Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3732–3740
Marseille, 20-25 June 2022

Enhanced Entity Annotations for Multilingual Corpora

Michael Strobl, Amine Trabelsi, Osmar Zaiane
University of Alberta, Lakehead University, University of Alberta
mstrobl@ualberta.ca, atrabels@lakeheadu.ca, zaiane@ualberta.ca

Abstract

Modern approaches in Natural Language Processing (NLP) require, ideally, large amounts of labelled data for model training. However, new language resources, for example, for Named Entity Recognition (NER), Co-reference Resolution (CR), Entity Linking (EL) and Relation Extraction (RE), naming a few of the most popular tasks in NLP, have always been challenging to create since manual text annotations can be very time-consuming to acquire. While there may be an acceptable amount of labelled data available for some of these tasks in one language, there may be a lack of datasets in another. WEXEA is a tool to exhaustively annotate entities in the English Wikipedia. Guidelines for editors of Wikipedia articles result, on the one hand, in only a few annotations through hyperlinks, but on the other hand, make it easier to exhaustively annotate the rest of these articles with entities than starting from scratch. We propose the following main improvements to WEXEA: Creating multi-lingual corpora, improved entity annotations using a proven NER system, annotating dates and times. A short evaluation of the annotation quality of WEXEA is added.

Keywords: Wikipedia, Entity Annotations, Distant Supervision

1. Introduction

Understanding natural language and storing information in a structured way at human level has been an issue for decades, e.g. see the SHRDLU program (Winograd, 1972), which was able to conduct a dialogue in English with a human to move blocks by the computer in a virtual box. More recent approaches often involve the population of a Knowledge Graph (KG), as described in (Zhang et al., 2016), which explicitly (or implicitly for end-to-end approaches) involves the following sub-tasks:

1. Named Entity Recognition (NER): The detection of Named Entities (NE). The most common dataset is derived from news data (Sang and De Meulder, 2003) and aims to detect four entity types: PERSON, LOCATION, ORGANIZATION, MISC.
2. Co-reference Resolution (CR): CR aims to detect co-reference clusters of entities mentioned throughout a document, not necessarily NEs.
3. Entity Linking (EL): EL links NEs to an existing entity in a Knowledge Base (KB), e.g. Wikipedia[1].
4. Relation Extraction (RE): RE systems are able to detect relations, typically, between NEs and their co-references recognized by NER and CR systems.

All of these tasks ideally require large amounts of data on which to train models, which is typically limited, especially if languages other than English are of interest. Wikipedia, as-is, is already able to provide datasets in the following ways:

1. NER: If a class-article-mapping is available, e.g. from (Ghaddar and Langlais, 2017), existing annotations in Wikipedia can be tagged with their corresponding type and used for NER models to be trained on. However, Wikipedia is not exhaustively annotated, and training such models on partially annotated data is challenging and leads to worse results than if fully annotated datasets are used, e.g. see (Jie et al., 2019) or (Mayhew et al., 2019).
2. CR: Wikipedia sometimes contains annotations of the same entities within the same article, even though authors are discouraged from linking the same entity per article more than once. Still, a dataset with such co-references for some entities can be created and a model can be trained on this dataset, even though more exhaustive annotations in Wikipedia would be beneficial.
3. EL: Wikipedia already has entities linked as hyperlinks, e.g. (Guo and Barbosa, 2018) created a dataset where they leveraged hyperlinks as links between entities and their articles in Wikipedia.
4. RE: Using Distant Supervision, as applied by (Riedel et al., 2010), and a KB, such as DBpedia (Bizer et al., 2009), an RE dataset can be created through considering a specific relation between a specific pair of entities, as it appears in the KB and sentences expressing this relation if both entities are present.

