# Comparison of Multilingual Entity Linking Approaches

**Ivelina Bozhinova**
Solutions Unit, Ontotext AD
79 Nikola Gabrovski St
Sofia, Bulgaria
ivelina.bozhinova@ontotext.com

**Andrey Tagarev**
Research Unit, Ontotext AD, Sofia, Bulgaria
and
Computer Science, University of Sheffield
andrey.tagarev@ontotext.com

## Abstract

Despite rapid developments in the field of Natural Language Processing (NLP) in the past few years, the task of Multilingual Entity Linking (MEL) and especially its end-to-end formulation remains challenging. In this paper we aim to evaluate solutions for general end-to-end multilingual entity linking by conducting experiments using both existing complete approaches and novel combinations of pipelines for solving the task. The results identify the best performing current solutions and suggest some directions for further research.

## 1 Introduction

Entity linking (EL) (Hoffart et al., 2011), (Cucerzan, 2007) is the task of mapping mentions in unstructured text to entities in an existing Knowledge Base (KB). It has drawn the attention of many researchers in the past few years due to its application in different areas of NLP, including Question Answering (De Cao et al., 2019), (Yin et al., 2016), (Wang et al., 2021), Relation Extraction (Baldini Soares et al., 2019), Dialogue (Chen et al., 2017a), (Bordes et al., 2017), (Wen et al., 2017) and Biomedical systems (Bhowmik et al., 2021), (Zheng et al., 2015). Even though there has been a significant improvement in the field recently (Cao et al., 2021), (Wu et al., 2020), (Ayoola et al., 2022), the task of EL and especially in the cross-lingual (Ji et al., 2015), (McNamee et al., 2011), and MEL setups remain challenging. Different approaches have been proposed for solving this task, some of which are based on more traditional methods (Brank et al., 2017), (Delpeuch, 2020) and others exploit the recent discoveries in the field of natural language processing (Cao et al., 2021), (Wu et al., 2020), (Ayoola et al., 2022), (Botha et al., 2020). This paper will present experiments comparing the performance of various methods for a MEL task.

## 2 Multilingual Entity Recognition and Disambiguation Methods

In EL, also known as named-entity recognition and disambiguation (NERD) words of interest in an unstructured text are mapped to corresponding unique entities in an existing target KB. Formally it can be defined as the task of linking a given entity mention $m$ in a given context $c$ to the corresponding entity e in a KB. For the multilingual definition a set of languages L is added and the context is defined as language specific (context $c$ of language $l$). It also requires a multilingual KB. As the name NERD suggests, the task consists of two subtasks, namely named-entity recognition (NER) and entity disambiguation (ED). Two general groups of EL methods exist. One focuses on performing entity disambiguation but requires correctly annotated entities or at least entity spans in its input. The second takes plain text input and performs both recognition and disambiguation in one or more steps.

NER (Sundheim, 1995) is a fundamental task in NLP which consists of recognising entities in text, and identifying their types. In the past years, different approaches have been developed, including statistical machine learning methods (Zhou and Su, 2002), (Agerri et al., 2014), neural networks based ones (Strubell et al., 2017), (Xia et al., 2019) and a combination of both (Huang et al., 2015), (Chen et al., 2017b). The recent advances in the field of NLP introduced the application of richer contextual embeddings computed via Transformer models (He et al., 2021), (Devlin et al., 2019), (Vaswani et al., 2017) and have significantly improved the state-of-the-art (SOTA) of the task. In particular, these impressive results were achieved on benchmark datasets such as CoNLL03 (Tjong Kim Sang and De Meulder, 2003) and OntoNotes (Hovy et al., 2006). Nevertheless, it has been stated that the



*Proceedings of Recent Advances in Natural Language Processing*, pages 224–233
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_025

reason for this improvement lays not only on the model, but the fact that the benchmark datasets lack the presence of multiple practical challenges and these new models actually have problems detecting and classifying complex or unseen entities (Augenstein et al., 2017), (Meng et al., 2021). To address this issue, the dataset MultiCoNER (Malmasi et al., 2022) was developed which includes complex entity mentions with higher granularity on the type definition. With the introduction of the MultiCoNER2 task (Fetahu et al., 2023), which focus on tackling multilingual named entity recognition (NER) in fine-grained and noisy scenarios, a lot of promising approaches for general entity recognition have been proposed (García-Ferrero et al., 2023), (Tan et al., 2023).

mGENRE (De Cao et al., 2022), a MEL system based on autoregression, has emerged as the state-of-the art as measured on the major multilingual datasets. It is therefore a main focus of the experiments in our work.

