In Media Res: A Corpus for Evaluating Named Entity Linking with Creative Works

Adrian M.P. Braşoveanu
MODUL Technology GmbH
Vienna, Austria
adrian.brasoveanu@modul.ac.at

Albert Weichselbraun
Fachhochschule Graubünden
Chur, Switzerland
albert.weichselbraun@fhgr.ch

Lyndon J.B. Nixon
MODUL Technology GmbH
Vienna, Austria
lyndon.nixon@modul.ac.at

Abstract

Annotation styles express guidelines that direct human annotators by explicitly stating the rules to follow when creating gold standard annotations of text corpora. These guidelines not only shape the gold standards they help create, but also influence the training and evaluation of Named Entity Linking (NEL) tools, since different annotation styles correspond to divergent views on the entities present in a document. Such divergence is particularly relevant for texts from the media domain containing references to creative works. This paper presents a corpus of 1000 annotated documents from sources such as Wikipedia, TVTropes and WikiNews that are organized in ten partitions. Each document contains multiple gold standard annotations representing various annotation styles. The corpus is used to evaluate a series of Named Entity Linking tools in order to understand the impact of the differences in annotation styles on the reported accuracy when processing highly ambiguous entities such as names of creative works. Relaxed annotation guidelines that include overlap styles, for instance, lead to better results across all tools.

1 Introduction

Identifying all entities from texts and linking them to modern Knowledge Graphs (KGs) like DBpedia and Wikidata is the core task of Named Entity Linking (NEL) (Rosales-Méndez et al., 2018b). A particular challenge in NEL is disambiguating the entity references in the surface form (original text). For example, mentions such as *NYT* or *NY Times* or even *Times* could all be surface forms referencing the entity *The New York Times* and they can then be linked to the associated entries from well-known public KGs like DBPedia and Wikidata (i.e., dbr:The_New_York_Times or wd:Q9684 [1]). In addition to entities, the relations between them, as defined in the texts themselves or extracted from a KG, can also help enlighten the narratives. All of the NEL related tasks are supported by large communities which have a vested interest in advancing their Knowledge Extraction (KE) capabilities. The various good results for general NEL benchmarks (e.g., F1 scores of 0.80 (Rosales-Méndez et al., 2018a)) are encouraging and suggest that the tasks tested for in these benchmarks may be solved soon. However, the main focus of most reported evaluations to date has been on the correct annotation of several specific types of entities like Persons (PER), Locations (LOC) or Organizations (ORG). Sometimes fine-grained typing was also allowed for locations, therefore allowing us to use types like natural locations (LOC) to depict naturally occurring locations like mountains or rivers, facility (FAC) to describe buildings or infrastructure like bridges or airports and Geo-Political Entities (GPE) to describe cities, region or countries (Ji et al., 2017). Only in the last half-decade the focus has slowly shifted towards expanding the typing to encompass a wider array of types, from dates and products to games, sports, books or movies. A large subset of the fine-grained types of entities found naturally in texts, especially in the media domain, are focused on what can generically be called *Works* or sometimes *Creative Works* and which would encompass all creative works (manifestations of creative effort by a creator or group of creators), from books or songs, to movies, television episodes and video games, for example. Collective works with unknown authors (e.g., religious works, folklore), festivals, concerts or sports events can also be considered creative works, even though they may be labelled differently (e.g., a concert might be labelled as an *Event*, whereas its audio or video recording could be labelled as a *Creative Work*).

[1] *dbr* represents the namespace abbreviation of DBpedia resources, *wd* for Wikidata resources

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 355–364

https://doi.org/10.18653/v1/P17

The current generation of automated annotators provide a rather mixed set of results for media-related annotations due to inconsistent annotation guidelines related to this domain in the past. For example, a TV show like *Star Trek: Deep Space Nine* will either get no result or two different results, one marking the franchise (*Star Trek*) and the second the titular space station (*Deep Space Nine*), but will probably not be annotated with the full correct entity (*Star Trek: Deep Space Nine*) by many automated annotator tools. Such an annotation could alternatively be considered correct in our view, if instead of the space station, the second annotation would point to the TV show itself. Of course, if we examine the characters from the respective TV shows, we can run into similar problems, as some characters might be described by multiple resources, each of them highlighting a particular performance (e.g., *James Bond* is described both as literary character - dbr:James_Bond_(literary_character) - and Person dbr:James_Bond). In order to counteract such issues, we have developed a corpus focused around annotating such media-related entities. We share this corpus, together with the corresponding annotation guidelines, as well as a set of alternative annotations, especially for nested entities [2]. Since we plan to continuously update this corpus with new partitions every few months, one of its core concepts is the fact that for each ten partitions created there will be a summary partition, a partition that will combine annotations similar to those from the previous nine partitions. If users aim at performing a quick evaluation, they will be able to use this summary partition which we call 'core set' in order to have a first set of results and decide upon their preferred fine-tuning procedure.

