-KBP2011 dataset is 13.1 [20]. Reference [16] used upward of 30 candidates.

C. Baseline models

The baseline models for this work are based on various Transformer-based models (IceBERT-igc [2], ConvBERT-small, ConvBERT-base, ELECTRA-small, and ELECTRA-base [21]) that have been pre-trained on Icelandic data and made available on HuggingFace. We fine-tuned these models to classify whether an entity mention in the MIM-GOLD-EL corpus matches the description of an entity retrieved from Wikipedia. The process was set up as a sequence classification task, with each model being evaluated using 10-fold cross-validation. The performance of these models was evaluated using a candidate set of up to 16 candidates for each mention in the test set.

To set up ED as a sequence classification task, we first retrieved the textual description of the entities from Wikipedia and paired them with entity mentions from MIM-GOLD-EL (sentence-level). We then fine-tuned models to classify whether the entity mention matched the textual description or not. In this manner, a mention-description pair can be evaluated and it receives a score that can be interpreted as how well the description matches the mention. That score can be used to rank multiple candidates for a given mention.

D. Rule-based baseline

In order to investigate how important it is to use machine-learning to solve the task, we implemented a simple rule-based baseline. Given an example from our evaluation set, it selects, from the set of candidate entitites, those entities whose Wikipedia titles share the longest common substring with the entity mention in the example. Then, from this set of 1–3 candidates, it selects the candidate whose Wikipedia page has the highest number of incoming links from other pages (the in-degree of the node in the network). This method does not use the context from the text that surrounds the entity mention at all. For the rule-based ranking, we do the following:

1) Compute the longest common subsequence of the entity mention and all Wikipedia page titles.
2) Normalize results by

$$\min\left(\text{length(mention)}, \text{length(title)}\right).$$

3) Choose the highest scoring result if the scores are greater than 0.5. We break score ties by choosing shorter titles unless the mention has 1-2 letters, then we break ties by choosing candidates that start with the same letters as the mention.

In the special case when all scores are smaller than 0.5, we use gestalt pattern matching and pick the highest scoring title. In case that results in a tie, we choose a candidate starting with the same letter as the mention.

E. IceLUKE for Entity Disambiguation

Pre-training: LUKE uses RoBERTa [22] as a base pre-trained model and the pre-training was continued on entity-annotated data obtained from Wikipedia.

We retrieved an Icelandic Wikipedia dump from January 1, 2022[9] and used it to create the pre-training data for Icelandic. We used IceBERT-igc[10], which is trained using the RoBERTa architecture, as our base pre-trained model, and the Icelandic pre-training corpus, described in Section III-B, for continued pre-training of the model for ED. We followed the two-stage process described in [16] using the same hyperparameters. We call the resulting model IceLUKE for ED.

Fine-tuning: We fine-tuned and evaluated IceLUKE for ED both for local and global contexts using the sentence-level and document-level versions of MIM-GOLD-EL, respectively, and with and without candidates from alias table lookup.

Additionally, we fine-tuned and evaluated IceLUKE for ED using two filters, a coarse filter based on the standard four entity types and a finer filter based on the fiNE types, see Section III-A. Note that we omit sentences and documents exceeding 512 tokens.

F. Code and Hardware

For training and evaluation of all models, we used Colab Pro+ and a Tesla V100-SXM2-16GB GPU. The code and files for fine-tuning and evaluation is publicly available on GitHub[11].

V. Results

In this section, we first present the accuracy of IceLUKE in comparison with other ED models. Second, we compare the evaluation results for different versions of IceLUKE, using confidence-order and the default approach, with and without using an ET filter, and compare it with the rule-based approach.

Table I shows the results for ELECTRA-base, ConvBERT-base, IceBERT-igc, and IceLUKE when fine-tuned and evaluated using the sentence-level version of MIM-GOLD-EL.