## 3 Evaluated Approaches

In our work we experiment with three different end-to-end EL approaches: a classical approach, Wikifier (Brank et al., 2017), and two systems that we build as a combination of mGENRE (De Cao et al., 2022) with a multilingual NER model and a multilingual EBD method.

### 3.1 Wikifier

Wikification is a simple approach for multilingual text annotations, a process in which a text is annotated with relevant concepts from Wikipedia. Each Wikipedia article is treated as a Wikipedia concept and the relations between the concepts are expressed by the links between the articles. Wikipedia is large, multilingual and contains general knowledge and is therefore a popular choice for a target database in different entity linking approaches.

We experiment with a Wikifier (Brank et al., 2017) based on page rank and global disambiguation which provides a semantic annotation in 100 languages. Instead of trying to detect separate entities in the text and then map them to corresponding Wikipedia concepts, Wikifier sees the text as a whole and aims at finding suitable annotations which are supported by multiple mentions of the text. In this way it follows the intuition that most mentions in a text should be similar and related to common topics. Based on the page rank of a concept and its support by mentions in the text, a decision is made if it is a suitable annotation for the given text. When returning the final list of annotations for a text, Wikifier does not return the exact mention match for the concept, but a list with all mentions in the text which support the annotation. Wikifier is available as a public web service which we used for our experiments.

### 3.2 mGENRE Disambiguation

mGENRE (multilingual GENRE) (De Cao et al., 2022) is a system for general MEL, which predicts the label of the corresponding entity in a multilingual KB from left to right, token-by-token using autoregression which enables it to effectively cross-encode mention and entity labels to capture more interactions than the standard dot product between mention and entity vectors. It is also capable of fast search in KBs even for mentions that are not part of mention tables and without need of large-scale vector indices. In contrast to most MEL approaches which implement a single representation for each entity, mGENRE maps against entities in multiple languages and with that enables exploiting relations between mention in text and target name.

It also works in a zero-shot setting for languages without any training data, since it processes the target language as a latent variable and marginalises it during prediction. mGENRE ranks each element in a knowledge base by computing a score with an autoregressive formulation. It is based on a fine tuned mBART (Liu et al., 2020) architecture. Beam search is used to pre-select top-k linking candidates for each entity. GENRE employs a prefix tree (trie) to enable constrained beam search and then generate only valid entity identifiers. In order to extend GENRE in multilingual settings, the authors use canonical entity representation and multilingual entity representation for training and marginalisation during testing and inference.

mGENRE has achieved SOTA results for MEL on several datasets ( Mewsli-9 (Botha et al., 2020), TR2016 (Tsai and Roth, 2016), KBP2015 (Ji et al., 2015)) and is currently the best general MEL system so we have decided to use it in our experiments and combine it with a suitable entity (boundary) detection algorithms. In our experiments, we apply the pre-trained mGENRE model provided by its authors which is fine-tuned an mBART (Liu et al., 2020) model that had been pre-trained on

125 languages using Wikipedia hyperlinks in 105 languages.

### 3.3 EBD + mGENRE Entity Disambiguation

The entity boundary detection (EBD) (García-Ferrero et al., 2023) is a transformer-based multilingual masked language model pre-trained on text in 100 languages (Conneau et al., 2020), and works as follows: Given unlabelled text as input, it predicts the boundaries of a named entity by analysing the structure of the input sentence. This task is presented as a sequence labelling task in which the model predicts for each token if it is part of an entity or not by classifying it in one of the categories: "B-ENTITY", "I-ENTITY", and "O", where "B-ENTITY" stands for beginning of an entity, "I-ENTITY" is for inside an entity and "O" means no part of entity.