The expression *media annotations* is typically used to refer to video or audio annotations, therefore in the context of this paper we have selected the expression *media entity annotations* to signify the annotation of texts created and/or published by various media conglomerates which contain entities of type Creative Work.

We consider our work as an important step in the road towards building culturally-aware AIs since a significant part of human culture is represented by the creative works humanity has produced, AIs are needed that can correctly annotate and interpret information about the referenced creative works.

The rest of the paper is organized as follows: Section 2 showcases somewhat similar work and articles that discuss some of the ideas that we followed during our design process; Section 3 describes the design of the corpora; Section 4 evaluates several tools on some of the partitions included in our corpora; whereas the final section provides a reflection upon the insights gained from designing a gold standard that focuses on creative works.

2 Related Work

Since this dataset uses several concepts that are not widely used in NEL evaluation datasets (e.g., lenses, core sets, multiple annotation styles), this section also includes explanations for those concepts.

After examining a rich set of media annotation papers we have discovered that very few papers actually discuss how to correctly annotate media assets with creative work entities. Early results were focused on the correct semantic annotation of text extracted from the multimedia content (e.g., transcription or subtitles), as presented in a set of papers that originated from the LinkedTV project like (Li et al., 2013; Nixon and Troncy, 2014), but using the classical entity types well supported by NEL tools (Person, Organization, Location). More recent media annotation work discussed cross-modal annotations (Zhang et al., 2017) and story identification (Nixon et al., 2019). None of these projects or papers focused on the correct identification of creative works or the various naming variants involved in creative works referencing. An episode name like *The Trouble with Tribbles* can be an episode of *Star Trek*, but also an instance of a chapter, comic book or book in the same franchise. Even though the existing research would help us annotate the respective episode's video, it will not really help us contextualize it, if mentioned in a press release, for example.

Due to the shortcomings of current approaches, we have also examined several other avenues, including corpora that contain creative works or historical entities that might otherwise be difficult to identify by the general public and were created for different purposes but in somewhat similar situations (e.g., no corpora available, specific domain, etc). The digital humanities domain, and in particular historical documents and literary criticism, are perhaps the closest scenarios to our use case

[2]The corpus is available online at: https://github.com/modultechnology/in_media_res.

that come to mind. Several works in this domain like the historical documents from the Impresso collection (Ehrmann et al., 2020), the multilingual news corpora MeanTime (Minard et al., 2016) and Dekker's work on extracting small snippets of literary criticism from social media (Dekker et al., 2018) have served as a starting point in our journey, helping us to define our annotation guidelines.

Some early ideas about naming variations (in particular (Ehrmann et al., 2017), (Rosales-Méndez et al., 2019) and (Weichselbraun et al., 2019b)), and nested annotations ((Ju et al., 2018) and (Ji et al., 2017)) have also shaped our understanding of the difficulties of correctly capturing names referencing entities, regardless of the domain, and have led us to multiple works on annotation styles and lenses. The idea of data lenses comes from multiple places, but it was generally inspired by photography where different lenses are used in order to get different views on an object. In NEL, and in NLP in general, the idea is to enable different views onto the same dataset. During the last few years, lenses have increasingly been used to transform data between different views (Rajkumar et al., 2013). In the NLP domain, lenses have traditionally been implemented as annotation sets that reflect the view of a human annotator upon the specific data, being an important feature of annotation packages like GATE (Maynard, 2009), as well as for implementation of large parallel corpora typically used for multilingual settings (e.g., (Iranzo-Sánchez et al., 2019). In the Semantic Web (SW), ontologies have traditionally been used to act as lenses over data eventually leading to an entire field of study: Ontology-Based Data Access (Calvanese et al., 2015). More recently, lenses have been used to create multiple views over chemistry data (Batchelor et al., 2014), help manage large data sets (Lenzerini, 2018), or understand big data and AI workloads (Gao et al., 2018).

With respect to entity linking, an early example of lenses was the concept of approximate matching implemented in tools like Neleval (Hachey et al., 2014) and Gerbil (Röder et al., 2018). The computation of partial matches was done based on the number of overlapping characters or entity types with respect to the measures that can be optimized (e.g., precision or recall). Many optimization strategies will fail to correctly identify either connections between entities (e.g., parent) or nested entities. What is missing and why, can generally be discovered during the error analysis phase of the evaluations, as shown in (Braşoveanu et al., 2018b) and (Stanislawek et al., 2019). Understanding error classification schemes like the ones described in these publications can be the key towards moving the field forward.