If done carefully, exhaustive entity annotations for Wikipedia as done by Wikipedia EXhaustive Entity Annotations (WEXEA) (Strobl et al., 2020) can add a large amount of data for each of these tasks. In addition, datasets extracted by WEXEA can be considered

[1] https://www.wikipedia.org/

as a starting point for manual data annotation since it presumably needs minimal refinement by human annotators, compared to starting data annotations from scratch.
Although, WEXEA has a number of shortcomings that our approach improves upon. Therefore, our contributions are:

- A number of improvements to WEXEA: (1) Dates and times are now included through using the SUTime library(Chang and Manning, 2012). (2) Better NER through using the CoreNLP toolkit (Manning et al., 2014) instead of a rule-based NER based on regular expressions. Instead of entities of an unknown type (in case no Wikipedia article matching these entities was found), these entities are now typed. (3) Parser improvements: Resolving more Wikipedia templates frequently appearing in articles.
- So far, WEXEA was only applied to the English version of Wikipedia and due to some language-specific restrictions, e.g. regarding the NER used, it was not straightforward to apply the same approach to other languages. We relaxed these restrictions and applied WEXEA to the English, French, Spanish, and German Wikipedia.
- Generally WEXEA now can provide the following language resources for multiple languages (as long as an NER is available): Exhaustively annotated Wikipedia, RE datasets using Distant Supervision (need to be refined manually, ideally), NER datasets (article-type-mapping needs to be available), EL datasets, CR datasets (especially for the article entity, since these are often mentioned), parsed Wikipedia articles, dictionaries for hyperlinks, aliases, redirects, categories and many more.

The remainder of this article is structured as follows: Section 2 provides information about similar existing approaches and the original WEXEA. In Section 3 our method is described in detail with an evaluation in Section 4, and our conclusions are presented in Section 5.

2. Related Work

In this section we provide some information on similar systems to WEXEA as well as an overview of it.

2.1. Previous Work

WiNER (Ghaddar and Langlais, 2017) processes the English Wikipedia similarly to WEXEA, focusing on NER datasets. The corpus is publicly available[2] but only contains typed annotations; the original Wikipedia annotations are removed. This basically eliminates the possibility of extracting CR, EL and RE datasets.

The system in (Klang and Nugues, 2018) is a similar approach and focuses on EL and indexing and visualizing Wikipedia for English and Swedish. The annotation quality is unknown and except for the visualizations, which are accessible online[3], annotations can neither be downloaded nor reproduced.

2.2. WEXEA

Wikipedia is a valuable source of information, which can be transformed into training data, for example, for NER (see (Ghaddar and Langlais, 2017)) or EL (see (Guo and Barbosa, 2018)). These datasets take advantage of the quality, size as well as the hyperlink structure of Wikipedia. Editors can add hyperlinks for mentions of entities to their corresponding articles in text in order to add more information for the reader. Due to their unambiguity, these links can be used for dataset creation for NER (considering each article has a corresponding NER type) or EL or potentially for datasets for RE systems, when manually refined or Distant Supervision is used. However, the major drawback of a raw Wikipedia dump is that it is not exhaustively annotated.
Wikipedia policy is discouraging editors from adding more than one hyperlink per entity, i.e. the first mention of an entity is linked to its article, subsequent mentions are not, assuming the reader already knows which entity is mentioned. Furthermore, the entity an article is about is never linked within the same article and popular entities are not linked as well since, again, it is obvious which entity is being talked about. Also sometimes an entity may not have an article in Wikipedia, in which case no annotation is possible.
Consider the following example with existing hyperlinks in blue, illustrating the issue[4]:

> "Tony Hawk was born on May 12, 1968 in **San Diego, California** to Nancy and Frank Peter Rupert Hawk, and was raised in San Diego."

Tony Hawk himself is not linked since this sentence is part of his article. The entity "San Diego" is linked the first time, but not the second time and his parents are not linked since there is no corresponding article in Wikipedia.
WEXEA (Strobl et al., 2020) aims to solve this issue and produces an exhaustively annotated dataset based on the English Wikipedia. The main assumption is that adding annotations to Wikipedia articles is much easier due to its structure than automatically creating high-quality annotations for open text , e.g. editor guidelines indicate that non-linked entities typically refer to either the article entity, a previously mentioned (and linked) entity or a popular entity, which is easy to link to its

[2] http://rali.iro.umontreal.ca/rali/?q=en/wikipedia-main-concept

[3] http://vilde.cs.lth.se:9001/en-hedwig/

[4] From Tony Hawk's Wikipedia article: https://en.wikipedia.org/wiki/Tony_Hawk

	Wikipedia	WEXEA
Articles	2,952,439	2,952,439
Mentions total	64,654,332	265,498,363
- avg per sentence	0.38	1.56
- avg per article	21.90	89.93

Table 1: Annotation statistics for WEXEA compared to raw Wikipedia. WEXEA contains significantly more entity mentions, more mentions per sentence and per article.

article. A number of rules are used to achieve this as well as a neural EL system (Gupta et al., 2017), trained on Wikipedia data, is used in case of ambiguous entities (after rules were applied). Therefore, WEXEA provides a procedure (including a codebase) on how to process a raw Wikipedia dump to generate an exhaustively annotated dataset, which can be used for downstream tasks.