The approach is based on a multilingual XLM-RoBERTa-large model (Conneau et al., 2020) with a linear token classification layer on top of each token representation. Its is based on the sequence labelling implementation of the Huggingface open-source library (Wolf et al., 2020). Five different independent models have been trained and then a majority vote has been used as the ensemble strategy at inference time. No trained model was available, however, there were instructions and code available on how to replicate the training of the models. Therefore we followed these instructions and trained five different models, choosing the best one afterwards using the same strategy described in the paper.

The boundaries detected by the EBD model are then processed using the mGenre model presented in the previous subsection.

### 3.4 SpaCy multilingual NER + mGENRE Entity Disambiguation

SpaCy (Honnibal and Montani, 2017) is an open-source Python library focusing on advanced NLP. Currently SpaCy supports more than 70 languages and provides pre-trained pipelines for NER. SpaCy comes with a separate pipeline for each of the languages. While a multilingual model exists, it is quite small and limited so individual language pipelines need to be used. However, since it is one of the most used and reliable libraries for NLP (Lorica and Nathan, 2021) we consider it an interesting candidate for performing the entity recognition part of an end-to-end entity linking system.

SpaCy returns the start and end indices for each annotation so it can be combined with mGENRE EL in the same way as the EBD model described previously.

## 4 Experiments

### 4.1 Datasets

In our experiments we use two datasets, one freely available multilingual dataset Mewsli9 (Botha et al., 2020), which contains mentions linked to Wikidata and one custom dataset, consisting of documents in three languages extracted from the Database of Known Fakes (DBKF) (Tagarev et al., 2021). The choice of Mewsli-9 is justified by the fact that mGENRE has already been tested on it and therefore using Mewsli-9 will allow us to compare the entity disambiguation of Wikifier and mGENRE. On the other hand, Mewsli-9 is an entity disambiguation dataset in which not all mentions have been tagged and therefore it is not a suitable dataset for testing end-to-end entity linking. For this reason, instead of using another entity disambiguation dataset, we chose to compare overall performance of the three approaches on a small selection of text from the DBKF (Tagarev et al., 2021). It is multilingual, it contains fact checking news articles on recent events which can be more challenging to link to a KB. These texts would give a better view on how the tested systems perform in a real world scenario.

#### 4.1.1 Mewsli9 Dataset

Mewsli-9 (Botha et al., 2020) (short for "Multilingual Entities in News, linked") is a large multilingual dataset which contains nearly 300,000 mentions across 9 languages from different language groups (English, German, Spanish, Arabic, Serbian, Japanese, Turkish, Persian, Tamil). The dataset is freely available and each mention is linked to a WikiData item, which makes the dataset suitable for our experiments.

An interesting feature of the dataset is that it contains many entities that lack English Wikipedia pages and which are thus not accessible to a lot of cross-lingual systems. Mewsli-9 consists of 289,087 entity mentions (with no predefined splits) which are to be found in 58,717 originally written news articles from WikiNews, covering different genres. In contrast to other multilingual datasets, which cover only European languages (e.g. VoxEL (Rosales-Méndez et al., 2018)), the Mewsli-9 cor-

pus contains languages which represent five language families and six orthographies. The dataset is however not balanced between the languages.

#### 4.1.2 DBKF Dataset

Apart from Mewsli9 we also use a small selection of debunks from the Database of Known Fakes (DBKF) (Tagarev et al., 2021) consisting of 90 documents in three languages, English, German and Spanish. The test dataset contains two document types, claims and claim reviews. Claims are short texts describing a (false) claim and reviews are whole debunking articles. The documents are not annotated with ground truth annotation, which means that during evaluation only precision could be measured. An approximation of recall can be estimated based on the total number of unique valid annotations produced by the three systems.

### 4.2 Experimental Design

As the goal of the paper is to explore and compare end-to-end entity linking systems, we have defined two types of experiments covering different parts of the tested approaches.

The first is to run Wikifier on Mewsli-9 dataset. Since mGENGRE achieves state-of-the-art results on Mewsli9 and we want to allow comparison between the two approaches, we have decided to test Wikifier on Mewsli-9. Such an experiment focuses on evaluation of the entity disambiguation part, but we also try to analyse the overall performance based on the results.

