Mining and Leveraging Background Knowledge for Improving Named Entity Linking

Albert Weichselbraun
Swiss Institute for Information Research
University of Applied Sciences Chur
Chur, Switzerland
albert.weichselbraun@htwchur.ch

Philipp Kuntschik
Swiss Institute for Information Research
University of Applied Sciences Chur
Chur, Switzerland
philipp.kuntschik@htwchur.ch

Adrian M.P. Braşoveanu
Swiss Institute for Information Research
University of Applied Sciences Chur
Chur, Switzerland
adrian.brasoveanu@htwchur.ch

ABSTRACT

Knowledge-rich Information Extraction (IE) methods aspire towards combining classical IE with background knowledge obtained from third-party resources. Linked Open Data repositories that encode billions of machine readable facts from sources such as Wikipedia play a pivotal role in this development.

The recent growth of Linked Data adoption for Information Extraction tasks has shed light on many data quality issues in these data sources that seriously challenge their usefulness such as completeness, timeliness and semantic correctness. Information Extraction methods are, therefore, faced with problems such as name variance and type confusability. If multiple linked data sources are used in parallel, additional concerns regarding link stability and entity mappings emerge.

This paper develops methods for integrating Linked Data into Named Entity Linking methods and addresses challenges in regard to mining knowledge from Linked Data, mitigating data quality issues, and adapting algorithms to leverage this knowledge.

Finally, we apply these methods to Recognyze, a graph-based Named Entity Linking (NEL) system, and provide a comprehensive evaluation which compares its performance to other well-known NEL systems, demonstrating the impact of the suggested methods on its own entity linking performance.

CCS CONCEPTS

• **Information systems** → Incomplete data; Inconsistent data; Extraction, transformation and loading; Data cleaning; Entity resolution; • **Computing methodologies** → Information extraction;

KEYWORDS

Knowledge-rich Information Extraction, Named Entity Linking, Linked Data Quality, Information Extraction, Semantic Technologies, Natural Language Processing

1 INTRODUCTION

Named Entity Linking (NEL) is an **Information Extraction (IE)** technique that identifies mentions of named entities in textual content and grounds these mentions to a Knowledge Base such as Wikipedia, Google Knowledge Graph [25], DBpedia [10] and Wikidata [28]). NEL is also quite often formulated as a problem of **Knowledge Base Population (KBP)** in which a system is required to extract triples from a text corpora that represent the available knowledge on one or all entities (traditionally those that belong to the classes Person, Organization or Location) in the respective corpora [9]. This description of the problem also hints to the interdisciplinary and challenging nature of NEL, as in order to successfully address it, it is often required to combine methods from different fields, especially Natural Language Processing (NLP), Semantic Web (SW) or Machine Learning (ML).

In recent years, due to the emergence of Open Data portals, a new class of IE methods that leverages background knowledge extracted from Linked Data to improve the performance of IE has gained traction in application areas like sentiment analysis, fake news assessment or NEL. These recent knowledge-intensive methods benefit from the open availability of comprehensive information from open Knowledge Bases like Wikipedia, DBpedia, Wikidata or ConceptNet [24], but they also pave the way for novel strategies for improving the performance of IE methods. Such strategies may, for instance, (i) address the quality of the available background knowledge by mitigating data quality problems (ii) increase the coverage and amount of the available knowledge by improving the knowledge mining processes (iii) boost the efficiency of how the extracted knowledge is used within the NEL method.

Using background knowledge also opens the way for applying more advanced paradigms to information extraction and NLP tasks that draw upon semantics or even pragmatics rather than syntactic word-based techniques [2]. Linked Data sources provide access to billions of computer-readable statements encoding background knowledge and, therefore, plays a pivotal role in moving from simple text representations to more sophisticated ones. Nevertheless, the diversity of the Linked Data ecosystem, heterogeneous data quality standards and outdated datasets considerably complicate its use for Information Extraction.

1.1 Contributions

The main contributions of this work are

(1) the introduction of methods for mining background knowledge from Linked Data, mitigating data quality issues and adapting NEL algorithms to leverage this knowledge and to maximize its impact on the information extraction process (Section 3).
(2) the application of the suggested strategies and techniques to Recognyze, a state of the art NEL component that draws upon background knowledge from Linked Data repositories such as DBpedia, GeoNames and WikiData (Section 4)
(3) a comprehensive evaluation and discussion of the presented strategies and their impact on the performance of the NEL performance achieved by Recognyze (Section 5).

2 RELATED WORK

The following discussion of related work describes the state of the art in NEL and provides an overview of research on Linked Data quality issues.

2.1 Named Entity Linking Systems

Named Entity Linking [9] is becoming more important as more and more entities are present on the web with their official URLs, DBpedia URIs [11] or Wikidata [28]. Traditionally, only three main classes of entities: **Person - PER, Organization - ORG, LOC - Location** were included in the NEL competitions. Location is considered by far the most difficult class since it has a lot of conflicts with other classes (e.g. street names often contain people names, there are many confusions between organizations and the buildings in which they are located, etc.). More recently, location has been split into three classes for the TAC KBP challenges [9]: **Natural Locations - LOC** like mountains, rivers or lakes; **Geo-Political Entities - GPE** like countries, regions, cities, streets, and **Facilities - FAC** like airports, road infrastructure, parks or buildings. Besides splitting Location into multiple classes based on the type of location entity, new classes are added each year, for example, **Event - EVENT** or **Product - PROD**. Such an expansion is necessary in order to advance the state of the art and has become standard in the last NEL competitions. Each year, several challenges are organized, the most important being the NIST's TAC-KBP [9] traditionally organized for English, Chinese and Spanish text. Recently competitions also expanded the number of languages from three to 13, piloting even languages of some smaller countries (e.g. Albania). We have chosen to work only on the three classic classes (PER, ORG, LOC) as our current focus is on data quality mitigation strategies for improving NEL performance.

The most successful NEL systems can be included into the following three classes: **graph-based disambiguation models** (e.g., AIDA [8], [14] and AGDISTIS [26]) which exploit the links between the entities included in the text with the intention of using these relations for disambiguation. **statistical models** including mixtures of Conditional Random Fields models - (e.g., ADEL [18] or DBpedia Spotlight [3]) exploit classic Machine Learning approaches. More recently, **neural models** (e.g., the Convolutional Semantic Similarity model for NEL proposed by Francis-Landau [6]) are used in order to jointly resolve the detection and resolution of links. Regardless of the model that is globally used for disambiguation, all NEL tools need to link the entities to a target Knowledged Base, therefore they need to exploit the relations between the entities or the graph structure of Linked Open Data. Our own tool, Recognyze, builds upon the graph-based disambiguation method.