More recently, lenses have been used to understand entity annotations (Braşoveanu et al., 2018a) and to study the effect of relaxed annotation schemas that include more types than usual (Rosales-Méndez et al., 2018a). A theoretical treatment of automatic procedures for building lenses is presented in (Weichselbraun et al., 2019a). The paper showcases how to automatically build lenses by following simple rules (e.g., expand the surface form to the maximum length - the longest mention - or reduce the surface form to the minimal possible length - the shortest mention). Using the longest mention leads to an annotation schema with less entities, whereas using the shortest mention will lead to more entities being annotated. Intermediary schemas are also possible, if we take into account the possibilities of combining the previous schemas or if we add additional rules. We test some of the lenses discussed in this paper, as it can be seen in Section 3.

Our In Media Res corpora contains partitions from multiple domains. Due to this aspect, it was decided to introduce core sets partitions at every nine partitions. By adding these core sets, evaluators can later build a summary dataset for the whole corpus, regardless of the number of partitions contained in it. The concept of core sets was borrowed from computational geometry and robotics. A core set is a small engineered subset of a very large dataset that retains its properties as accurately as possible (e.g., a set of points that approximate the shape of a figure). Its size typically depends on the desired accuracy, not necessarily on the size of the original data set. Core sets have been successfully applied to a variety of problems, from dimensionality reduction of massive data sets (Feldman et al., 2016), to vector summarization (Feldman et al., 2017) or compression of neural networks (Baykal et al., 2018).

3 The In Media Res Corpus

Since most corpora used in ground truth annotation are focused on news media or tweets that describe current events, there is no wide agreement on annotation styles for entities of type Creative Work.

We, therefore, started our corpus design by testing various annotation styles and creating an annotation guideline for Creative Work entities. After we identified a solution, we created multiple partitions from various domains.

The name of the corpus comes from the Latin expression *in media res* that refers to narratives (e.g., books or movies) that start in the middle of the story. Similarly, the created corpus has been designed with the aim of understanding different annotation styles, and rather than starting from zero builds upon prior research in this area. Also, the fact that most of the annotated documents are related to media (e.g., franchises, books, TV shows) has been one of the reasons why we have selected this name.

3.1 Annotation Styles

While classic corpora are quite good for identifying people, organizations or locations, there are less adequate corpora to help with evaluating works (e.g., books, TV shows, music, etc) or events. This has been the main reason why we have decided to create a corpus focused mostly on creative works to help us fine-tune media domain document annotations.

As opposed to the classic entity types in NEL evaluations (e.g., Person, Organization, Location), the annotation of documents from the media domain (TV, radio, film etc.) raises specific challenges. In addition to the core entity types like *Person*, *Organization* and *Location*, media domain document annotation also needs to support a fourth large class of entities: *Creative Work* or *Work* [3]. This class encompasses a large selection of entities that might be classified as creative works, from books and songs, through to games, movies, TV Shows and entire media franchises. It is important to note that, if we leave the temporal attributes aside (e.g., new positions for a person, key people for a company, new episodes for a TV show), there is a lot of variation when it comes to the main attributes of this entity type as opposed to the three core types (Person, Organization, Location). A TV show might have some executive producers, a production company, some actors starring in it, as well as a set of episodes, each with their own list of directors, writers or stars; a book will have some author(s), publisher, and awards or links to a book series; and a song might have an interpret, author, music producer, and so on. As it can be seen, it is difficult to find common attributes between the various sub-classes, except for the fact that they are all types of creative works that were published in some format or medium in a certain period of time. It can even be argued that all these creative works should be modeled as their own entity types, but in order to perform such a fine-grained extraction, it is important to first identify the large class to which the entities belong. Adding works can also lead to a high number of false positives, as often fictional characters (e.g., *James Bond, Harry Potter*) might share names with real people, as well as with their own media franchises which can encompass different sets of series (e.g., books, tv shows, movies, comics, etc); fictional characters might be based on real people (e.g., see the recent trend of music biopics based on *N.W.A., Queen* or *Elton John* or TV shows like *Narcos* who often fictionalize real characters by changing their names or changing the events in which they participate) or the name of a work is later used for a different work (e.g., again the example of music biopics is relevant). Table 1 showcases some of the issues encountered by annotators such as AIDA (Hoffart et al., 2011), DBpedia Spotlight (Daiber et al., 2013) and Recognyze (Weichselbraun et al., 2019b) when performing named entity linking on works. As illustrated in the table, each annotator tends to return a different set of results based on its settings. Unfortunately, the best settings for annotators are not published, therefore, even when NEL annotators are integrated into benchmarking systems like Gerbil (Röder et al., 2018) or (Odoni et al., 2018) it is still difficult to understand if the obtained results really represent the best possible outcome. When examining these differences between tools, we decided to use lenses as a method for further investigating overlaps and partial matches.