The second is to compare the three end-to-end entity linking solutions on the DBKF extract. The three solutions compared, as described in Section 3, are Wikifier, EBD + mGENRE and SpaCy + mGENRE.

## 5 Results

### 5.1 Results on Mewsli-9

We first evaluate the performance of Wikifier on the Mewsli-9 dataset in order to compare performance with mGENRE disambiguation. The results are shown in Table 1. Clearly, applying Wikifier on the dataset provides an immediate challenge in that Wikifier doesn't simply link already annotated entities but discovers them within the text. This leads to a significant mismatch in recognised entities between Wikifier and the gold standard (Note: here a partial overlap is treated as two mismatches).

In order to compare the performance of the algorithm to the existing approaches, we define a precision score that is applied only to entities that are in the gold standard and recognised by Wikifier. This means the results are not completely comparable but they are calculated over a subset of the Mewsli-9 annotations.

Table 2 shows the results of running the mGENRE model on Mewsli-9 (De Cao et al., 2022). While technically the numbers for accuracy over the whole dataset are lower than the precision of Wikifier, it is important to consider that the Wikifier precision is only calculated on a subset of the annotation.

At this point we need to consider the two major concerns with our approach to evaluating Wikifier on the Mewsli-9 dataset. They both stem from the fact that the Named Entity Recognition (NER) has a significant mismatch. Immediately relevant is the issue with gold standard entities that are not recognised by Wikifier. Referencing Table 1 again, we see that Wikifier in fact fails to precisely recognize over 40% of all annotation in the gold standard.

On the other hand is the concern that Wikifier recognizes many concepts that are not part of the gold standard and cannot be evaluated. Actually there are almost three times as many entities tagged by Wikifier than can be found in the gold standard and it is important to understand what is in there. We have expected this behaviour, since as already mentioned Mewsli-9 is a EL datasets in which not all mentions are tagged. In order to achieve a fair comparison of the three tested systems, we proceeded with manually evaluated experiments on our custom dataset.

### 5.2 Results on Manual Evaluation

For the next part of the experiments we annotated all 90 documents from our custom dataset with all three systems of interest. We then randomly selected a subset of all annotations (200 per system) that were annotated by multiple annotators reaching agreement. The evaluation included two judgements- entity recognition and entity disambiguation. For the first step, we defined three possibilities, exact, partial and false as we also want to examine if entities which are not exactly detected by the first step of an approach can be correctly linked to Wikidata by the Entity Disambiguation part of the systems. In other words, check whether the ED step is capable of fixing errors of NER or

| Lang | Errors | Only WF | Only GS | Both | Precision | Recall | F1 |
|---|---|---|---|---|---|---|---|
| en | 1235 | 173483 | 38560 | 41093 | 0.96 | 0.51 | 0.67 |
| de | 1378 | 173910 | 21114 | 43807 | 0.96 | 0.67 | 0.79 |
| es | 1187 | 152925 | 22495 | 33240 | 0.96 | 0.59 | 0.73 |
| ar | 37 | 42846 | 3442 | 3166 | 0.98 | 0.43 | 0.60 |
| fa | 9 | 1925 | 214 | 307 | 0.97 | 0.57 | 0.72 |
| ta | 28 | 9156 | 1588 | 1098 | 0.97 | 0.41 | 0.58 |
| tr | 78 | 7015 | 3272 | 2464 | 0.96 | 0.42 | 0.59 |
| ja | 134 | 108708 | 16563 | 17741 | 0.99 | 0.51 | 0.68 |
| sr | 543 | 68643 | 13982 | 21687 | 0.97 | 0.61 | 0.75 |
| all | 4629 | 738611 | 121230 | 164603 | 0.97 | 0.57 | 0.72 |

Table 1: Results form running Wikifier over the Mewsli-9 dataset.