Table 1 shows annotation statistics from WEXEA, compared to raw Wikipedia. Overall, the increase in the number of found mentions is very large.

Due to many more annotations from WEXEA compared to Wikipedia, the authors of (Strobl et al., 2020) compared dataset creation using Distant Supervision from both corpora. For Distant Supervision (e.g. see (Riedel et al., 2010)) it is assumed that if a pair of entities is related in a KB, e.g. DBpedia (Bizer et al., 2009), sentences containing both entities may express these relations between them. In order to create such datasets it is crucial to consider corpora with already annotated entities, which can be linked to the KB. WEXEA contains many more annotations than Wikipedia, resulting in more sentences containing pairs of entities of interest.

However, WEXEA has two main drawbacks:

- Language dependency: WEXEA was developed for the English Wikipedia and, for example, the rule-based NER uses a set of words for frequent sentence starters in English, which cannot be easily provided in multiple languages.
- Rule-based NER: Due to Wikipedia's size, a rule-based NER was used to find non-annotated entities. However, state-of-the-art NER approaches work presumably much better.

We present improvements to WEXEA mitigating these drawbacks in Section 3.

3. Method

In this section we propose methods to overcome issues existing in WEXEA.

3.1. Named Entity Recognition

WEXEA tries to exhaustively annotate entities in Wikipedia and takes advantage of hyperlinks added by the editors of each article. While these hyperlinks can be considered as properly annotated entities, such annotations are far from being exhaustive. In addition, aliases of these entities are found using regular expressions and annotated accordingly. However, popular entities, as well as entities not part of Wikipedia, are not linked to any article, and WEXEA aims to find them using a rule-based NER.

This rule-based NER mainly looks for a sequence of words starting with capital letters (with a few exceptions for words such as "of" in "University of Alberta"). This decision was made since Wikipedia contains a large amount of text and a model-based NER would have taken too much time to process all of it. It becomes obvious that there are two major drawbacks of such an NER system:

- Many entities are missed, e.g. "Germany national football team", and for languages such as German, with all nouns capitalized, many false positives would be recognized.
- A great benefit of commonly used NER systems is the output of a type for each found entity. WEXEA is unable to provide such a type even though it may be useful for using the dataset to train new NER models.

We found that the CoreNLP toolkit (Manning et al., 2014) with appropriate settings can provide accurate annotations while still running through all of Wikipedia within a reasonable amount of time. This method leads to the following advantages:

- Typed annotations: LOC, PER, ORG, MISC as well as numerical and temporal entities. Specifically temporal entities, such as times and dates, can be important for certain relations, such as "birthDate" and are therefore included.
- The NER system contained in CoreNLP is proven to work (Finkel et al., 2005).
- The SUTime library (Chang and Manning, 2012), which is part of CoreNLP, is responsible for recognizing temporal entities.

Therefore, the CoreNLP toolkit provides a variety of improvements and, in addition, is proven to work.

3.2. Multi-Language

WEXEA currently only supports English, i.e. only the English Wikipedia would result in meaningful datasets. Originally, the rule-based NER relied on a list of frequent sentence starter words in English in order to avoid the annotation of such words at the beginning of a sentence since they are always capitalized. While this is an obvious language dependency, there are a few more subtle dependencies in WEXEA:

- As previously mentioned, the rule-based NER may not be able to reliable detect entities for languages other than English, e.g. all nouns are capitalized in German, but not all of them correspond to interesting entities worth annotating.
- CR in WEXEA is based on the detection of pronouns and types of entities as co-references, e.g. "the company" can refer to "General Electric". Co-references other than pronouns are extracted from Wikipedia article names (some contain the type in brackets, e.g. "Audrey (band)"[5]) or the Yago KB (Mahdisoltani et al., 2015). However, these co-references can only be used in English, otherwise they would have to be translated.
- Wikipedia markup contains a few language-dependent keywords, e.g. "category"[6], "template"[7], "file" or "image" (both are used mostly to reference images). WEXEA aims to extract articles without Wikipedia markup (except hyperlinks to other articles) and, therefore, needs to be aware of these keywords. And even though they are typically not visible to the reader, they are still different for each language-version of Wikipedia.