For the current evaluation, we have considered the following annotation styles based on (Weichselbraun et al., 2019a) (here illustrated on the annotation of the text snippet *Star Trek: Picard* and its associated DBpedia resource $dbr : Star\_Trek : \_Picard$):

1. The annotation style $\emptyset MIN$ disregards overlapping entities and extracts the minimum number of entities: $m^{dbr:Star\_Trek:\_Picard}_{\text{[Star Trek: Picard]}}$, i.e. links the snippet to the *Star Trek: Picard* DB-

[3] represented by http://dbpedia.org/ontology/Work (abbreviated as dbo:Work) in DBpedia or https://schema.org/CreativeWork in the schema.org vocabulary

Example	AIDA	Spotlight	Recognyze
Sir Patrick Stewart OBE	1: Patrick Stewart	1: Patrick 2: Stewart 3: OBE	1: Patrick Stewart 2: OBE
Star Trek: Deep Space Nine	1: Star Trek	1: Star Trek 2: Deep Space Nine	1: Star Trek 2: Star Trek: Deep Space Nine
The British Broadcast Corporation (BBC)	1: British Broadcast Corporation 2: BBC	1: British 2: BBC	1: British Broadcast Corporation 2: BBC
Seinfeld	1: Seinfeld	1: Seinfeld	1: Seinfeld

Table 1: Understanding differences between annotator results. Numbers were added in order to clarify which entities were retrieved.

pedia entity.

2. The annotation style *∅MAX* also ignores overlapping entities but extracts the maximum number of entities from a given text snippet: $m^{dbr:Star\_Trek}_{\text{[Star Trek]}}, m^{dbr:Star\_Trek:\_Picard}_{\text{[Picard]}}$.

3. The annotation style *OMAX* allows for overlaps and, again, will aim to extract the maximum number of entities whenever possible: $m^{Star\_Trek:\_Picard}_{\text{[Star Trek: Picard]}}, m^{dbr:Star\_Trek}_{\text{[Star Trek]}}$.

It has to be noted that two of the styles (*∅MAX* and *OMAX*) can also extract the person rather than the TV show *Picard* as a separate entity, but since this result would have a different type (*Person* instead of *Work*, as it points to the Jean-Luc Picard character from the same franchise) and different link (e.g., $dbr : Jean - Luc\_Picard$ instead of $dbr : Star\_Trek : \_Picard$), it would be an incorrect result that is automatically removed in our implementation. The presented rules only consider borderline cases, even though combinations of them can also be used within a corpus. A corpus which would not apply the OMAX rule, for example, might lose the extended reference to *Sir Patrick Stewart OBE* and only return *Patrick Stewart* or end up removing the references to the actor's titles (e.g., *Sir*, *OBE*). We consider OMAX annotation rule to be the best, as it essentially merges the other annotation styles. The advantage of using these rules comes from the fact that they can easily be automated. By using them we have generated alternative annotations (also known as lenses) for all of our dataset partitions. One such example is provided for the core set of our dataset in the evaluation from Section 4. The annotations in Table 2 illustrate the gold standard results for the different lenses described in this section. The ∅MIN results are somewhat closer to the expected full annotation (e.g., the one presented in the example column) and only introduce minor variations. One of our assumptions was that lenses such as OMAX should even the playing field by reducing the penalty of overlaps in terms of false positives, while also offering some clues on what kind of surface forms are picked up more frequently by the various annotators.

3.2 Partitions and Statistics

The corpora currently has 10 sections each with 100 documents. We plan to add more partitions in time to cover different domains and therefore test different algorithms for domain adaptation in Named Entity Linking.

At every 900 documents, we include a numbered partition called *general* (e.g., general-1, general-2, etc) which will contain some documents from each of the domains covered in these 900 documents, therefore representing a summary or a core set of the previous set of partitions. This is done in order to create a large core set of the entire corpora. The General core sets can be used both as smaller standalone corpora, as well as small test beds in order to decide if certain partitions are useful.