[9]https://dumps.wikimedia.org/iswiki/20220101/

[10]https://huggingface.co/mideind/IceBERT-igc

[11]https://github.com/bennigeir/ice-luke

TABLE I
ACCURACY FOR ICELUKE FINE-TUNED AND EVALUATED ON THE SENTENCE-LEVEL VERSION OF MIM-GOLD-EL USING UP TO 16 CANDIDATES PER ENTITY MENTION OBTAINED ONLY FROM WIKIPEDIA SEARCH. COMPARISON IS PROVIDED AGAINST BASELINE MODELS.

Model	Accuracy (%)
ELECTRA-base	74.7
ConvBERT-base	74.7
IceBERT-igc	73.3
Rule-based	78.6
IceLUKE	**88.4**

TABLE II
ED EVALUATION ON A HOLDOUT SET OF THE DOCUMENT-LEVEL VERSION OF MIM-GOLD-EL. RESULTS REPORTED FOR TWO TYPES OF CANDIDATE GENERATION, WIKIPEDIA SEARCH (W) AND ALIAS TABLE LOOKUP (A), AND WITH AND WITHOUT FILTERING ENTITY TYPES (CF=COARSE FILTERING, FF=FINE FILTERING). BOLD NUMBERS INDICATE HIGHEST PERFORMANCE IN A GIVEN COLUMN. THE CO COLUMN REFERS TO CONFIDENCE ORDERING OF DISAMBIGUATION. THE RULE-BASED BASELINE CORRESPONDS TO RANKING THE CANDIDATES WITH A PROXY FOR POPULARITY.

Setup	Acc. (%)	Acc. (%, CO)
IceLUKE		
- w	93.3	93.7
- w + cf	**95.0**	95.0
- w + ff	94.9	95.1
- w + a	93.2	93.4
- w + a + cf	94.8	94.9
- w + a + ff	**95.0**	**95.2**
Rule-based		
- w	84.0	-
- w + cf	85.3	-
- w + ff	85.7	-
- w + a	82.8	-
- w + a + cf	84.0	-
- w + a + ff	84.5	-

We observe that IceLUKE achieves an accuracy of 88.4%, outscoring all other models by a large margin. For candidate generation, we only used Wikipedia search and no ET filtering.

Table II shows the accuracy of IceLUKE evaluated on the document-level version of MIM-GOLD-EL using two different approaches to capture context. The default approach resolves entities in the order in which they appear in the text. The confidence-order approach uses a greedy algorithm that starts by disambiguating those entity mentions that have the most confident prediction of all unresolved entity mentions. The highest scoring method, confidence-order, achieves an accuracy of 95.2%.

Note that the results for IceLUKE in Table I are different from those presented in Table II, since we do not use the same version of MIM-GOLD-EL.

Table III shows the performance on the different sources of data that make up the document-level test set for MIM-GOLD-EL. IceLUKE with fine-grained ET achieves higher accuracy than the rule-based baseline for all but two sources.

TABLE III
ACCURACY RESULTS FOR ICELUKE (CONFIDENCE ORDER, WITH A FINE-GRAINED ENTITY TYPING FILTER) AND FOR THE SIMPLE RULE-BASED BASELINE, ON THE DOCUMENT-LEVEL VERSION OF MIM-GOLD-EL, SHOWING ACCURACY FOR THE DIFFERENT SOURCES OF TEXT WHICH MAKE UP MIM-GOLD-EL.

Source	Support	IceLUKE (%)	Rule-based (%)
Blogs	268	98.1	83.5
Books	96	91.7	70.8
Newspaper *Fbl*	237	93.3	82.3
Laws	14	100.0	100.0
Newspaper *Mbl*	307	96.1	77.5
Radio news scripts	20	80.0	85.0
School essays	46	97.9	95.7
The Icelandic WoS	112	95.5	84.8
Websites	32	90.6	96.8

VI. DISCUSSION

In this study, we proposed an approach to ED for Icelandic that utilizes the NER4EL method and a new, entity-aware LM, which we call IceLUKE, fine-tuned on the MIM-GOLD-EL corpus.

In our experiments, we saw a great benefit in using IceLUKE to disambiguate entities when compared with models without an entity-aware mechanism. One limitation of our study is that we did not compare the models on the document-level version of MIM-GOLD-EL. However, given the large difference in accuracy between the models (13.7%, see Table I), we would not expect the other models to perform better than IceLUKE with more context.