The NEL field is going through a period of consolidation, with significant surveys appearing in the last years. Derczynski et al. [4] analyze the NEL pipelines used for short texts (e.g., tweets, microblogs) and offer solutions for better pre-processing (e.g., language identification, POS tagging, normalization). Rizzo et al. [20] summarize the lessons learned during the several editions of the NEEL Challenge with a focus on changes to the annotation methodology, corpus analysis, emerging trends in the design and evaluation of NEL system. The work also includes a long analysis of the evaluation measures (e.g. scorers) used during these challenges and is notable for the inclusion of all the major systems that were launched in the last five years. Ozdikis et al. [17] focus strictly on location detection techniques from short texts (e.g. Twitter) and analyze the best algorithms for jointly estimating the real location of Twitter events. The scalable architecture used for geoparsing and geosemantics extraction in the EU REVEAL project [13] included features like the tweet content, position of the terms, part-of-speech (POS) sets, and 3-gram feature sets that combined named entities with their POS tags. It is clear that the key to obtaining good results is to focus on continuously improving the NEL pipelines.

Understanding and classifying the NEL results is also a growing research topic in itself. GERBIL [27] integrates multiple tools and publishes evaluation results to a web interface without offering additional explanations for each mention. TAC-KBP neleval [7] provides a simple solution that is used for both reporting the results and offering a primary error analysis for all mentions (e.g., correct link, wrong link, extra link, and so on).

2.2 Linked Data Quality Issues

DBpedia which is derived from Wikipedia is considered one of the most prominent Linked Data sources. It contains structured information that has been extracted from Wikipedia by automatic exploration tools together with manually crafted property mappings. A major part of creating the original semi-structured knowledge base, as well as the machine readable property mappings relies on human labor [10] which is by its very nature individual, costly and error-prone and, therefore, a significant source for data quality issues.

Zaveri et al. [31] provide a comprehensive summary of different data quality issue dimensions that can arise in Linked Data sources and conduct a systematic review of the existing approaches to assess these quality issues, analyzing 21 relevant papers published between 2002 and 2012. They classify the data quality issues based on error correction strategies into errors that can be solved by (i) amending the extraction framework, (ii) correcting the property mapping, or (iii) adjusting the semi-structured knowledge base. Ristoski and Paulheim [19] suggest to deal with data problems in a separate data preprocessing step that handles missing values, identifies incorrect data, eliminates duplicates and performs conflict resolution. Weichselbraun and Kuntschik [29] discuss the impact of these data quality issues on knowledge extraction methods and investigate different mitigation strategies for the corresponding dimensions. They suggest to integrate these strategies into graph mining and information extraction methods and provide real-world use cases of such mitigation strategies.

A taxonomy of the various errors that can be found in NEL evaluations can be found in Braşoveanu et al, [1]. The article identifies several large categories of errors that plague today's evaluations: Knowledge Base errors (KB), Dataset errors (DS), Annotator Errors (AN), NIL Clustering Errors (NIL) or even Scorer Errors (SE). Since our work is not currently focused on NIL Clustering, we have only considered the other four error classes in this work.

This paper builds upon this research by introducing knowledge extraction and data quality mitigation strategies for Linked Data and discussing how the extracted knowledge may be used to refine Named Entity Linking algorithms.

3 METHOD

Unfolding the full potential of background knowledge for NEL components requires (i) innovations on the algorithmic level, i.e. methods which are capable of capitalizing the available information that are complemented by (ii) strategies for mining relevant background knowledge, e.g., by mitigating quality issues and transforming knowledge into data structures suitable for machine learning and information extraction algorithms.

Data quality mitigation strategies particularly focus on the following four data quality dimensions described by Zaveri et al. [31]:

(1) *Completeness* in terms of population completeness (i.e. the coverage in terms of individuals within the dataset) and property completeness (i.e. that all properties relevant to a particular individual are available in the dataset).
(2) *Relevancy* that requires the extracted knowledge to be relevant to the application domain
(3) *Semantic accuracy* which refers to the correctness of the available data
(4) *Timeliness*, i.e. the recency of the extracted knowledge

Figure 1 outlines the relation between background knowledge and algorithmic improvements of NEL algorithms and how techniques that mine and mitigate quality issues further contribute towards the creation of more sophisticated NEL methods.

3.1 Optimize data mining

3.1.1 Domain-specific queries. Adapting data mining to the application domain increases the background knowledge's relevancy and therefore boosts the precision of subsequent NEL tasks.

Historical entities such as dissolved countries and empires, for example, often trigger ambiguities with their successors. Since most applications focus on a particular temporal period, filtering entities by their dissolution date or year improves the efficiency of the disambiguation process. In the experiments presented in this paper, for instance, we removed all entities that have already been dissolved by checking if a *dbo:dissolutionDate* or *dbo:dissolutionYear* property is present.

Table 1 lists a number of historical geographic entities and their corresponding dissolution date. The example also demonstrates the heterogeneity of the available data. For Yugoslavia, for instance, no dissolution date has been recorded in DBpedia as of January 2018 and for the Japanese colonial empire only the dissolution year is available.

Table 1: Examples for historical entities and the corresponding dissolution data or year recorded in DBpedia.

historical entity	dissolution date/year
Austria-Hungary	1918-11-11
Colonial Brazil	1815-12-16
Japanese colonial empire	1945
Roman Empire	1453-05-29
West Germany	1990-10-03
Yugoslavia	-

3.1.2 Domain-specific entity filters. Linked Data sources such as DBpedia often organize entities in extensive type hierarchies that spawn manifold types (*rdf:type*) and subjects (*dct:subject*). Restricting knowledge mining to types that are of actual relevance to the domain or evaluation task considerably reduces the amount of total ambiguities and, therefore, the overall precision.

For the evaluations presented in Section 5, for instance, we only considered organizations (ORG), persons (PER) and locations (LOC). Adapting the knowledge mining by removing irrelevant entity types such as music bands, movies, television series and fictional characters increased the relevancy of the mined knowledge and improved the NEL component's performance.

3.1.3 Combine heterogeneous data sources. Integrating complementing linked open data sources addresses completeness issues since the additional data sources might yield further individuals as well as properties for each individual. In our experiments, for instance, combining DBpedia with Wikidata has proven to be particularly beneficial for mining additional name variants for named entities. GeoNames, in contrast, is a valuable source for adding hierarchical information (country, state, administrative unit, etc.) to geographic entities in DBpedia.