In order to overcome these shortcomings, we made a number of improvements supporting multiple languages, which are described below.
CoreNLP is the main entity detection tool now, which is available in the following languages: English, Arabic, Chinese, French, German and Spanish[8]. We tested WEXEA on English, French, German and Spanish, even though more languages can be introduced through training new models.
The SUTime library for detecting dates and times is available in English and Spanish. We extended the rule-set for German and French, which mainly involved translating weekdays and months as well as adjusting regular expressions for detecting dates. It can be adjusted for more languages of interest or left out.
For each of the currently used three languages, other than English, we translated the aforementioned language-dependent keywords used for Wikipedia markup, which involves around 20 keywords.
The EL system used for mentions of entities with multiple candidates was trained on English Wikipedia text (Gupta et al., 2017) and therefore cannot be used for other languages. In this case a greedy EL system is used, which links the candidate with the highest prior probability, i.e. the one that appears the most often with the detected mention in Wikipedia. For example, if the entity "New York" is found and the set of candidates consists of the articles for "New York City" and the "New York Yankees", the former would be chosen since the hyperlink count from "New York" to the article "New York City" is higher than for the "New York Yankees". In addition, as previously mentioned, the CR system used cannot directly be applied to other languages. Hence pronouns are translated from English to the currently used language, and English entity names are still used, other co-references from Yago are removed since they would have to be translated including the corresponding definite articles used. This reduces the set of entities with such co-references, although still some of them can be detected for languages other than English.

3.3. Parser Improvements

We improved a few issues the parser previously had:

- WEXEA did not parse the "convert"-template, which contains information about numbers, their units and how they can be converted, in order to display the right value and unit, depending on what is commonly used in the country of the reader, e.g. imperial vs. metric units. This template is parsed now and included in the resulting articles, if mentioned.
- The "language"-template was not parsed either since WEXEA generally removed all templates. However, this template often appears in the first sentence in case the entity is, for example, typically written in another alphabet, e.g. Greek or Cyrillic. Completely removing the template often resulted in a non-grammatical first sentence.

4. Evaluation

In this section we present statistics of the datasets we created with WEXEA for the English, German, French and Spanish versions of Wikipedia and a visualization of an article in English.

4.1. Dataset Creation

Wikipedia is large[9] and processing all articles can be time-consuming, depending on the language used. WEXEA has two main steps[10]:

1. Dictionary creation, article split (storing each article separately), resolving templates, removing non-content articles, e.g. redirects, lists, categories etc. ($\sim$ 5h).

[5] https://en.wikipedia.org/wiki/Audrey_(band)
[6] https://en.wikipedia.org/wiki/Help:Category
[7] https://en.wikipedia.org/wiki/Help:Template
[8] https://stanfordnlp.github.io/CoreNLP/human-languages.html

[9] 6,383,000+ articles in the English Wikipedia as of January 14, 2022: https://www.wikipedia.org/
[10] Runtimes are based on dataset creation on a AMD Ryzen 3700X with 64G main memory and the English Wikipedia (version with the most articles).

2. Removing non-named entity annotations and adding annotations through the alias and redirect dictionaries as well as running the CoreNLP toolkit and finding acronyms ($\sim$ 2d).

Especially step (2) takes a significant amount of time due to CoreNLP annotating each article. However, once all necessary dictionaries are created in (1), articles can be prioritized in step (2), e.g. only processing the most popular articles, to speed up the process. Wikipedia for languages other than English usually contains significantly less articles and, therefore, runtime will be faster.

4.2. Visualization

Figure 1 shows a paragraph from Queen Victoria's English Wikipedia article[11] with the original in Figure 1a and the corresponding one generated by WEXEA in Figure 1b. WEXEA annotated many more entities than the original paragraph contains. Multiple annotations of Queen Victoria can be found compared to no annotations in Wikipedia since this is her article, and editors are not supposed to add links to the same article.

4.3. WEXEA Statistics

Table 2 provides basic statistics of WEXEA when run on the English[12], German[13], French[14] and Spanish[15] Wikipedia versions. The number of relevant articles[16] and sentences as well as the number of original annotations in Wikipedia and all annotations in WEXEA ("Entities total") are shown in total, averaged per sentence and article.

