Name Variants for Improving Entity Discovery and Linking

Albert Weichselbraun
Swiss Institute for Information Science, University of Applied Sciences Chur
Pulvermühlestrasse 57, 7000 Chur, Switzerland
albert.weichselbraun@htwchur.ch

Philipp Kuntschik
Swiss Institute for Information Science, University of Applied Sciences Chur
Pulvermühlestrasse 57, 7000 Chur, Switzerland
philipp.kuntschik@htwchur.ch

Adrian M. P. Braşoveanu[1]
MODUL Technology GmbH, Am Kahlenberg 1, 1190 Vienna, Austria
adrian.brasoveanu@modul.ac.at

Abstract

Identifying all names that refer to a particular set of named entities is a challenging task, as quite often we need to consider many features that include a lot of variation like abbreviations, aliases, hypocorism, multilingualism or partial matches. Each entity type can also have specific rules for name variances: people names can include titles, country and branch names are sometimes removed from organization names, while locations are often plagued by the issue of nested entities. The lack of a clear strategy for collecting, processing and computing name variants significantly lowers the recall of tasks such as Named Entity Linking and Knowledge Base Population since name variances are frequently used in all kind of textual content.

This paper proposes several strategies to address these issues. Recall can be improved by combining knowledge repositories and by computing additional variances based on algorithmic approaches. Heuristics and machine learning methods then analyze the generated name variances and mark ambiguous names to increase precision. An extensive evaluation demonstrates the effects of integrating these methods into a new Named Entity Linking framework and confirms that systematically considering name variances yields significant performance improvements.

2012 ACM Subject Classification Information systems → Incomplete data; Information systems → Inconsistent data; Information systems → Extraction, transformation and loading; Information systems → Entity resolution

Keywords and phrases Named Entity Linking, Name Variance, Machine Learning, Linked Data

Digital Object Identifier 10.4230/OASIcs.LDK.2019.14

Funding The research presented in this paper has been partially conducted as part of the DISCOVER (`https://www.htwchur.ch/discover`) and the MedMon (`https://www.htwchur.ch/medmon`) projects, funded by Innosuisse. The work presented here has also been partially supported through the InVID project (`https://www.invid-project.eu/`) funded through the European Union's Horizon 2020 Research and Innovation Programme under GA No 687786 and through the FFG project EPOCH.

[1] Corresponding author

2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 14; pp. 14:1–14:15
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

State of the art Named Entity Linking (NEL) systems [14] link mentions of named entities in textual content such as newspaper articles and tweets to the corresponding entities in Knowledge Bases (KB). Many of these systems excel at identifying entities in the canonical form presented in a Knowledge Base and some also accept variations (e.g., abbreviations, alternative names), but most systems do not necessarily take into account name variance, especially if it is not available in the target KB (e.g., DBpedia, Geonames). This limitation significantly lowers recall, since name variances such as *Joe Kennedy* rather than *Joseph Kennedy*, *IBM Research* or even only *IBM* for *IBM Zurich Research Laboratory*, and *SoCal/NoCal* for *Southern/Northern California* are frequently used, especially in less formal settings such as social media.

This article focuses on assessing the effect of name variance across domains, and introduces the following strategies for addressing this problem:

(i) *Obtain name variances by combining knowledge repositories.* Blending KBs requires aligning the entity identifiers used within them, triggering quality issues due to errors caused by the necessary ontology alignment tasks [14]. However, this issue can be avoided, if the links between KBs are exploited (e.g., by collecting name variants from multiple KBs, but linking them to the most used KB). The approach presented in this paper, therefore, uses graph mining to extract name variances and to integrate them into the target knowledge base.

(ii) *Algorithmic name variance generation* derives name variances from existing names by applying heuristics such as reducing the number of tokens (e.g. shorten *IBM Zurich Research Laboratory* to *IBM* or *IBM Zurich*), changing token alignment (*IBM Research* or *IBM Laboratory*), and substituting selected tokens with frequently used synonyms (e.g. *IBM Labs*).

(iii) *Name Analyzers* focus on boosting precision by marking ambiguous name variances. This paper discusses two name analyzer implementations: a) a heuristics entropy-based algorithm where tokens known to belong to certain entity types (e.g., prefixes or suffixes for organizations and locations, title for people, etc.) contribute higher entropy scores which are used for identifying ambiguous names; b) a machine learning implementation that uses support vector machines (SVM) and features that are inspired by the heuristic algorithm.

The first two approaches are targeted at increasing recall, whereas the third one improves precision. The reference implementation of the algorithms discussed in this paper draws upon Recognyze Lite, a graph-based NEL framework.

The rest of this paper is organized as follows: Section 2 describes the state of the art in graph disambiguation and the computation of name variance; Section 3 formalizes the generation and enrichment of named entity graphs for graph disambiguation and presents the architecture used to implement the suggested name variance strategies. Section 4 presents a comprehensive evaluation of the impact of name variance on NEL and discusses these results. The paper concludes with Section 5 which provides an overview of the presented and future work.