3.1.4 Extract link anchor text from Wikipedia. Roth et al. [22] have shown that expanding queries with Wikipedia anchor text of links that point to the corresponding entities (e.g. "U.S. president" to "Donald Trump") significantly improved recall of their slot filling approach. Slot filling is another popular subtask of TAC-KBP challenges, but unlike linking, systems are required to fill in slots with various properties of selected entities instead of simply returning the entities and links. Inspired by this approach we, therefore, extracted the anchor text and linked entities from the 1 December 2017 Wikipedia dump and removed all anchor texts that did not point to a unique named entity. Afterwards we created a Linked Data source that connects DBpedia entities to these link anchor texts using the *skos:altLabel* property. The Recognyze graph mining component then integrated the available knowledge with the existing DBpedia dataset.

The created knowledge source has a number of benefits for Named Entity Linking: (i) the extracted anchor text captures a wide range of name variations that occur in actual sentences [22], addresses the differences between recorded entity names and name variants used in informal texts and News articles, and improves the completeness of the background knowledge. (ii) The name variations have been extracted from a data source that is much more recent (December 2017 versus April 2016) than the current DBpedia version and, therefore, also captures facts that are not yet available in DBpedia dumps which provides benefits in terms of timeliness of the available knowledge. Table 2 illustrates some of the expansions that have been obtained by using this technique.

3.2 Mitigate data quality issues

Data quality issues within the Linked Data sources are another major concern. Recognyze uses a powerful graph mining, natural language processing and pre-processing pipeline to improve the *semantic accuracy*, *completeness*, *relevancy* and *timeliness* of the available information (Section 4).

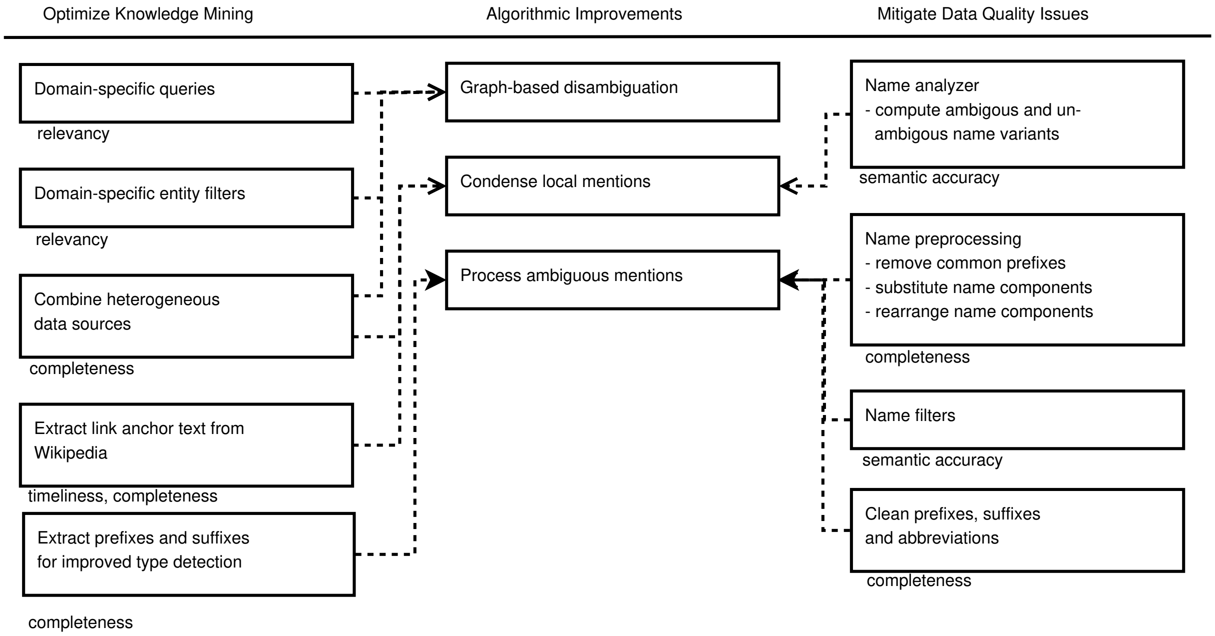
Optimize Knowledge Mining
Algorithmic Improvements
Mitigate Data Quality Issues
Domain-specific queries
relevancy
Domain-specific entity filters
relevancy
Combine heterogeneous data sources
completeness
Extract link anchor text from Wikipedia
timeliness, completeness
Extract prefixes and suffixes for improved type detection
completeness
Graph-based disambiguation
Condense local mentions
Process ambiguous mentions
Name analyzer
- compute ambigous and un-ambigous name variants
semantic accuracy
Name preprocessing
- remove common prefixes
- substitute name components
- rearrange name components
completeness
Name filters
semantic accuracy
Clean prefixes, suffixes and abbreviations
completeness

Figure 1: Improving Named Entity Linking with background knowledge

Table 2: Alternative mention names obtained by extracting link anchor text from Wikipedia

link target text	DBpedia entity
St. Peterskirche	St. Peter, Zurich
Rui Shijō	Rui Shijo
Héctor Veira	Héctor Rodolfo Veira
Fakir Khana Art	Gallery Fakir Khana
Nysaker	Näsåker
UNBRO	United Nations Border Relief Operation
US SOF	United States special operations forces
Jeremy Crispian Stanley	Jeremy Stanley
SV Innsbruck	Sportsverein Innsbruck
Parsimonious	Parsimony
San Vicente Mártir	Valencia Catholic University Saint Vincent Martyr

Data quality is important for Knowledge Bases due to multiple factors, but the ones that are most likely to cause issues are: (i) name variance; (ii) class (type) confusability; (iii) lack of stable links; and (iv) entity mappings.

Name variance is the problem of finding all the names that refer to a single entity within a collection of texts. Several cases of variance have been described in the literature: a) known aliases ("Robert Gailbraith", a pseudonym used by J.K. Rowling; "John Barron" for "Donald Trump"); b) hypocorisms or common aliases ("Bobby" for Robert, "Liz" for "Elizabeth"); c) abbreviations ("JFK" for both "John F. Kennedy" and "John F. Kennedy International Airpot"); d) multilingual names ("Austria" can have different names or spelling depending on the language: in German it will be "Österreich", in French "Autriche", or "Ausztria" in Hungarian); e) partial matches (names of royal figures often fall under this rule; e.g., you will more often find links to "Prince Charles" instead of "Charles, Prince of Wales"). Additionally, each entity type might have its own name variance rules. People names frequently include titles ("Senator", "Judge", "HRH", etc.) or nicknames. Organization names are often abbreviated through different methods that might involve: classic abbreviations (e.g., "NBA"), cutting suffixes (e.g., "Corp" or "Inc"); removing country or branch names ("Sony Europe" might often be referred simply as "Sony"); combining parts of words (e.g., "Nortel" for "Northern Telecom"). Locations have more problems with name variances than the other classes due to overlap and assimilation (e.g., people and organization names often contain location references), but can still include place qualifiers (e.g., N/E/S/W, "So" for "Southern"); abbreviations (e.g., "OH" for "Ohio"); embeddings or nested entities (e.g., "New York Stadium"); possessive names (e.g., "Hawaii's Waikiki"); addresses (e.g., "221B Baker Street").