We have started by collecting several sentences from the Wikipedia abstracts of 100 articles about creative works (as classified by their respective DBPedia resource). The initial set of entities contained books, TV shows, media companies, YouTube influencers and media franchises. Several entity types were annotated, include Person (PER), Orga-

Example	ØMIN	ØMAX	OMAX
Sir Patrick Stewart OBE	1: Sir Patrick Stewart OBE	1: Sir 2: Patrick Stewart 3: OBE	1: Sir Patrick Stewart OBE 2: Sir 3: OBE
MLB Advanced Media (MLBAM)	1: MLB Advanced Media (MLBAM)	1:MLB Advanced Media 2: MLBAM	1: MLB Advanced Media (MLBAM) 2: MLBAM
Burbank, California	1: Burbank, California	1: Burbank 2: California	1: Burbank, California 2: California
Seinfeld	1: Seinfeld	1: Seinfeld	1: Seinfeld

Table 2: Understanding differences between lenses. Numbers were added in order to clarify which entities were retrieved.

nization (ORG), Location (LOC), Work (WORK), Event (EVENT) or Other (OTHER). The corpus was annotated by two annotators following our annotation guideline. The human annotators were asked to use the ØMIN lens for creating the initial annotations, therefore disregarding overlapping entities and selecting the minimum possible number of matches. A judge was available for questions during the whole annotation process and helped solve disagreements after the annotation process has been completed. The rest of the lenses (e.g., ØMAX, OMAX) have later been automatically generated using a Python script. The judge has then verified the resulting annotations in order to eliminate mistakes. This process was iterative, therefore most of the errors reported being eliminated from the script until the end of the process. The resulting corpus was exported into multiple formats, including CSV and NIF.

The remaining texts were collected from the open source repositories TVTropes[4] and WikiNews[5]. Currently the following partitions are available (we indicate the sources in parentheses):

- *Franchises* (TV Tropes) set is focused on big multimedia franchises like *Marvel Cinematic Universe* or *Star Wars* and the creative works in various formats (movies, TV shows, books, video games) that support them. Due to the fact that many NEL tools are not trained to correctly recognize creative works and due to the popularity prior settings used by various algorithms, this partition is generally considered difficult for current NEL tools.
- *RegionalTV* (TV Tropes) set contains texts about European TV Shows.
- *EuroFilm* (TV Tropes) is focused on classic and modern European films. We have typically included five to fifteen movies for the selected countries (France, Germany, Austria, Switzerland, U.K., Netherlands, Italy, Denmark, Sweden).
- The *WebMedia* (TV Tropes) set was built around YouTube influencers. Due to the dynamic nature of the YouTube platform, the channels of some of the annotated entities may not exist in the future. Therefore, the content from this partition should be considered time-sensitive and relatively difficult. For the current version of the corpus the various types of YouTube subcultures[6] were not annotated, but we consider including such types in future versions.
- *News* (Wiki News) collects general interest News on a variety of topics. As expected, the level of difficulty for this partition is medium, since most tools were trained for such content.
- *Politics* (Wiki News) encompasses general politics News related to elections, political events (e.g., Syrian Conflict, Arab Spring) and war-related News. In some cases there

[4] https://tvtropes.org/
[5] https://en.wikinews.org/
[6] https://tvtropes.org/pmwiki/pmwiki.php/UsefulNotes/Subcultures

Category	Count
Partitions	10
Documents per partition	100
Documents	1000
Total entities	3422
Total entities (ØMIN)	3422
Total entities (ØMAX)	3655
Total entities (OMAX)	3809

Table 3: Basic statistics.

might be overlaps between this partition and the News partition, but this is simply due to the fact that they include the same types of entities.

- *Business* (Wikipedia) includes documents on corporations from the tech, medical and media domains. The difficulty level is medium.
- *Climate* (Wikipedia) contains coverage on Climate Change, sustainability and related entities (e.g., Greenpeace, Al Gore, Greta Thunberg).
- *Entertainment News* (Wikipedia) is a partition related to celebrity News during early 2020.
- *Core set (General)* (Wikipedia) partition contains the core set of the first 9 partitions, including short texts from domains like general news, politics, franchises, TV or movies. The level of difficulty is generally medium.

Some basic statistics about this corpus can be found in Table 3. Most of the documents have one to three sentences and can be considered equivalent to DBpedia abstracts, even though they were collected from various sources. Due to the nature of the collected information (e.g., franchises, TV shows, books) the early partitions often draw upon abstracts. Later partitions, in contrast, are randomly selected from the actual content of the articles. This assures that the collection is heterogeneous and that it can later be used for testing multiple use cases (e.g., media-related entities, news media, policies, etc).