EBD and with that can improve the overall performance. The second step includes evaluation of ED in which again three categories were defined: correct, wrong and invalid entity. The latter category is defined when no entity to link exists within the span. The results for all evaluated systems can be seen in Table 3. It is important to note here that the columns presenting the ED results ("ED(ve)" and "ED(vp)") show the accuracy of the ED only on the correctly recognised entities, exact and partially. In this way we want to assess the ED of each system independently from its mention detection part. Column "end-to-end EL" presents the overall accuracy of each system and is the best indicator for the performance of the whole system. Since our custom data is not previously annotated, we cannot formally analyse the recall of the entity linking performed. However, we could infer an estimated recall based on the results that we have combined with the total number of annotations for the whole dataset for each system (presented in column "Total number of annotations"). From the results presented in 3 we can conclude the following:

| Lang | Accuracy |
|---|---|
| ar | 94.7 |
| de | 91.5 |
| en | 86.7 |
| es | 90 |
| fa | 94.6 |
| ja | 89.9 |
| sr | 94.9 |
| ta | 92.9 |
| tr | 90.7 |
| micro | 90.2 |
| macro | 91.8 |

Table 2: Reported results of mGenre model on Mewsli-9 dataset.

- SpaCy produces the highest number of annotations, however also the highest number of incorrect ones. The general performance of the SpaCy + mGENRE system on the manually annotated annotations is also lowest. We assume that the recall for the system is quite high, however its low accuracy makes it less reliable in comparison to the other two systems.

- When linking exactly extracted entities, mGENRE performs very well and combined with EBD achieves results comparable with the ones reported in the paper (around 90% accuracy). In a combination with SpaCy, on the other hand it performs worse (80% accuracy). We suspect the reason is that SpaCy detects many annotations of types date and cardinal, which are then wrongly linked to unrelated Wikidata items by mGENRE. mGENRE also works well with partial entities (around 80% in both systems) which is a good indicator that mGENRE is capable of "fixing" errors with respect to the extraction of the mention.

- EBD has a very low score when considering the exact matches (66%), however it achieves a very good result of over 90% correctly recognised entities when we loosen the restriction on correctness and allow partially matched entities. The overall performance of the EBD-mGENRE systems in terms of accuracy is also satisfactory (75%), but notably lower than the overall accuracy achieve by Wikifier (86%).

| EL System | NER (e) | NER (p) | ED (ve) | ED (vp) | end-to-end EL | Total number of annotations |
|---|---|---|---|---|---|---|
| WF | 88,5 | 99 | 93,2 | 88,3 | 87,5 | 482 |
| SpaCy | 62 | 79,5 | 81,2 | 78 | 63 | 2398 |
| EBD | 66 | 92 | 90 | 82 | 75,5 | 618 |

Table 3: Accuracy in % for all end-to-end EL systems for each step. The first column is the name of the EL system, WF for Wikifier, SpaCy for SpaCy + mGENRE, and EBD for EBD + mGENRE. Column NER(e) shows the percentage of exactly recognised entities, column NER(p)- partially recognised entities. Columns ED(ve) and ED(vp) describe the results for the Entity Disambiguation part for valid exactly recognised and valid partially recognised entities, respectively. The column end-to-end EL shows the overall performance of the system and the last column presents the total number of annotations for each model on all documents.

- Wikifier achieves the best accuracy results in single components of the system and also end-to-end. This result is expected since Wikifier is not a true EL system. It does not link a concrete part of the text (mention) to an entity in a KB, but instead it sees the text as a whole and finds Wikipedia article which are related to the it. Wikifier, however, produces the lowest number of annotations overall (482) which means the inferred upper bound on recall is quite low (e.g. we estimate EBD annotated 90 additional accurate concepts over the dataset).

- We also noticed that Wikifier has difficulties detecting entities in short text. For 11 of the 90 documents, Wikifier produced no annotations. All these 11 documents are short documents (one or two sentences) in English. For comparison the other two systems found annotations in 87 (EBD + mGENRE) and 90 (SpaCy + mGENRE) documents.

- EBD + mGENRE seems like a good balance between precision and recall. However, its law accuracy requires further improvement.

### 5.3 Effect mGENRE Linking Threshold

| t | -0.2 | -0.3 | -0.4 | -0.5 |
|---|---|---|---|---|
| missed | 11.9 | 6.6 | 4.6 | 1.3 |
| fixed | 77.1 | 66.7 | 58.3 | 50 |

Table 4: Comparison of the trade off between correct missed and wrong fixed for different mGENRE thresholds. The ”missed” row analyses the percentage of correctly recognised (both exact and partial) and linked entities which would be discarded for each threshold value presented in the columns. The ”fixed” column present the percentage of the wrongly extracted or linked entities which are discarded when applying the corresponding threshold.