Class (type) confusability is often related to the problem of name variance. As already mentioned, we notice frequent clashes between the main three classes due to name reusability (e.g., "King George Street" - people names used for street names, "Boston University" - geographical names used for street or company names). Therefore, it is always important to pay attention to the hints offered by the text related to the respective entity. In many cases confusability between classes appears due to the fact that multiple entity types are attributed to the same entity. For instance, when multiple main entity types have been added to a DBpedia entity we have an instance of this type of error.

Stability of links is the third factor that forces us to consider the quality of a KB due to issues like redirects, missing information, or old links. A stable link is a link that has not changed during the last Knowledge Base updates. It is really important to be able to asses this property of links, as, for example, when we are performing an evaluation we would like to know that the link has not changed between the DBpedia version used for annotating the gold standard and the current DBpedia version, for example. A common problem with current KBs is the frequency of RDF dumps publishing (e.g., six months for official DBpedia version, weekly for Wikidata versions) which also relates to the data quality issue of *timeliness*. DBpedia publishes major versions every six months, whereas Wikidata, for example, publishes weekly RDF dumps which contain more recent information.

Entity mappings generally refer to the manual or automated mappings created between the properties from Wikipedia and third-party KB like DBpedia or Wikidata. They can change slightly between versions and they can also sometimes yield weird results (e.g., a year that is typed as a person). Such bad mappings need to be reported to the KB maintainer, but fixing them can take weeks or months (depending on when the next dump is released). Entity mappings errors can be caused by the statistical biases of the algorithms that perform the extractions based on mappings, as well as by extension of the mapping rules into domains they were not supposed to cover, but they need to be fixed if we want reliable Knowledge Extraction services.

By examining all these factors and considering that these are only some of the most well-known problems, the need for a clear and strong data quality mitigation strategy arises, especially if we want to provide clean updated results to user queries. The remainder of this subsection provides some details about the components we have built in order to create such a strategy. All the strategies presented in this article are geared towards improving the disambiguation of entities with a larger number of name variances. Class confusability is addressed directly by the graph-disambiguation techniques discussed in the next subsection. Link stability and entity mappings are mostly problems that need to be fixed in Knowledge Bases, the a NEL tool simply returning whatever the KBs contained [1].

3.2.1 *Name analyzer.*

Name variance is addressed by the name analyzer component and all the subsystems related to it. Language and entity-type-specific name analyzers assess whether name variants are considered unambiguous (i.e. unique enough to refer to a mention of a named entity) or ambiguous. The analyzers use complex algorithms such as entropy metrics and heuristics for classifying names into these two categories [30]. If unambiguous names occur within a text they are considered mentions of the corresponding named entity. Ambiguous names, in contrast, need to be disambiguated in a prior step.

3.2.2 *Name preprocessing.*

Recognyze's name preprocessing components focus on increasing the completeness of the extracted name variances by splitting input strings s into tokens $t_i = \{t_1, ...t_n\}$ that are then used to generate additional name variants $n^1, ...n^m$ by

(1) removing common prefixes $\{t_1, ...t_i\}$ such as "U.S.", "United States", "European" which produces name variances such as "Department of Agriculture' for the "United States Department of Agriculture" and "Environment Agency" for "European Environment Agency". Adding the removed prefixes as context information allows using this information for subsequent disambiguation processes.
(2) substituting synonyms by replacing tokens $\{t_i, ...t_j\}$ with synonyms, if available. For instance, the name "United States Department of Commerce" can lead to the additional name variants "U.S. Department of Commerce" and "US Department of Commerce".
(3) drawing upon common name patterns for creating name variants by rearranging tokens. This allows us to automatically add name forms such as "Justice Department" from "Department of Justice".
(4) creating name variants for uppercase names and international names using the Unicode Normalization Form Canonical Decomposition (NFD) yields additional normalized name forms such as "Nestle" for "Nestlé".

Considering these additional name variants in the NEL process, improves the likelihood of grounding such name forms and, therefore, recall.

3.2.3 *Name filters.*

Name filters improve semantic accuracy by removing name variants that are too general to be used within the disambiguation process. Such filter terms include stop words, numbers and denonyms such as Londoner, New Yorker, Austrian, Swiss and American. Where such demonyms are known to also be famous organization names (e.g., New Yorker) or locations (e.g., Wells), they are typically not removed, but treated similarly to an ambiguous name.

3.2.4 *Extract prefixes, suffixes and abbreviations.*

Extracting prefixes, suffixes and abbreviations from Linked Data fields such as *dbo:abstract*, *rdfs:comment* and *dbp:caption* addresses property completeness issues. The prefix and suffixes extraction component draws upon language-specific dictionaries and provides disambiguation constraints such as named entity types that help in improving disambiguation performance.

The abbreviation extraction component uses heuristics for identifying potential abbreviations and adds them to the list of name variants for the underlying entity yielding abbreviations such as "CBM" for "Commodore Business Machines" or "IBM" for "International Business Machines".

3.3 Algorithmic improvements

3.3.1 *Graph-based disambiguation.*

Graph-based disambiguation is only possible due to Linked Data which provides relations between entities. When analyzing a text, a small Knowledge Graph (KG) is built which contains all the possible entities that appear on a text based on the found surface forms (candidates). Progressively, more and more candidates are eliminated due to their weak links with the main entities from the texts. For example, if the text contains a reference to Paris, this can point to multiple entities like Paris, France or Paris, Texas or even Paris, the prince of Troy. Let's assume that the entity mentioned in the text is Paris, France, then the rest of the entities should have some links (e.g., *locatedIn* or *bornIn* or others) that help infer that this refers to the Paris mentioned in the text. This method was found to perform the best, while also being quite robust and efficient [8]. Of course, in some cases, links between entities might not be enough for performing a clear disambiguation (e.g., a politician from a city visits another city). In such scenarios, several other indicators, like a popularity prior, might help. Exploiting the links between entities can also sometimes lead to strange side effects due to the fact that Wikipedia (and by association DBpedia and Wikidata since they are based on Wikipedia) contains lots of lists. Such lists would inevitably contain links to many entities from a text and, therefore, need to be filtered out when performing graph-based disambiguation (e.g., by using regular expression filters). The main improvement to the graph-based disambiguation algorithms implemented in Recognyze consists in combining data from heterogeneous data sources (e.g., multiple Knowledge Bases like DBpedia and Wikipedia, textual content, lexicons), as explained in the next paragraphs.