2 Related Work

The state-of-the-art and open issues in NEL are described in the overview of the TAC-KBP tasks each year [14]. Depending on the task and features that are used (e.g., strong or weak typing and/or linking, classification or clustering evaluation, etc.), NEL tasks can be defined and evaluated in multiple ways as explained in [29], [10] or [14]. The most general situation is called NERLC (Named Entity Recognition Linking and Classification) and involves detecting not just the entities (NER), but also the links (NEL) and associated types (NEC) [14].

Knowledge Graph (KG) disambiguation is currently considered among the most effective approaches towards NEL. Several graph disambiguation NEL tools have been listed among the top performers in NLP competitions (e.g., TAC-KBP [14], OKE [19]): AIDA [11], HITS [9], Babelfy [17], AGDISTIS [29] or the multilingual version of AGDISTIS called MAG [18]. Competing approaches include statistical disambiguation (e.g., ADEL [21] or DBpedia Spotlight [4]) and neural models (e.g., Ensemble Nerd [3] for NEL).

Almost all the NEL systems have to provide at least a basic algorithm (or alternatively a set of features) for addressing the name variance problem. Some of the recently applied methods include: query expansion [8], mention-entity similarity based on keyphrases or syntax and entity-entity coherence (Milne-Witten) in AIDA [11], maximum entropy (ME) [22], synset expansion in Sematch [32], string matching via Levenshtein distances [13], Knowledge Base Embeddings [28], and ensemble neural networks [3]. Several systems that use hybrid approaches have also been developed. The HITS system [9] uses a heuristic that includes a rule-based approach for abbreviations, considers Wikipedia redirects for most common aliases, and calls to Wikipedia search functions for less common name variants. The LIEL system [26] uses language independent features like mention-entity pair features (text-based, KB link properties, Wikipedia page titles, etc.) and entity-entity pair features (overlap, title co-occurence, etc). All of these approaches struggle with missing abbreviations, names that originate in other languages, partial matches, etc. Maximum entropy [22], has been applied in Named Entity Recognition (NER) setups, therefore improvements on top of it might be needed for NEL. Popularity prior [11] is not a good metric for new entities. Synset expansion [32] can in theory help match almost all the name variance cases provided they are covered by existing KBs which rarely happens in practice. Knowledge Base embeddings [28] are dependent upon KB data quality.

Mining for name variants by combining modern KBs helps improving the coverage of entities and their name variants, but a single KB rarely provides all the information we need. DBpedia [16] does not contain special fields for name variants, but they can be collected from different fields (e.g., *dbp:wikiPageDisambiguates*, *dbo:wikiPageRedirects*, *dbp:acronym*, etc). Wikidata [6] has less factual triples for each entity than DBpedia since it has been curated manually, but it provides more triples and many name variants for each entity (through the "also known as" field). Wikidata is ideal for identifying named entities, whereas DBpedia excels at obtaining additional information about a particular entity. JRC-Names [5] is a multilingual KB that provides lists of entities and their name variants. It focuses mostly on spelling variations and covers persons and organizations, but currently does not contain any triples for locations. Geographical KBs (e.g., LinkedGeoData [27]) can also be considered good sources of name variants, provided the users are only interested in locations and are willing to combine the names from multiple fields and languages. Improving the coverage of entities and their name variances is a good technique for improving NEL, but when the entities or their name variants are missing from KB it might be best to use the entire Internet as background knowledge as described in [1].

It has to be noted that the problem of name variances is not limited to NEL or Knowledge Base Population (KBP) systems, but rather is also relevant to any field that requires matching records or names such as ontology alignment, word sense disambiguation, data linkage or slot filling tasks [12].

3 Method

This section describes Recognyze Lite, a new NEL framework that focuses on increasing recall through the use of name variance while mitigating its impact on precision. It provides a formalization of the graph generation and enrichment problem covering the tasks of adding name variances to the knowledge graph and using name analyzers for marking ambiguous name variances. Recognyze Lite provides a flexible, multi-KB NEL system that, among others, utilizes relations between entities from any given linked data source to disambiguate between correct and false candidate mentions in an unknown text.

3.1 Graph disambiguation

Similar to Usbek et al. [29] we define our approach as follows: Given a knowledge base K as a directed graph $G = (V, E)$ with vertices V and edges E. Recognyze Lite uses SPARQL queries to obtain a sub-graph $G' = (V', E')$ with the following properties:

1. $s \in V'$ and $o \in V'$, where s refers to a resource and o either indicates a resource or a literal (i.e. in this case a name used to identify a named entity)
2. for every pair $(s, o) \in E \Rightarrow \exists p : (s, p, o)$ which is denoted to as an RDF triple in G'.

The named entity disambiguation process comprises multiple sub-tasks: (i) Directed Acyclic Word Graphs (DAWGs) [25] provide fast text search within the input documents to identify candidate entities by locating mentions of their name variances. (ii) A controlled vocabulary is applied to search for potential affixes that hint on relevant entity types. (iii) These affixes are then used to remove candidate mentions that do not match the type implied by the affix. (iv) The remaining candidate entities are then linked using multiple disambiguation algorithms in sequence. In this sub-task, the relations between the candidate mentions, as well as the significance of a single mention are used to determine the best fitting network of entities. (v) Finally, Recognyze Lite transforms the accepted entities into the desired output format.