WEXEA extracts a number of potentially useful dictionaries and articles of a specific type, with statistics presented in Table 3:

- Article: Number of articles (in order to put numbers into perspective).
- Categories (assigned): Wikipedia contains a category hierarchy[17] to group together entities of a similar subject. Each article can have multiple categories attached, hence more category assignments than articles can be found.
- Disambiguation pages: These pages often contain lists of articles, which can be referred to by the name of the disambiguation page[18].
- Lists: List articles contain a list of articles of the same type[19].
- Aliases: Each hyperlink in Wikipedia contains an anchor text (i.e. an alias) and the entity it refers to (not necessarily the same). Therefore this dictionary contains aliases and linked articles, important for EL in order to create a list of candidates per entity found in text.
- Links: Some EL systems, e.g. see (Guo and Barbosa, 2018), consider the links of a candidate to other entities in a document in order to link an entity to a candidate. This dictionary contains all links extracted through hyperlinks from an article to the linked entity.
- Redirects: This dictionary provides alternative names for many Wikipedia articles, e.g. "New York state" refers to the article with name "New York (state)". Whenever an article in Wikipedia is renamed, a redirect page is created (if it is ambiguous). Editors can also add redirects manually.
- Given names and surnames: Specific templates and categories mark articles about a "given name" or "surname". These articles are stored in these dictionaries and can be used to identify entities of type person.

Table 4 breaks up the annotation types of WEXEA:

- Annotations: Main annotations from Wikipedia.
- Article Entity: The article's entity, mentioned many times, but typically not annotated.
- Popular entities: Matching one of the 10,000 most popular articles in Wikipedia (after sorting by each article's number of hyperlinks), which are often not annotated due to the assumption the reader knows which article is mentioned.
- Single candidate: Annotations with a single candidate, typically referring to articles previously linked in the same article.
- Multi candidate: Multiple candidates are available, solved by the EL system.
- Co-references: Pronouns (found by the CoreNLP NER) and co-references for many entities based on their type.
- Acronyms: Acronyms and corresponding entities, which were not already annotated.
- NER: All other entities found and typed by the NER (including numerical entities only available in English).

[11] https://en.wikipedia.org/wiki/Queen_Victoria

[12] Wikipedia dump enwiki-20210220

[13] Wikipedia dump dewiki-20220101

[14] Wikipedia dump frwiki-20220101

[15] Wikipedia dump eswiki-20220101

[16] Not all article entities are considered as Named Entities, therefore the shown number of articles is lower than the one found on https://www.wikipedia.org/.

[17] https://en.wikipedia.org/wiki/Help:Category

[18] Disambiguation page of "New York": https://en.wikipedia.org/wiki/New_York

Victoria's father was Prince Edward, Duke of Kent and Strathearn, the fourth son of the reigning King of the United Kingdom, George III. Until 1817, Edward's niece, Princess Charlotte of Wales, was the only legitimate grandchild of George III. Her death in 1817 precipitated a succession crisis that brought pressure on the Duke of Kent and his unmarried brothers to marry and have children. In 1818 he married Princess Victoria of Saxe-Coburg-Saalfeld, a widowed German princess with two children—Carl (1804–1856) and Feodora (1807–1872)—by her first marriage to Emich Carl, 2nd Prince of Leiningen. Her brother Leopold was Princess Charlotte's widower. The Duke and Duchess of Kent's only child, Victoria, was born at 4:15 a.m. on 24 May 1819 at Kensington Palace in London.[1]

(a) Original paragraph from Wikipedia. Multiple entities are not annotated.

Victoria's father was Prince Edward, Duke of Kent and Strathearn, the fourth son of the reigning King of the United Kingdom, George III. Until 1817, Edward's niece, Princess Charlotte of Wales, was the only legitimate grandchild of George III. Her death in 1817 precipitated a succession crisis that brought pressure on the Duke of Kent and his unmarried brothers to marry and have children. In 1818 he married Princess Victoria of Saxe-Coburg-Saalfeld, a widowed German princess with two children—Carl (1804–1856) and Feodora (1807–1872)—by her first marriage to Emich Carl, 2nd Prince of Leiningen. Her brother Leopold was Princess Charlotte's widower. The Duke and Duchess of Kent's only child, Victoria, was born at 4:15 a.m. on 24 May 1819 at Kensington Palace in London.