3.3.2 Condensed multiple mentions. News articles often contain multiple mentions of persons, organizations or locations that are introduced with a longer name variance such as "President Donald Trump". Later mentions of the same person, in contrast, only use short versions of the same name such as "President Trump" and "Trump". Usbeck et al. [26] leverage this information by assigning name variances that are substrings of other names to the same entity.

Leveraging the information on name variances obtained from the knowledge mining components, algorithms can not only condense such simple cases but also mentions that are abbreviations or other name variances of that name (e.g. "Mr. Trump" in the example above, or even "VW" for "Volkswagen").

Algorithm 1 outlines one simple strategy for merging multiple mentions. This process also reduces the number of potential candidate entities and therefore increases the accuracy of the disambiguation process.

Ambiguities such as mentions of Bill and Hillary Clinton in one article can be either addressed by skipping the condensation process or by using more advanced approaches that, for instance, consider constraints on entity types (e.g. due to the use of prefixes such as Mr., Mrs.).

Algorithm 1 Disambiguation: condense local mentions

```
procedure CondenseMentions($mentions$)
    $mentions \leftarrow$ sortByLength(mentions)
    for each Mention $m1$ in $mentions.ascendingOrder$ do
        for each Mention $m2$ in $mentions.descendingOrder$ do
            if $m2.surfaceText$ contains $m1.surfaceText$ then
                $m1.candidateEntities \leftarrow m2.candidateEntities$;
                $break$;
            end if
        end for
    end for
    return $mentions$
end procedure
```

3.3.3 Affixes for improved type detection and disambiguation. Prefixes and suffixes obtained from Linked Data provide valuable information for improving disambiguation algorithms. Many prefixes and suffixes such as "Dr.", "Inc." and "city" imply that the corresponding mention refers to a particular type such as a person, organization and location. Type is a very important clue about the correctness of an answer, since if the entity from a text is designated as Person, for example, the returned entity needs to have the same type. Leveraging this information, therefore, allows for a better disambiguation of mentions that refer to different types (the "Kingdom of Jordan" versus "Mr. Jordan") or even entities within a type (e.g. "General Marshall" versus "Prof. Marshall").

4 NAMED ENTITY LINKING

This section discusses the adaption of the methods introduced in Section 3 to the Recognyze Named Entity Linking component. Recognyze utilizes Linked Data for searching, disambiguating and linking mentions of entities in documents. We describe how Recognyze mines background knowledge, mitigates data quality issues and performs disambiguation. Section 5 then demonstrates the effectiveness of this approach.

4.1 Recognyze

Recognyze [30] draws upon one or multiple Linked Data repositories for background knowledge. Entity linking profiles specify (i) the used data repositories, (ii) the queries (SPARQL) for retrieving information from these repositories, (iii) the background knowledge acquisition pipeline which consists of different kinds of filters, preprocessors and analyzers, as well as (iv) the disambiguation algorithms that are used to identify and ground named entity mentions to their respective resource in the repository. Its background knowledge acquisition pipeline utilizes different methods to generate and validate name variants to maximize the extend and quality of information received from repositories. Preprocessors and analyzers are encapsulated within an simple communication interface which allows a flexible configuration and therefore optimization of the system as a whole.

Figure 2 illustrates Recognyze as a comprehensive system consisting of two essential subprocesses: (i) a background knowledge acquisition process that focuses on maximizing coverage and quality of the extracted knowledge, and (ii) an information extraction process which uses the extracted and refined knowledge to detect and ground mentions in a text to the corresponding entities in the knowledge source. Preprocessors clean the input data and generate additional name variants that, in this form, have no natural existence in the used Linked Data repository. Analyzers assess the generated name variants and determine whether they are unambiguous enough to count as a name (for example full names such as "Donald Trump"), if they should be threatened as ambiguous names (as for example tokens of a full name as "Donald") or even should be taken as a context term (as "Government" in this example). Filters at various positions within the knowledge acquisition pipeline allow to remove unwanted resources and misleading name variants during the process. A profile then describes the summary of the extracted name variants, context information, relations and links between entities, as well as the configuration of the information

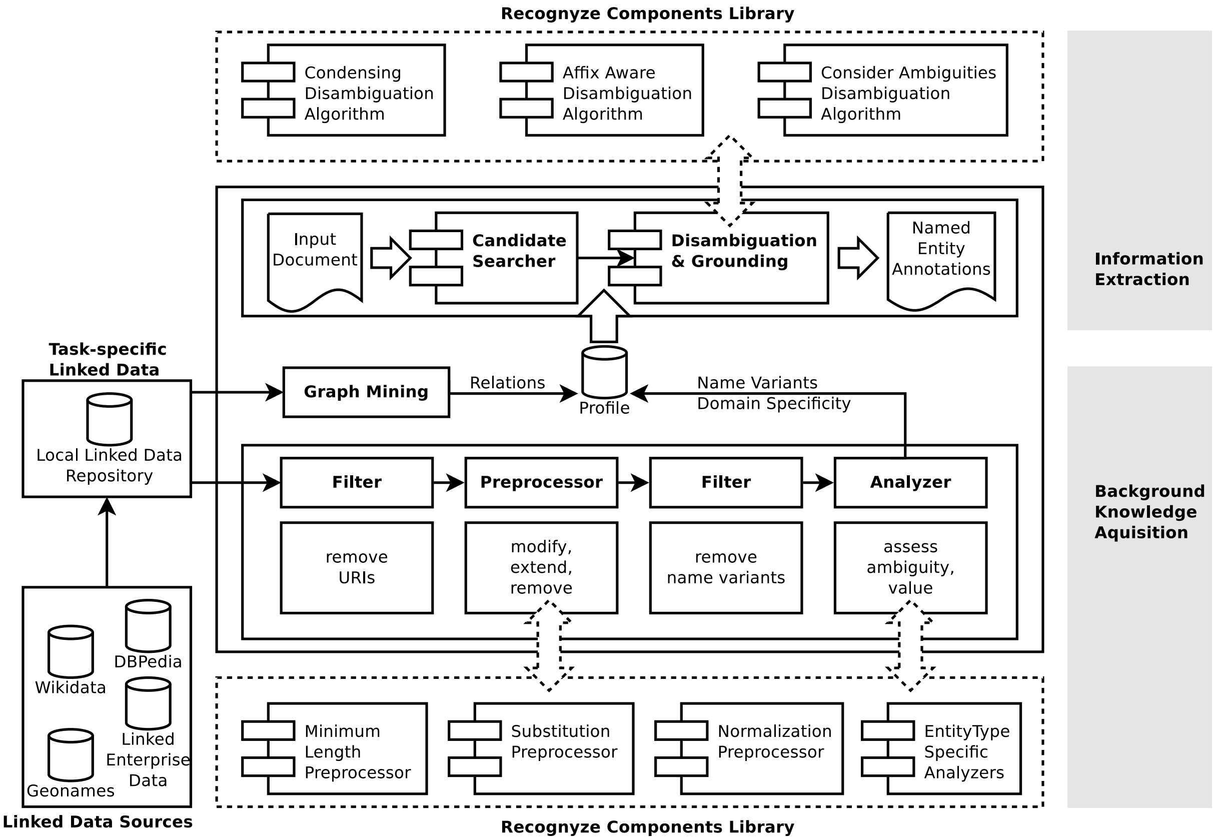
Recognyze Components Library
Condensing Disambiguation Algorithm
Affix Aware Disambiguation Algorithm
Consider Ambiguities Disambiguation Algorithm
Input Document
Candidate Searcher
Disambiguation & Grounding
Named Entity Annotations
Information Extraction
Task-specific Linked Data
Local Linked Data Repository
Graph Mining
Relations
Profile
Name Variants
Domain Specificity
Filter
Preprocessor
Filter
Analyzer
remove URIs
modify, extend, remove
remove name variants
assess ambiguity, value
Background Knowledge Aquisition
Wikidata
DBPedia
Linked Enterprise Data
Geonames
Linked Data Sources
Minimum Length Preprocessor
Substitution Preprocessor
Normalization Preprocessor
EntityType Specific Analyzers
Recognyze Components Library