For filtering entities, we saw a small benefit. However, for fine-grained entity type filtering, we were surprised by the little benefit it provided when compared to a coarse filter. This might be because the evaluation sets do not contain sufficiently varied entity types, even though they are sampled from a diverse set of sources. Furthermore, IceLUKE might be sufficiently good at disambiguating, such that it can tolerate having a coarse filter. However, we saw a small benefit of using fine-grained filtering for the rule-based approach, possibly indicating that weaker ranking models benefit more from it.

We observed that filtering using the fine types only improves accuracy when the ranking model confuses entities within the `Misc` category. For example, if IceLUKE ranks Orion spacecraft (*Vehicle*) highest when the true entity mention is the Orion nebula (*Celestial*) then we would see a benefit from filtering when the ET model predicts Celestial.

For the English Wikipedia, with its millions of entities, it matters more to have fine-grained entity types. For Icelandic, there are not many `Foods` that share names with `Vehicles`, i.e. it rarely happens that the same candidate set has entities that are both `Foods` and `Vehicles`. However, fine-grained filtering should not be disregarded, since with a larger KB, we would expect to see a greater benefit from ET.

Disambiguating using nothing else than string matching between entity mention and candidate title, and the in-degree of the Wikipedia page as a prior, resulted in a relatively high accuracy. In fact, it exceeded the performance of all models except IceLUKE on the sentence-level version of MIM-GOLD-EL. Incorporating such priors into a LM has the potential to further aid in disambiguation.

While our approach has shown promising results, there are several limitations. One is the use of the MIM-GOLD-EL corpus, which is a relatively small corpus that was specifically created for Icelandic EL tasks. Therefore, the results may not generalize well to other datasets or real-world applications. Another limitation is that the model is only able to link entities to the Icelandic Wikipedia, which may not always be the most relevant or up-to-date KB for a specific entity. Furthermore, the candidate generation method is based on the Wikipedia search API, which can affect reproducability. Future work could involve training the model on a larger and more diverse corpus, and also incorporating other KBs to improve the performance of the model.

Another limitation concerns comparisons to work on English EL. To have a fair comparison to the English version of LUKE, it would have been ideal to use the same approach, i.e. to choose the 30 highest-ranked entities according to the mention-entity prior $\hat{p}(e|m)$ from [23]. The mention-entity prior is computed using the ratios of the number of hyperlinks into a Wikipedia page that have the mention as anchor text. The problem is, however, that the Icelandic Wikipedia has only 55 thousand articles, whereas the English one has 6.6 million[12]. Therefore, there simply are not any mentions that link to ≥ 30 different pages. Only 1.7% of the mentions in the examples used in the MIM-GOLD-EL evaluation sets link to ≥ 5 pages. Therefore, we relied upon the alias table, which assigns no more than a few candidates to each entity mention, and the Wikipedia search, which finds up to 16 candidates for an entity mention.

VII. CONCLUSION

In this study, we presented a novel approach to ED for Icelandic, which has low resources in the domain of EL, utilizing insights from the NER4EL method and a new entity-aware LM, IceLUKE. Our results showed a significant improvement in ED accuracy compared to baselines, with our best approach achieving 95.2% accuracy. We observed a minor improvement in the use of fine-grained entity filtering after generating candidates, and great improvements of using IceLUKE in the candidate ranking step.

In conclusion, our study provides a step towards more effective and efficient EL systems for low-resource languages and highlights the potential for entity-aware LMs in such domains.

ACKNOWLEDGMENTS

This work was funded by the Icelandic Strategic Research and Development Program for Language Technology 2021, grant no. 200075-5301.

[12]https://en.wikipedia.org/wiki/List_of_Wikipedias

References

[1] S. L. Ingólfsdóttir, Á. A. Guðjónsson, and H. Loftsson, "Named Entity Recognition for Icelandic: Annotated Corpus and Models," in *Statistical Language and Speech Processing*, L. Espinosa-Anke, C. Martín-Vide, and I. Spasić, Eds. Cham: Springer International Publishing, 2020, pp. 46–57.