(b) Corresponding paragraph from the file generated by WEXEA with exhaustive entity annotations. In order to simplify, only the anchor text for hyperlinks is shown, the linked entity is left out. Hyperlinks (blue) refer to an entity in Wikipedia, for all other annotations (green) no entity in Wikipedia can be found (mostly dates, times, numbers as well as other Named Entities). The original annotation of "succession crisis" (third sentence) was removed since it does not correspond to a Named Entity.

Figure 1: Same paragraph from Queen Victoria's English Wikipedia article as well as her corresponding article generated by WEXEA. Hyperlinks/Annotations in blue, linked articles are omitted.

Language	English		German	
Type	Wikipedia	WEXEA	Wikipedia	WEXEA
Articles	2,676,086		1,929,698	
Sentences	148,866,723		83,426,382	
Entities total	62,026,078	320,142,453	45,383,570	149,776,874
- avg per sentence	0.42	2.15	0.54	1.80
- avg per article	23.18	119,63	23.52	77.62
Language	French		Spanish	
Type	Wikipedia	WEXEA	Wikipedia	WEXEA
Articles	1,568,460		1,137,844	
Sentences	64,214,837		43,794,620	
Entities total	26,113,542	97,482,667	17,059,545	74,824,888
- avg per sentence	0.41	1.52	0.39	1.71
- avg per article	16.65	62.15	14.99	65.76

Table 2: Annotation statistics for WEXEA compared to Wikipedia, for English, German, French and Spanish.

Since all annotations are marked accordingly, unwanted annotations, e.g. numerical entity types, can be filtered out.

4.4. Entity Annotations

Table 5 shows a small evaluation of a subset of the main entity annotations from WEXEA (apart from NER annotations) for 20 randomly selected Wikipedia articles. Overall, the annotation quality for article entities, popular entities and single candidate entities is excellent. Multi-candidate entities do not appear often, a large number of articles needs to be evaluated to achieve a higher confidence level. The CR system of WEXEA is sometimes mislead, carrying an incorrect entity through a number of sentences.

[19]List of sovereign states: `https://en.wikipedia.org/wiki/List_of_sovereign_states`

4.5. Datasets for NLP tasks

WEXEA corpora can be used for multiple NLP tasks, such as NER, EL, CR and RE, and here we describe how datasets can be extracted or how these corpora can be used as-is. Since WEXEA does not provide gold-standard annotations, these datasets can be used as a starting point to pre-train models or manually annotate data, especially if high-quality training data is very limited or unavailable (specifically for languages other than English).

4.5.1. Named Entity Recognition

Models for NER are typically trained for token classification, e.g. see (Lample et al., 2016), i.e. every token in the training set needs to be properly annotated. Wikipedia cannot be used due to many missing annotations and partially annotated datasets for NER can significantly reduce model performance, e.g. see (Jie et al., 2019) or (Mayhew et al., 2019). WEXEA can fill

	English	German	French	Spanish
Articles	2,676,086	1,929,698	1,568,460	1,137,844
Categories (assigned)	21,067,025	11,702,075	9,509,203	5,641,208
Disambiguation pages	303,508	311,151	106,963	58,845
Lists	111,103	83.664	27,229	55,992
Aliases	4,620,268	3,019,351	2,315,007	1,818,029
Links	104,820,886	62,790,995	47,263,448	33,434,730
Redirects	4,811,019	1,444,249	1,089,731	1,434,897
Given names	10,482	7,103	1,327	976
Surnames	35,600	6,318	2,379	142

Table 3: Further annotation statistics for WEXEA and all languages.

	English	German	French	Spanish
Annotations	62,026,078	45,383,570	26,113,542	17,059,545
Article Entity	19,873,610	11,215,066	6,507,156	4,695,562
Popular entities	14,507,645	7,667,027	4,110,588	3,502,393
Single candidate	42,833,403	27,162,472	10,228,398	8,786,032
Multi candidate	2,861,934	894,593	316,171	223,584
Co-references	36,300,537	2,604	1,092	1,140
Acronyms	745,470	207,876	84,113	117,951
NER	140,993,776	57,243,665	43,808,292	32,023,229
Total	320,142,453	149,776,874	91,169,353	66,409,439

Table 4: Breakup of mention types found by WEXEA for all languages.