Figure 2: The Linked Data driven background knowledge acquisition- and the information extraction process as two significant parts of the Recognyze Named Entity Linking component.

extraction pipeline on top and, therefore, builds the interface between the two subprocesses. The information extraction pipeline that builds upon the extracted knowledge performs two major tasks: (i) the recognition of candidate mentions in text, whereby candidates refer to an observed mention that may or may not represent a particular entity, and (ii) the disambiguation and grounding of these candidates, where unlikely candidate mentions are discarded and mentions are linked to the best fitting entity in the Knowledge Base.

4.2 Optimizing data mining

The configuration of a profile in Recognyze allows the free definition of SPARQL queries that are used in the background data acquisition phase. This yields two mayor advantages: (i) data can be excluded or exclusively included on a query level which allows us to remove entities based on the application domain and features in the repository (e.g. historical resources as described in Section 3.1.2); (ii) data from multiple repositories can be combined with each other (Section 3.1.4). This allows not only combining existing linked open data sources with each other but also integrating data stored in any repository accessible via SPARQL. Figure 3 demonstrates an example query that leverages these advantages. Our approach has been tested with both full Knowledge Base dumps, as well as with slices that contained the needed entity types. The slicing can currently be done directly via our engine, but for extremely large KBs it is often recommended to use a separate slicing tool (e.g., RDFSlice or Torpedo [12]).

4.3 Mitigating data quality issues

Recognyze utilizes simple interfaces between the main component and its components library that allow the adaption of its background knowledge acquisition pipeline to application domains and evaluation tasks as described in Section 3.2. Depending on the use case, preprocessors might, for instance, require a certain string length (*MinimumLengthPreprocessor*), filter names without letters, remove or replace invalid characters in names (*NormalizationPreprocessor*), and replace parts of a name variant with synonyms (*SubstitutionPreproceesor*). Analyzers use more complex algorithms such as entropy metrics [30] that allow the assessment of name variants to determine whether a name variant is considered too ambiguous to stand on its own.

4.4 Algorithmic Improvements

Recognyze allows the adaption of multiple disambiguation and grounding algorithms in a similar way as discussed in the previous section. Candidate mentions are subsequently processed by a

```
### DBpedia

SELECT DISTINCT ?s ?name WHERE {
  { ?s rdfs:label ?name.
    ?s a dbo:Place. }
  UNION
  { ?s rdfs:label ?name.
    ?s a dbo:Location. }
  FILTER NOT EXISTS
  { ?s dbo:dissolutionDate ?date. }
}

### Wikidata

SELECT DISTINCT ?s ?alternativename WHERE {
    ?wikidata owl:sameAs ?s.
    ?wikidata rdfs:label ?alternativename.
    FILTER(lang(?alternativename) = "en")
    FILTER(regex(str(?s), "dbpedia"))
}
```

Figure 3: Example queries for a simple profile that combines and enriches DBpedia entities relevant to the application domain (top) with additional name variants obtained from Wikidata (bottom).

pipeline which consists of multiple disambiguation algorithms, each building upon the result of its predecessor. The following evaluation utilizes three disambiguation algorithms together, each of them implementing a different mitigation strategy: (i) *CondensingDisambiguationAlgorithm* removes possible resources from a candidate mention if that mention represents a substring of another mention in the same text (as discussed in Section 3.3.2). (ii) *AffixAwareDisambiguationAlgorithms* remove possible resources from a candidate mention based on whether the resource's entity type corresponds to constraints imposed by affixes in the text (see Section 3.3.3). We observed this as particularly effective to differentiate between location and person entities. (iii) *ConsiderAmbiguitiesDisambiguationAlgorithm* determines if candidate mentions representing the same entity have been already disambiguated and grounded by previous algorithms and performs this grounding otherwise. This is especially effective for roles like "President" or abbreviations such as "VW".

5 EVALUATION

5.1 Datasets

Our evaluation draws upon three English gold standard datasets. We have used the English dataset **Reuters128** which contains 128 texts from Reuters and is part of the the **N3 collection** [21] that comprises three smaller corpora in German and English focusing on classic and recent News media. In these datasets the surface forms of the entities point towards the most popular entities bearing the respective name. We only used the three annotated entity types (PER, ORG, LOC). We have only used the Reuters128 segment of the collection since it was older than the other segments and a good candidate for testing older entities, being based on the classic Reuters dataset from 1980s. It was important to do this, as for example if we examine organizations, in such a long interval (close to 40 years), they can often expand, merge or disappear.

OKE2015 [15] and **OKE2016** [16] are two datasets used during the SemEval at ESWC conferences. They contain short sentences (less than 200 sentences each dataset) which quickly describe one subject (e.g., short DBpedia abstracts). We have only used the datasets for task one (NEL) and also selected the three classic entity types.

Reuters128 covers older events, the texts being extracted from the classic Reuters dataset that has been originally published in the 1980s, whereas the OKE challenges cover more recent events, but the texts are somewhat encyclopedic in nature, even though shorter. We used these datasets in order to have a balanced view over the results, therefore not only old or new results, but rather both.

5.2 Tools

Besides our own tool (Recognyze), the evaluation draws upon the following three NEL systems: **AIDA** [8], **Babelfy** [14] and **Spotlight** [3]. We have selected these systems as they were the closest to Recognyze in terms of philosophy and goals.