4 Evaluation

This section describes the tools used during the evaluation, its design and a discussion around results.

4.1 Evaluation Design and Results

The following tools have been used during our general-purpose evaluation:

- **DBpedia Spotlight** (Daiber et al., 2013) is a statistical NEL engine that was originally built as a demo for showcasing DBpedia's capabilities and has been ported to multiple languages. The statistical models from Spotlight are really good for larger Knowledge Extraction or WSD challenges where all words need to be linked to their respective KG entities, but they are not necessarily fine-tuned for typed NEL tasks.
- **AIDA** (Hoffart et al., 2011) uses graph-based disambiguation algorithms and is considered one of the best NEL engines focused around Wikipedia linking.
- **Recognyze** (Weichselbraun et al., 2019b) is a multi-KG (e.g., DBpedia, Wikidata, Wikipedia) graph-based disambiguation engine focused on the issue of name variance.

We have created two different sets of evaluations. The first one (see Table 4) contains the results for the core set partition of the corpus. This evaluation also lists the results for different annotation styles (lenses) as they were presented in this paper. The second evaluation (see Table 5) compares the results of several tools on the entire corpora.

As it can be seen in Table 4, the multiple annotation styles had an impact on almost all of the evaluated tools. This suggests that most of the tools do seem to perform better when considering these annotation styles in succession, with Spotlight and Recognyze gaining up to 4%. It is interesting to note that the rules seem to improve the recall of DBpedia, Spotlight and Recognyze in all cases, whereas precision is not impacted by OMAX styles for Spotlight. There might be a need for multiple evaluations in a future publication to establish the full impact of these guidelines, but since such annotation styles can automatically be generated from any dataset following the outlined rules, they are definitely worth investigating.

As expected the results on the core set (Table 4) and the entire corpus (Table 5) are similar. They are between 2% and 3% lower for each tool, which due to the higher number of entities should be considered a good results. This also indicates that the content of the core set was indeed carefully chosen

Corpus	*System*	*mP*	*mR*	*mF*1	*MP*	*MR*	*MF*1
Core set	AIDA	0.47	0.48	0.47	0.43	0.48	0.43
ØMIN	Spotlight	0.53	0.43	0.48	0.35	0.42	0.37
(480 entities)	Recognyze	**0.61**	**0.52**	**0.56**	**0.52**	**0.50**	**0.51**
Core set	AIDA	0.49	0.48	0.49	0.45	0.48	0.44
ØMAX	Spotlight	0.55	0.43	0.48	0.35	0.40	0.36
(507 entities)	Recognyze	**0.62**	**0.54**	**0.58**	**0.55**	**0.52**	**0.53**
Core set	AIDA	0.49	0.48	0.49	0.45	0.48	0.44
OMAX	Spotlight	0.51	0.57	0.54	0.51	**0.58**	0.52
(527 entities)	Recognyze	**0.65**	**0.61**	**0.64**	**0.61**	0.57	**0.59**

Table 4: Core set experiments with multiple lenses (*m* - micro; *M* - macro; *p* - precision; *r* - recall; *F*1 - F1).

Corpus	*System*	*mP*	*mR*	*mF*1	*MP*	*MR*	*MF*1
	AIDA	0.50	**0.49**	0.50	0.46	0.47	0.46
All partitions	Spotlight	0.65	0.48	0.54	0.66	**0.50**	0.50
(3809 entities)	Recognyze	**0.69**	**0.49**	**0.57**	**0.70**	0.48	**0.56**

Table 5: Results on the entire corpora - with overlaps - OMAX lens (*m* - micro; *M* - macro; *p* - precision; *r* - recall; *F*1 - F1).

to reflect the content of the whole dataset up to this point in time.

4.2 Discussion

NEL performance on the *In Media Res* corpus is considerably lower than the results obtained on traditional data sets. This was expected due to the large amount of errors introduced by adding creative works to the corpora. Also as expected, the tools were not able to distinguish well between a character and the franchise that bears its name or offer good results on the YouTube influencer partition of the corpus. The influencer partition is especially difficult due to the fact that some of the works mentioned there (e.g., YouTube channels that were shut down or early gigs for famous influencers) are NIL (i.e. entities that have not been included in Wikipedia or related KGs such as DBpedia and Wikidata).

While some media franchises are well-covered by Wikipedia (e.g., *Harry Potter*) and related KGs, others are not. In such cases a good approach towards improving coverage and results might be leveraging Linked Data extracted from dedicated wikis such as *Memory Alpha* (covering the *Star Trek* franchise), *Wookieepedia* (covering *Star Wars*) or *Marvel Database* (covering both *Marvel Comics* and the *Marvel Cinematic Universe*). Most of the fandoms organize around such wikis and also many of them are published through *Fandom*[7] and similar wiki engines. Some of the information from these wikis is also collected in Linked Data form through DBkwik (Hertling and Paulheim, 2018).