[2] V. Snæbjarnarson, H. B. Símonarson, P. O. Ragnarsson, S. L. Ingólfsdóttir, H. Jónsson, V. Thorsteinsson, and H. Einarsson, "A Warm Start and a Clean Crawled Corpus – A Recipe for Good Language Models," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 4356–4366.

[3] S. Tedeschi, S. Conia, F. Cecconi, and R. Navigli, "Named Entity Recognition for Entity Linking: What Works and What's Next," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2584–2596. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.220

[4] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6442–6454. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.523

[5] S. R. Friðriksdóttir, V. Á. Eggertsson, B. G. Jóhannesson, H. Daníelsson, H. Loftsson, and H. Einarsson, "Building an Icelandic Entity Linking Corpus." in *Proceedings of the workshop 'Dataset Creation for Lower-Resourced Languages', at the 13th International Conference on Language Resources and Evaluation (LREC 2022).*, Marseille, France, 2022.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[7] N. Kolitsas, O.-E. Ganea, and T. Hofmann, "End-to-End Neural Entity Linking," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 519–529. [Online]. Available: https://www.aclweb.org/anthology/K18-1050

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[9] J. Raiman, "DeepType 2: Superhuman Entity Linking All You Need Is Type Interactions," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, pp. 8028–8035, 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20774

[10] W. Shen, Y. Li, Y. Liu, J. Han, J. Wang, and X. Yuan, "Entity Linking Meets Deep Learning: Techniques and Solutions," *CoRR*, vol. abs/2109.12520, 2021. [Online]. Available: https://arxiv.org/abs/2109.12520

[11] M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum, "AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables," *Proceedings of the VLDB Endowment*, vol. 4, no. 12, pp. 1450–1453, 2011.

[12] N. Botzer, Y. Ding, and T. Weninger, "Reddit entity linking dataset," *Information Processing & Management*, vol. 58, no. 3, p. 102479, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457320309687

[13] J. A. Botha, Z. Shan, and D. Gillick, "Entity Linking in 100 Languages," *ArXiv*, vol. abs/2011.02690, 2020.

[14] N. De Cao, L. Wu, K. Popat, M. Artetxe, N. Goyal, M. Plekhanov, L. Zettlemoyer, N. Cancedda, S. Riedel, and F. Petroni, "Multilingual Autoregressive Entity Linking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 274–290, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.16

[15] N. De Cao, G. Izacard, S. Riedel, and F. Petroni, "Autoregressive Entity Retrieval," *arXiv preprint arXiv:2010.00904*, 2020.

[16] I. Yamada, K. Washio, H. Shindo, and Y. Matsumoto, "Global Entity Disambiguation with BERT," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 3264–3271. [Online]. Available: https://aclanthology.org/2022.naacl-main.238

[17] H. Loftsson, J. H. Yngvason, S. Helgadóttir, and E. Rögnvaldsson, "Developing a PoS-tagged corpus using existing tools," in *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, ser. LREC 2010, F. M. T. Sarasola, Kepa and M. L. Forcada, Eds., Valetta, Malta, 2010.

[18] S. Steingrímsson, S. Helgadóttir, E. Rögnvaldsson, S. Barkarson, and J. Guðnason, "Risamálheild: A Very Large Icelandic Text Corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018.

[19] K. Bjarnadóttir, K. I. Hlynsdóttir, and S. Steingrímsson, "DIM: The Database of Icelandic Morphology," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland: Linköping University Electronic Press, 2019, pp. 146–154.

[20] H. Ji, R. Grishman, and H. Dang, "Overview of the TAC2011 Knowledge Base Population Track," in *TAC 2011 Proceedings Papers*, 2011.

[21] J. F. Daðason and H. Loftsson, "Pre-training and Evaluating Transformer-based Language Models for Icelandic," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 7386–7391.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv*, vol. abs/1907.11692, 2019.

[23] O.-E. Ganea and T. Hofmann, "Deep Joint Entity Disambiguation with Local Neural Attention," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2619–2629. [Online]. Available: https://aclanthology.org/D17-1277