DBpedia Spotlight is well-known within the Semantic Web and NLP communities for being one of the first tools to use DBpedia and offers semantic approaches to the named entity recognition and disambiguation problems. It was built around a vector space model and is available through a public endpoint. Due to the fact that DBpedia Spotlight is slowly becoming a general information extraction tool, we have only selected the named entities from the run, otherwise the number of wrong links would have been too high.

Babelfy was one of the first graph disambiguation tools that worked in a multilingual setting and it was built around the idea of word sense disambiguation. It offers a free web service with a limited number of requests and the option to evaluate it for research purposes. Again only the named entities were selected. While typically Babelfy offers Babelnet links, we have only selected the results that offered DBpedia / Wikipedia links.

AIDA was the first graph disambiguation tool that provided superior performance. Regardless of it being run with the local or graph disambiguation algorithms, AIDA has been one of the top performers in NEL since 2011.

To the best of our knowledge, most of the current Deep Learning tools that extract named entities do not offer DBpedia links, therefore they were not included in the current evaluations.

5.3 Results and Discussion

It can easily be seen from Table 3 that Recognyze offers the best results overall out of the selected systems, mainly due to the fact that it has the best recall. The precision results are different for each dataset, therefore Recognyze, AIDA and Spotlight each get a top precision result. This suggests that there is no one strategy to rule them all when it comes to precision. However since three of the studied systems (Recognyze, AIDA, Babelnet) use graph-based disambiguation techniques, and two of them get the best results

Table 3: Recognyze evaluation against competitors

	Reuters128			OKE2015			OKE2016		
	P	R	F1	P	R	F1	P	R	F1
Recognyze	0.487	0.515	0.500	0.633	0.543	0.585	0.602	0.414	0.491
AIDA	0.532	0.429	0.475	0.505	0.406	0.450	0.574	0.423	0.487
DBpedia Spotlight	0.504	0.485	0.494	0.610	0.362	0.455	0.638	0.343	0.446
Babelnet	0.321	0.225	0.264	0.399	0.264	0.318	0.519	0.340	0.411

overall (Recognyze, AIDA), it can easily be concluded that this type of disambiguation strategy is quite effective. The differences between tools are higher on the old datasets (Reuters128), but smaller on the more recent ones (OKE2015, OKE2016), therefore suggesting that datasets and systems can become dated or suffer various regression issues.

Several well-known issues (based on [1]) were identified: gold standard problems (e.g., wrong annotation spans such as *U.K* annotated instead of *U.K.* or missing annotation links like *Jerzy Urban*) and KB issues (e.g., some bad redirects) among them. While the graph disambiguation methods generally worked well, it seems the confusion between cities that are also region capitals and the region names still persists across most annotator tools (e.g., *N.Y.* and *N.Y. City* or *São Paulo* and *São Paulo (state)*). The Recognyze error that appeared the most during evaluations was related to shortened name variants. While the links were mostly correct, the surface form spans were not. In order to fix this we have improved our name analyzer component during the evaluations, therefore today this only happens in limited number of cases.

It has to be noted that each tool builds its Knowledge Graph in a different way, therefore besides the effective algorithms that were used, the graph construction techniques can also be considered to have an impact on the result. Babelfly uses its own KG (Babelnet), AIDA also exploits the Wikipedia texts, Spotlight leverages only DBpedia, whereas Recognyze uses a system of filters to clean the graph.

6 OUTLOOK AND CONCLUSIONS

The research presented in this paper discussed (i) the use of Linked Data as background knowledge for Named Entity Linking; (ii) knowledge mining strategies for improving the completeness, relevancy and timeliness of Linked Data obtained from theses sources; (iii) methods for mitigating data quality issues (e.g., name variances, type confusability, link stability, mappings) in the available data sources, and (iv) improvements to NEL algorithms that leverage this knowledge. Afterwards, we focused on (v) the Recognyze NEL system that implements the introduced approaches and (vi) performed a comprehensive evaluation that demonstrates their efficiency for NEL against a suite of other well-known systems.

The evaluation results not only demonstrate the impact of the discussed modifications on Recognyze's performance but also show how information extraction methods can benefit from background knowledge. Due to the success of the LD quality mitigation strategies w.r.t. the name variance issue especially, we are considering extending the list of Knowledge Bases we will support for the future. Some of the additional KBs we are currently studying include JRC-Names [5] and Google Knowledge Graph [25]. If there is a need to include further data types (e.g., products, events), we will also add domain-specific datasets that include additional information about such types, as the large KBs do not currently cover them well.

While the core of the paper is focused on the data quality mitigation strategies, some lessons related to adjustments to the current graph-disambiguation methods are also presented (e.g., filters or additional information can improve results). Due to its complexity, NEL is one of the last domains that is not yet seriously affected by the current Deep Learning craze. We, therefore, think that improvements such as the ones introduced in this paper are worthy of consideration whenever new systems are implemented.

Future work will focus on (i) advancing research on mitigation strategies and (ii) investigating means to address the problem of domain and Knowledge Base evolution and its interaction with information extraction methods. For instance, if a mention of "U.S. president" is grounded against DBpedia, the grounding depends on the chosen KB version. Identifying consistent strategies for versioning Knowledge Bases and information extraction artifacts needs to be an important cornerstone for a reliable handling of knowledge evolution and other temporal effects relevant to information extraction.

Acknowledgments

The research presented in this paper has been conducted as part of the DISCOVER Project (www.htwchur.ch/discover), funded by the Swiss Commission for Technology and Innovation (CTI). The work presented here has also been partially funded through the InVID project (http://www.invid-project.eu/) funded through the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 687786.

REFERENCES

[1] Adrian M.P. Braşoveanu, Giuseppe Rizzo, Philipp Kuntschick, Albert Weichselbraun, and Lyndon J.B. Nixon. 2018. Framing Named Entity Linking Error Types. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (7-12), Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Paris, France, 266–271. http://www.lrec-conf.org/proceedings/lrec2018/summaries/612.html

[2] Erik Cambria and Bebo White. 2014. Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine* 9, 2 (May 2014), 48–57. https://doi.org/10.1109/MCI.2014.2307227

[3] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTICS'13)*. ACM, Graz, Austria, 121–124. https://doi.org/10.1145/2506182.2506198

[4] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Inf. Process. Manage.* 51, 2 (2015), 32–49. https://doi.org/10.1016/j.ipm.2014.10.006

[5] Maud Ehrmann, Guillaume Jacquet, and Ralf Steinberger. 2017. JRC-Names: Multilingual entity name variants and titles as Linked Data. *Semantic Web* 8, 2 (2017), 283–295. https://doi.org/10.3233/SW-160228

[6] Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). The Association for Computational Linguistics, San Diego, CA, USA, 1256–1261. https://doi.org/10.18653/v1/N16-1150

[7] 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*. ACL, Baltimore, MD, USA, 464–469. https://doi.org/10.3115/v1/P14-2076