[7] https://www.fandom.com/

5 Conclusion and Future Work

The road towards designing culturally-aware AIs has started with the expansion of fields like digital humanities during the last decade. While some steps towards this goal were made, the selected topics still depend on funding and on the researcher's own goals, as this is a relatively new field. Even so, we find it surprising that there was a lack of guidance in the NEL community related to name variation and nested entities for creative works. The corpus and concepts introduced in this paper are a first attempt to address this issue.

The results of the various annotators on the *In Media Res* corpus are understandably lower than on corpora from traditional domains such as news articles and social media. While applying different annotation lenses improves the results for some annotators, it is clear that there is a need for more progress in this area.

We plan to continue maintaining the corpus and provide updates, new partitions, or lenses. We hope that these efforts will contribute to increased accuracy in the annotation of creative work entities and, therefore, aid annotation systems in taking a further step towards culturally-aware AIs.

Acknowledgments

This research has been partially funded through the following projects: the ReTV project (www.retv-project.eu) funded by the European Union's Horizon 2020 Research and Innovation Programme (No. 780656), and MedMon (www.fhgr.ch/medmon) funded by the Swiss Innovation Agency Innosuisse.

References

Colin R. Batchelor, Christian Y. A. Brenninkmeijer, Christine Chichester, Mark Davies, Daniela Digles, Ian Dunlop, Chris T. A. Evelo, Anna Gaulton, Carole A. Goble, Alasdair J. G. Gray, Paul T. Groth, Lee Harland, Karen Karapetyan, Antonis Loizou, John P. Overington, Steve Pettifer, Jon Steele, Robert Stevens, Valery Tkachenko, Andra Waagmeester, Antony J. Williams, and Egon L. Willighagen. 2014. Scientific Lenses to Support Multiple Views over Linked Chemistry Data. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*, pages 98–113. Springer.

Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, and Daniela Rus. 2018. Data-Dependent Coresets for Compressing Neural Networks with Applications to Generalization Bounds. *CoRR*, abs/1804.05345.

Adrian M. P. Braşoveanu, Lyndon J.B. Nixon, and Albert Weichselbraun. 2018a. StoryLens: A Multiple Views Corpus for Location and Event Detection. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS 2018)*, Novi Sad, Serbia. ACM.

Adrian M. P. Braşoveanu, Giuseppe Rizzo, Philipp Kuntschick, Albert Weichselbraun, and Lyndon J.B. Nixon. 2018b. Framing Named Entity Linking Error Types. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 266–271, Paris, France. European Language Resources Association (ELRA).

Diego Calvanese, Alessandro Mosca, José Remesal, Martín Rezk, and Guillem Rull. 2015. A 'historical case' of Ontology-Based Data Access. In *2015 Digital Heritage, Granada, Spain, September 28 - October 2, 2015*, pages 291–298. IEEE.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pages 121–124. ACM.

Niels Dekker, Tobias Kuhn, and Marieke van Erp. 2018. Evaluating Social Network Extraction for Classic and Modern Fiction Literature. *PeerJ Prepr.*, 6:e27263.

Maud Ehrmann, Guillaume Jacquet, and Ralf Steinberger. 2017. Jrc-names: Multilingual entity name variants and titles as linked data. *Semantic Web*, 8(2):283–295.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Ströbel, and Raphaël Barman. 2020. Language Resources for Historical Newspapers: The Impresso Collection. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 958–968. European Language Resources Association.

Dan Feldman, Sedat Ozer, and Daniela Rus. 2017. Coresets for Vector Summarization with Applications to Network Graphs. *CoRR*, abs/1706.05554.

Dan Feldman, Mikhail Volkov, and Daniela Rus. 2016. Dimensionality Reduction of Massive Sparse Datasets Using Coresets. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2766–2774.

Wanling Gao, Jianfeng Zhan, Lei Wang, Chunjie Luo, Daoyi Zheng, Fei Tang, Biwei Xie, Chen Zheng, Xu Wen, Xiwen He, Hainan Ye, and Rui Ren. 2018. Data Motifs: A Lens Towards Fully Understanding Big Data and AI Workloads. In *Proceedings of the 27th International Conference on Parallel Architectures and Compilation Techniques, PACT 2018, Limassol, Cyprus, November 01-04, 2018*, pages 2:1–2:14. ACM.