Further analysis suggests a method to improve the accuracy of the EBD + mGENRE system. Alongside the best linking candidate, mGENRE also returns a score. We decided to experiment with a threshold for this score and discard all annotations which return a score below the threshold. We hope to remove wrongly detected (or linked) entities while not losing many of the correctly recognized and linked ones. Table 4 presents the trade off between the discarded correct entities (column ”missed”) and the removed wrongly detected or linked entities (”column fixed”) for various thresholds. Our result show a clear connection between mGENRE score and correctness of the detected entities. We conclude that mGENRE is capable of fixing errors of the entity detection method it is combined with.

Table 6 presents the number of right an wrong annotations from the mGENRE system after the entities were discarded by the corresponding threshold as well as the system overall accuracy in each case. It is clear that with the implementation of the threshold, the EBD + mGENRE approach can match or even exceed the accuracy of Wikifier(87.5%).

With the discarded entities, the total number of annotations also declines. Table 6 shows expected number of annotations produced each threshold. We see that for t=-0.3 the EBD + mGENRE system has higher precision than Wikifier, for t=-0.5 it has a higher recall but for t=-0.4 it has the best trade-off in accuracy and recall with more annotations and higher accuracy than Wikifier.

## 6 Discussion

Our comparison between Wikifier and mGenre with respect to entity disambiguation shows that mGenre outperforms Wikifier on Mewsli-9. However, the linking accuracy of Wikifier is comparable to one reported for mGenre and the difference comes from the ER step. Based on analysis of the

| t | -0.2 | -0.3 | -0.4 | -0.5 | none |
|---|---|---|---|---|---|
| Accuracy | 0.92 | 0.89 | 0.87 | 0.86 | 0.75 |
| Annotations (right) | 133 | 141 | 144 | 149 | 151 |
| Annotations (wrong) | 11 | 16 | 20 | 24 | 48 |

Table 5: Accuracy of the end-to-end performance of the EBD + mGENRE system for different values of the mGENRE score. "Annotations (right)" and "Annotations (wrong)" present the number of correct and wrong annotations after applying the threshold in each case.

| t | -0.2 | -0.3 | -0.4 | -0.5 | none | Wikifier |
|---|---|---|---|---|---|---|
| total number of annotations | 404 | 455 | 493 | 527 | 618 | 482 |
| number of expected correct annotations | 371 | 404 | 428 | 453 | 463 | 419 |

Table 6: Accuracy of the end-to-end performance of the system for different values of the mGENRE score. "Annotations (right)" and "Annotations (wrong)" present the number of correct and wrong annotations after the by the threshold discarded annotations in each case.

overall performance of the three end-to-end systems, we conclude that SpaCy + mGENRE is the least reliable systems due to its very low accuracy and the fact that it detects many more mentions than the other two systems cannot overcome this issue. The other two systems both produce satisfactory results with each of them having different advantages and disadvantages. Wikifier has high accuracy for all components of the system but performs rather poorly on short texts and produces fewer annotations overall. EBD + mGENRE combined with a threshold achieves slightly higher accuracy than Wikifier while detecting more entities but the threshold selection is not part of the current training process. It also performs well on short texts while having some difficulties extracting entities from longer texts. EBD itself achieves underwhelming results when considering only exact matches, however including partial matches, the performance significantly improves. Fortunately, mGENRE is capable of "fixing" entity boundary detection errors and thus boosting the overall performance of the system. Improvements in the entity detection is the most promising approach for improving the overall solution.

## 7 Conclusion and Further Work

In this paper we attempted to explore and compare different end-to-end entity linking systems. Apart from testing existing systems, we also build our own solutions as a combination of the state-of-the-art entity disambiguation model mGENRE with suitable named entity recognition or entity boundary detection methods. Our results show that Wikifier is capable of entity disambiguation which is slightly worse that the one achieved by mGENRE. On the other hand its performance with respect to entity recognition is not satisfactory and requires significant improvement.