[8] 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*. ACL, Edinburgh, UK, 782–792. https://doi.org/10.3115/v1/D11-1072

[9] Heng Ji and Joel Nothman. 2016. Overview of TAC-KBP2016 Tri-lingual EDL and Its Impact on End-to-End KBP. In *Eighth Text Analysis Conference (TAC)*. NIST, Gaithersburg, Maryland, USA, Article 3, 15 pages. https://tac.nist.gov/publications/2016/additional.papers/TAC2016.KBP_Entity_Discovery_and_Linking_overview.proceedings.pdf

[10] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* 6, 2 (2015), 103–104. https://doi.org/10.3233/SW-140134

[11] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195. https://doi.org/10.3233/SW-140134

[12] Edgard Marx, Saeedeh Shekarpour, Tommaso Soru, Adrian M. P. Braşoveanu, Muhammad Saleem, Ciro Baron, Albert Weichselbraun, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, and Sören Auer. 2017. Torpedo: Improving the State-of-the-Art RDF Dataset Slicing. In *11th IEEE International Conference on Semantic Computing, ICSC 2017, San Diego, CA, USA, January 30 - February 1, 2017*. IEEE Computer Society, San Diego, CA, USA, 149–156. https://doi.org/10.1109/ICSC.2017.79

[13] Stuart E. Middleton and Vadims Krivcovs. 2016. Geoparsing and Geosemantics for Social Media: Spatiotemporal Grounding of Content Propagating Rumors to Support Trust and Veracity Analysis during Breaking News. *ACM Trans. Inf. Syst.* 34, 3 (2016), 16:1–16:26. https://doi.org/10.1145/2842604

[14] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics* 2 (2014), 231–244. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/291

[15] Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Darío Garigliotti, and Roberto Navigli. 2015. Open Knowledge Extraction Challenge. In *Semantic Web Evaluation Challenges - Second SemWebEval Challenge at ESWC 2015, Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers (Communications in Computer and Information Science)*, Fabien Gandon, Elena Cabrio, Milan Stankovic, and Antoine Zimmermann (Eds.), Vol. 548. Springer, Berlin, Germany, 3–15. https://doi.org/10.1007/978-3-319-25518-7_1

[16] Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Robert Meusel, and Heiko Paulheim. 2016. The Second Open Knowledge Extraction Challenge, See [23], 3–16. https://doi.org/10.1007/978-3-319-46565-4_1

[17] Ozer Ozdikis, Halit Oguztüzün, and Pinar Karagoz. 2017. A survey on location estimation techniques for events detected in Twitter. *Knowl. Inf. Syst.* 52, 2 (2017), 291–339. https://doi.org/10.1007/s10115-016-1007-z

[18] Julien Plu, Giuseppe Rizzo, and Raphaël Troncy. 2016. Enhancing Entity Linking by Combining NER Models, See [23], 17–32. https://doi.org/10.1007/978-3-319-46565-4_2

[19] Petar Ristoski and Heiko Paulheim. 2016. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web* 36 (jan 2016), 1–22. https://doi.org/10.1016/j.websem.2016.01.001

[20] Giuseppe Rizzo, Bianca Pereira, Andrea Varga, Marieke van Erp, and Amparo Elizabeth Cano Basave. 2017. Lessons learnt from the Named Entity rEcognition and Linking (NEEL) challenge series. *Semantic Web* 8, 5 (2017), 667–700. https://doi.org/10.3233/SW-170276

[21] Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 2014. N^3 - A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis (Eds.). ELRA, Paris, France, 3529–3533. http://www.lrec-conf.org/proceedings/lrec2014/summaries/856.html

[22] Benjamin Roth, Tassilo Barth, Michael Wiegand, Mittul Singh, and Dietrich Klakow. 2014. Effective Slot Filling Based on Shallow Distant Supervision Methods. *arXiv* 1401, 1158 [cs] (jan 2014), 0–0. http://arxiv.org/abs/1401.1158 arXiv: 1401.1158.

[23] Harald Sack, Stefan Dietze, Anna Tordai, and Christoph Lange (Eds.). 2016. *Semantic Web Challenges - Third SemWebEval Challenge at ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected Papers*. Communications in Computer and Information Science, Vol. 641. Springer, Berlin,Germany. https://doi.org/10.1007/978-3-319-46565-4

[24] Robert Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, Article 1072, 8 pages. http://www.lrec-conf.org/proceedings/lrec2012/summaries/1072.html

[25] Thomas Steiner, Ruben Verborgh, Raphaël Troncy, Joaquim Gabarró, and Rik Van de Walle. 2012. Adding Realtime Coverage to the Google Knowledge Graph. In *Proceedings of the ISWC 2012 Posters & Demonstrations Track, Boston, USA, November 11-15, 2012 (CEUR Workshop Proceedings)*, Birte Glimm and David Huynh (Eds.), Vol. 914. CEUR-WS, Aachen, Germany, Article 2, 4 pages. http://ceur-ws.org/Vol-914/paper_2.pdf

[26] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. 2014. AGDISTIS - Agnostic Disambiguation of Named Entities Using Linked Open Data. In *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014) (Frontiers in Artificial Intelligence and Applications)*, Torsten Schaub, Gerhard Friedrich, and Barry O'Sullivan (Eds.), Vol. 263. IOS Press, Amsterdam, The Netherlands, 1113–1114. https://doi.org/10.3233/978-1-61499-419-0-1113

[27] Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. 2015. GERBIL: General Entity Annotator Benchmarking Framework. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015* (May 18-22). ACM, Florence, Italy, 1133–1143. https://doi.org/10.1145/2736277.2741626

[28] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. https://doi.org/10.1145/2629489

[29] Albert Weichselbraun and Philipp Kuntschik. 2017. Mitigating linked data quality issues in knowledge-intense information extraction methods. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, WIMS 2017, Amantea, Italy, June 19-22, 2017*, Rajendra Akerkar, Alfredo Cuzzocrea, Jannong Cao, and Mohand-Said Hacid (Eds.). ACM, Amantea, Italy, 17:1–17:12. https://doi.org/10.1145/3102254.3102272

[30] Albert Weichselbraun, Daniel Streiff, and Arno Scharl. 2015. Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence. *International Journal on Artificial Intelligence Tools* 24, 2, Article 1 (2015), 24 pages. https://doi.org/10.1142/S0218213015400084

[31] Amrapali Zaveri, Dimitris Kontokostas, Mohamed A. Sherif, Lorenz Bühmann, Mohamed Morsey, Sören Auer, and Jens Lehmann. 2013. User-driven Quality Evaluation of DBpedia. In *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTICS '13)*. ACM, Graz, Austria, Article 1, 8 pages. https://doi.org/10.3233/SW-130102