Ben Hachey, Joel Nothman, and Will Radford. 2014. Cheap and Easy Entity Evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 464–469. The Association for Computer Linguistics.

Sven Hertling and Heiko Paulheim. 2018. DBkWik: A Consolidated Knowledge Graph from Thousands of Wikis. In *2018 IEEE International Conference on Big Knowledge, ICBK 2018, Singapore, November 17-18, 2018*, pages 17–24. IEEE Computer Society.

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, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 782–792.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchís, Jorge Civera, and Alfons Juan. 2019. Europarl-ST: A Multilingual Corpus For Speech Translation Of Parliamentary Debates. *CoRR*, abs/1911.03167.

Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, and Cash Costello. 2017. Overview of TAC-KBP2017 13 Languages Entity Discovery and Linking. In *Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, Maryland, USA, November 13-14, 2017*, page 4. NIST.

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A Neural Layered Model for Nested Named Entity Recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1446–1459. Association for Computational Linguistics.

Maurizio Lenzerini. 2018. Managing Data through the Lens of an Ontology. *AI Magazine*, 39(2):65–74.

Yunjia Li, Giuseppe Rizzo, José Luis Redondo García, Raphaël Troncy, Mike Wald, and Gary Wills. 2013. Enriching Media Fragments with Named Entities for Video Classification. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pages 469–476. International World Wide Web Conferences Steering Committee / ACM.

Diana Maynard. 2009. GATE: Bridging the Gap between Terminology and Linguistics. In *Proceedings of the 8th International Conference on Terminology and Artificial Intelligence, Toulouse, France, November 18-20, 2009*, volume 578 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Lyndon J. B. Nixon, Evlampios E. Apostolidis, Foteini Markatopoulou, Ioannis Patras, and Vasileios Mezaris. 2019. Multimodal Video Annotation for Retrieval and Discovery of Newsworthy Video in a News Verification Scenario. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part I*, volume 11295 of *Lecture Notes in Computer Science*, pages 143–155. Springer.

Lyndon J. B. Nixon and Raphaël Troncy. 2014. Survey of Semantic Media Annotation Tools for the Web: Towards New Media Applications with Linked Media. In *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, volume 8798 of *Lecture Notes in Computer Science*, pages 100–114. Springer.

Fabian Odoni, Philipp Kuntschik, Adrian M. P. Braşoveanu, and Albert Weichselbraun. 2018. On the importance of drill-down analysis for assessing gold standards and named entity linking performance. In *Proceedings of the 14th International Conference on Semantic Systems, SEMANTICS 2018, Vienna, Austria, September 10-13, 2018*, volume 137 of *Procedia Computer Science*, pages 33–42. Elsevier.

Raghu Rajkumar, Nate Foster, Sam Lindley, and James Cheney. 2013. Lenses for Web Data. *ECEASST*, 57.

Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. GERBIL - Benchmarking Named Entity Recognition and Linking Consistently. *Semantic Web*, 9(5):605–625.

Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2018a. 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*, volume 11137 of *Lecture Notes in Computer Science*, pages 170–186. Springer.

Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2019. NIFify: Towards Better Quality Entity Linking Datasets. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019.*, pages 815–818. ACM.

Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2018b. What Should Entity Linking link? In *Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management, Cali, Colombia, May 21-25, 2018.*, volume 2100 of *CEUR Workshop Proceedings*, page 15. CEUR-WS.org.

Tomasz Stanislawek, Anna Wróblewska, Alicja Wójcicka, Daniel Ziembicki, and Przemyslaw Biecek. 2019. Named Entity Recognition - Is There a Glass Ceiling? In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 624–633. Association for Computational Linguistics.

Albert Weichselbraun, Adrian MP Brasoveanu, Philipp Kuntschik, and Lyndon JB Nixon. 2019a. Improving Named Entity Linking Corpora Quality. *RANLP 2019*, pages 1328–1337.

Albert Weichselbraun, Philipp Kuntschik, and Adrian M. P. Brasoveanu. 2019b. Name Variants for Improving Entity Discovery and Linking. In *2nd Conference on Language, Data and Knowledge, LDK 2019, May 20-23, 2019, Leipzig, Germany.*, volume 70 of *OASICS*, pages 14:1–14:15. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik.

Lei Zhang, Andreas Thalhammer, Achim Rettinger, Michael Färber, Aditya Mogadala, and Ronald Denaux. 2017. The xLiMe system: Cross-lingual and Cross-modal Semantic Annotation, Search and Recommendation over Live-TV, News and Social Media Streams. *J. Web Semant.*, 46-47:20–30.