Enhanced Entity Annotations for Multilingual Corpora

Michael Strobl[1] Amine Trabelsi[2] Osmar Zaïane[1]

University of Alberta, Edmonton, Canada

Lakehead University, Thunder Bay, Canada

Overview

Improvements to WEXEA, a tool to exhaustively annotate entities in the English Wikipedia:

- Multi-language support (English, German, French and Spanish).
- Improved entity annotations using a proven NER system, annotating dates and times.
- Evaluation of the annotation quality of WEXEA.

Background: WEXEA

Wikipedia is a valuable source of information, which can be transformed into training data, for example, for NER or EL. However, due to Wikipedia policies for editors, e.g. no more than one hyperlink (annotation) per article mentioned, many potential annotations are missing. WEXEA aims to exhaustively annotate Wikipedia articles.

Consider the following example with existing hyperlinks in blue, illustrating the issue (see Tony Hawk's Wikipedia article):

> *"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 aims to solve this issue and produces an exhaustively annotated dataset based on the English Wikipedia.

Visualization

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]

Figure 1:Original paragraph from Queen Victoria's Wikipedia article. 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.

Figure 2:Corresponding paragraph from the file generated by WEXEA with exhaustive entity annotations.

Annotation Statistics

WEXEA can annotate many more entities than existing hyperlinks in Wikipedia. Resulting corpora can be used for multiple NLP tasks, such as NER, EL, CR and RE.

Lang.	English		German	
Type	Wikipedia	WEXEA	Wikipedia	WEXEA
Articles	2,676,086		1,929,698	
Sent.	148,866,723		83,426,382	
Entities	62,026,078	320,142,453	45,383,570	149,776,874
Lang.	French		Spanish	
Type	Wikipedia	WEXEA	Wikipedia	WEXEA
Articles	1,568,460		1,137,844	
Sent.	64,214,837		43,794,620	
Entities	26,113,542	97,482,667	17,059,545	74,824,888

Table 1:Annotation statistics for WEXEA compared to Wikipedia, for English, German, French and Spanish.

Relation Extraction

Using Distant Supervision, large datasets for RE can be extracted from WEXEA corpora with many more sentences compared to the original Wikipedia.

Lang.	English		German	
Type	Wikipedia	WEXEA	Wikipedia	WEXEA
Relations	1,458	2,322	616	928
Sent.	6,323,758	20,970,686	1,702,887	5,399,588
Lang.	French		Spanish	
Type	Wikipedia	WEXEA	Wikipedia	WEXEA
Relations	746	1,048	640	893
Sent.	3,392,807	6,577,523	1,862,071	4,689,909

Table 2: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 and extracted sentences total.

Improvements

- Replaced language-dependent NER and adapted Wikipedia markup parser to support three more languages.
- Now, CoreNLP NER for all entities, which cannot be found through matching aliases of linked articles. Available (among others) for English, German, French and Spanish.
- SUTime library for recognizing temporal entities (new rule sets for German and French, existing sets for English and Spanish).
- More Wikipedia templates parsed.

Entity Annotations Evaluation

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

Conclusion & Future Work

- 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, including temporal entities.
- WEXEA relies on a language-dependent Wikipedia keyword parser, making it challenging to adapt to new languages. For future work, this barrier should be removed.