Annotation type	Accuracy	Num.
Article entities	0.97	98
Popular entities	0.96	56
Candidate entities (single)	0.91	44
Candidate entities (multi)	0.67	3
Co-reference entities	0.69	67

Table 5: Accuracy and number of annotations for four different annotation types and 20 randomly selected articles from the English WEXEA corpus.

these annotation gaps and therefore the resulting corpora can be used for NER, if a article-type mapping can be provided. DBpedia contains type-relationships for each entity[20], which can be used here.

4.5.2. Entity Linking

Similar to the dataset created by (Guo and Barbosa, 2018), large datasets for EL can be extracted from WEXEA corpora without modification. Annotations linking entities to an article in Wikipedia can be used to train such systems. Furthermore, EL systems rely on alias dictionaries in order to create a set of candidates for each entity found in text. These alias dictionaries should be updated frequently since the set of articles/entities of interest increases constantly and WEXEA can be used to provide these dictionaries.

[20]For example, see Barack Obama's DBpedia article: `https://dbpedia.org/page/Barack_Obama`

4.5.3. Co-reference Resolution

CR systems aim to detect clusters of co-references for each mentioned entity, e.g. see (Lee et al., 2018). Therefore, datasets for CR need to contain spans of text corresponding to entity annotations and a unique identifier for each entity mentioned to link co-references to a unique entity. WEXEA does contain exhaustive entity annotations and unique identifiers for many annotations (except NER tagged annotations without articles attached). For example, the article entity is in average mentioned over seven times per article in the English WEXEA corpora (see Table 3 and 4), which corresponds to one entity cluster, which can be used for CR model training.

4.5.4. Relation Extraction

Many relations and corresponding entity pairs can be extracted from DBpedia. These pairs of entities can be found in Wikipedia and WEXEA corpora, and datasets for RE based on Distant Supervision can then be created. Table 6 shows the amount of data (sentences), which can be extracted from these corpora, for relations for which at least 100 sentences each can be found. Using WEXEA, many more relations, sentences and unique pairs of entities per relation can be found, when compared to Wikipedia. In case a large number of sentences can be extracted for a low number of unique pairs, an RE model may memorize relations for certain pairs rather than recognizing relations in text. WEXEA is able to extract a more significant number of unique pairs than Wikipedia contains.