Another significant outcome of our work is that a combination of entity boundary detection method with mGENRE and threshold filtering achieves the best overall performance on our custom dataset. In terms of Entity Disambiguation, mGENRE demonstrates comparably high results to the ones reported, which is an indicator for its reliability. Based on the separate results for entity (boundary) recognition and entity linking, we conclude that the performance of mGENRE regarding correctly detected entities (boundaries) is quite satisfactory and can be applied in real world applications.

For the improvement of the recall and precision of the end-to-end solution, improvements in the entity extraction is recommended. A possible future research direction in this field could be using Large Language Models, LLMs (Zhao et al., 2023) for named entity recognition (as proposed in (Wang et al., 2023), (Ashok and Lipton, 2023)). Apart from that, very recently, a transformer-based, end-to-end, one-pass multilingual system BELA (Plekhanov et al., 2023) was released. A comparison of this system to the solutions explored in this work would also be valuable.

## Acknowledgments

This work has received funding by the European Union under the Horizon Europe vera.ai: VERification Assisted by Artificial Intelligence project, Grant Agreement number 101070093.

## References


Rodrigo Agerri, Josu Bermudez, and German Rigau. 2014. Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dhananjay Ashok and Zachary C. Lipton. 2023. Promptner: Prompting for named entity recognition.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. Refined: An efficient zero-shot-capable approach to end-to-end entity linking.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Rajarshi Bhowmik, Karl Stratos, and Gerard de Melo. 2021. Fast and effective biomedical entity linking using a dual encoder.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant wikipedia concepts.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval.

Henry Y. Chen, Ethan Zhou, and Jinho D. Choi. 2017a. Robust coreference resolution and entity linking on dialogues: Character identification on TV show transcripts. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 216–225, Vancouver, Canada. Association for Computational Linguistics.

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2017b. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Syst. Appl.*, 72:221–230.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. 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)*, pages 2306–2317, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Antonin Delpeuch. 2020. Opentapioca: Lightweight entity linking for wikidata.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. Semeval-2023 task 2: Fine-grained multilingual named entity recognition (multiconer 2).

Iker García-Ferrero, Jon Ander Campos, Oscar Sainz, Ander Salaberria, and Dan Roth. 2023. Ixa/cogcomp at semeval-2023 task 2: Context-enriched multilingual named entity recognition using knowledge bases.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

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

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of tac-kbp2015 tri-lingual entity discovery and linking. *Theory and Applications of Categories*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

B. Lorica and P. Nathan. 2021. 2021 nlp survey report. Technical report, Gradient Flow.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Multiconer: A large-scale multilingual dataset for complex named entity recognition.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. Cross-language entity linking. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 255–263, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512, Online. Association for Computational Linguistics.

Mikhail Plekhanov, Nora Kassner, Kashyap Popat, Louis Martin, Simone Merello, Borislav Kozlovskii, Frédéric A. Dreyer, and Nicola Cancedda. 2023. Multilingual end to end entity linking.

Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2018. Voxel: A benchmark dataset for multilingual entity linking. In *The Semantic Web – ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II*, page 170–186, Berlin, Heidelberg. Springer-Verlag.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, Copenhagen, Denmark. Association for Computational Linguistics.

Beth M. Sundheim. 1995. Named entity task definition, version 2.1.

Andrey Tagarev, Krasimira Bozhanova, Ivelina Nikolova-Koleva, and Ivan Ivanov. 2021. Tackling multilinguality and internationality in fake news. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1380–1386, Held Online. INCOMA Ltd.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. Damo-nlp at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, page 142–147, USA. Association for Computational Linguistics.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models.

Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2021. Retrieval, re-ranking and multi-task learning for knowledge-base question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 347–357, Online. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,

Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval.

Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip Yu. 2019. Multi-grained named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1430–1440, Florence, Italy. Association for Computational Linguistics.

Wenpeng Yin, Mo Yu, Bing Xiang, Bowen Zhou, and Hinrich Schütze. 2016. Simple question answering by attentive convolutional neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1746–1756, Osaka, Japan. The COLING 2016 Organizing Committee.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Jin Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah Mcguinness, James Hendler, and Heng Ji. 2015. Entity linking for biomedical literature. *BMC medical informatics and decision making*, 15:S4.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 473–480, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.