Figure 2 shows the distribution of sentences for the

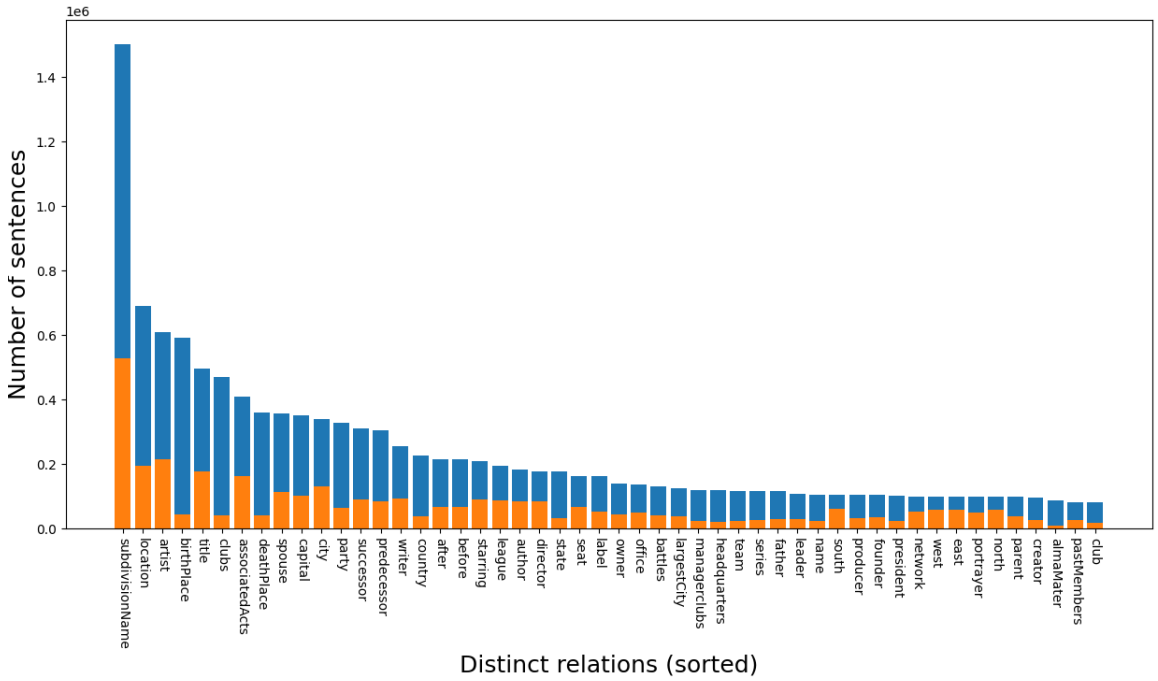

Figure 2: Number of sentences per relation (top 50) for the English Wikipedia (orange) and WEXEA (blue; in addition to Wikipedia).

Language	English		German	
Type	Wikipedia	WEXEA	Wikipedia	WEXEA
Relations	1,458	2,322	616	928
Sentences (total)	6,323,758	20,970,686	1,702,887	5,399,588
Unique pairs (avg)	1,017	1,540	709	1,935
Sentences (avg)	4,337	9,031	2,764	5,819
Language	French		Spanish	
Type	Wikipedia	WEXEA	Wikipedia	WEXEA
Relations	746	1,048	640	893
Sentences (total)	3,392,807	6,577,523	1,862,071	4,689,909
Unique pairs (avg)	971	1,289	651	1112
Sentences (avg)	4,548	6,276	2,909	5,252

Table 6: Statistics for datasets based on Distant Supervision created using Wikipedia or WEXEA corpora and DBpedia: Number of relations for which at least 100 sentences can be extracted, extracted sentences total, unique pairs of entities averaged per relation and average number of sentences per relation.

top 50 relations (based on the number of sentences extracted from WEXEA) for the English Wikipedia (orange) and WEXEA (blue). Again, many more sentences can be extracted from WEXEA for each of these relations.

5. Conclusion

We updated WEXEA to provide enhanced entity annotations and multilingual corpora. Improved Named Entity Recognition using the CoreNLP toolkit provides typed annotations for multiple languages. The SUTime library can detect temporal entities, which can be useful, for example, for Relation Extraction datasets. Datasets for multiple NLP tasks can be created from WEXEA corpora, which are publicly available[21]. A short evaluation of the annotation quality of a subset of WEXEA's annotation shows promising results.

We applied WEXEA to multiple Wikipedia language versions (English, German, French and Spanish). While the CoreNLP NER can be trained on datasets for other languages, the WEXEA parser still relies on several language-dependent keywords, making it challenging to adapt to new languages. For future work, we would like to remove this barrier to be able to provide enhanced corpora for all available Wikipedia versions.

[21] Codebase (including SUTime rules for German and French) and datasets for English, German, French and Spanish: `https://github.com/mjstrobl/WEXEA`

6. Bibliographical References

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, 7(3):154–165.

Chang, A. X. and Manning, C. D. (2012). Sutime: A library for recognizing and normalizing time expressions. In *Lrec*, volume 3735, page 3740.

Finkel, J. R., Grenager, T., and Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370.

Ghaddar, A. and Langlais, P. (2017). Winer: A wikipedia annotated corpus for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 413–422.

Guo, Z. and Barbosa, D. (2018). Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.

Gupta, N., Singh, S., and Roth, D. (2017). Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2681–2690.

Jie, Z., Xie, P., Lu, W., Ding, R., and Li, L. (2019). Better modeling of incomplete annotations for named entity recognition. 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 729–734.

Klang, M. and Nugues, P. (2018). Linking, searching, and visualizing entities in wikipedia. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

Mahdisoltani, F., Biega, J., and Suchanek, F. M. (2015). Yago3: A knowledge base from multilingual wikipedias. In *Conference on Innovative Data Systems Research*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mayhew, S., Chaturvedi, S., Tsai, C.-T., and Roth, D. (2019). Named entity recognition with partially annotated training data. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 645–655.

Riedel, S., Yao, L., and McCallum, A. (2010). Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Strobl, M., Trabelsi, A., and Zaiane, O. R. (2020). Wexea: Wikipedia exhaustive entity annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1951–1958.

Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Zhang, Y., Chaganty, A. T., Paranjape, A., Chen, D., Bolton, J., Qi, P., and Manning, C. D. (2016). Stanford at tac kbp 2016: Sealing pipeline leaks and understanding chinese. In *TAC*.