3.2 Name variance

Name variance is the problem of finding all the different names that represent a single entity within a collection of text. In theory, enriching G' with name variances improves recall, whereas adding name variance related features to the NEL extraction pipelines improves precision.

Several cases of variance have been described in the literature (e.g., [5] or [14]): (i) known aliases (*Robert Gailbraith*, a psudonym used by *J.K. Rowling*; *John Barron* for *Donald Trump*, *Mahatma Gandhi* for *Mohandas Karamchand Gandhi*) ; (ii) hypocorisms or common aliases (*Bobby* for Robert, *Liz* for Elizabeth); (iii) abbreviations (*JFK* for both *John F. Kennedy* and *John F. Kennedy International Airpot*); (iv) 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); (v) partial matches (names of royal figures often fall under this category; 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 can often include titles (*Senator*, *Judge*, 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 to simply as *Sony*); combining parts of words (e.g., *Nortel* instead of *Northern Telecom*). Locations have more problems with

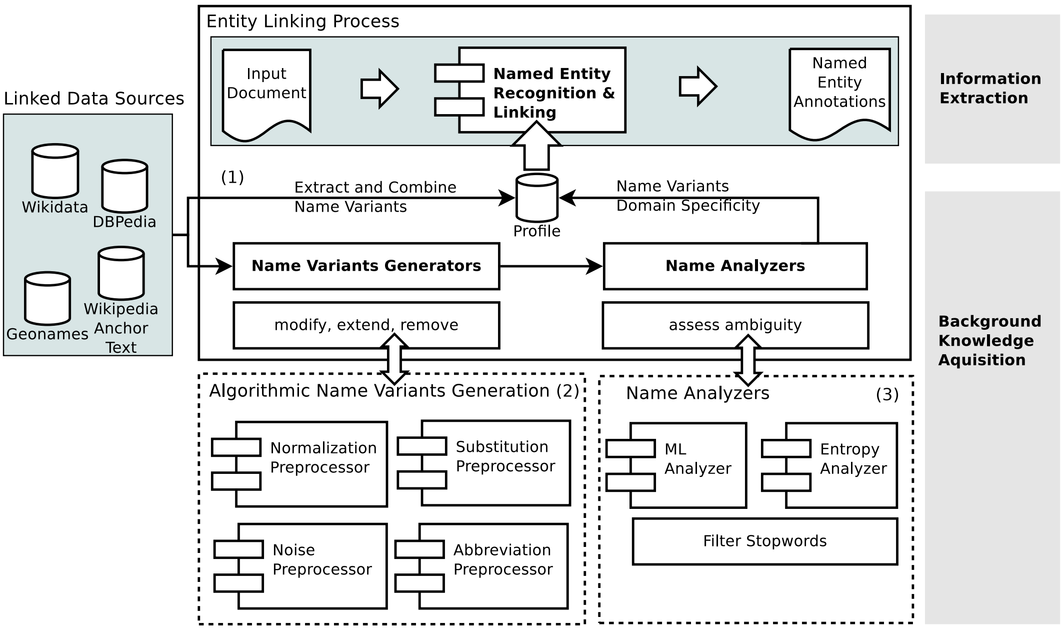


Figure 1 Name variance handling in Recognyze Lite: (1) combine name variants from multiple datasets; (2) algorithmic name variants generation; (3) name analyzers (entropy heuristic or machine learning (ML) based).

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*); regional abbreviations (e.g., *OH* for *Ohio*); embeddings or nested entities (e.g., *New York Stadium*); possessive names (e.g., *Hawaii's Waikiki*); and addresses (e.g., *221B Baker Street*).

If we take entity typing (e.g., Person – PER, organization – ORG, location – GEO, etc) into consideration, the variance problem can also include issues related to hyponyms and hypernyms [15] or even meronyms [7].

Recognyze Lite addresses the name variance problem in two ways: (i) by combining name variants from multiple datasets and (ii) by algorithmically deriving name variants from an entity's official names.

Name variances and the corresponding named entities are stored in a binary profile which is build from the knowledge base used for grounding entities. Recognyze Lite constructs knowledge graphs for NEL based on SPARQL queries that select relevant entity graphs and may comprise multiple knowledge bases (Section 3.3.1) such as DBpedia, Wikidata and GeoNames. A comprehensive preprocessing pipeline allows the analysis, manipulation and addition of name variances (Section 3.3.2), and the identification of name variances that would be harmful to the system's performance (Section 3.4).

3.3 Name variance for improving recall

3.3.1 Name variance through additional knowledge bases

The first approach for enriching the original graph draws upon further knowledge bases K_i and the corresponding graphs (V_i, E_i) to obtain tuples (s, p_j, o_k) where s is a resource in the knowledge graph G' $(s \in V')$ that is also available in knowledge base K_i $(s \in V_i)$. Adding edges $(s, o_k) \in E_i$ with relevant property types $p_j = \{p_1, ...p_n\}$ and the corresponding name variance $o_k \in V_i$ into G' enriches G' with these additional name variances.

Since such an approach might use SPARQL federation or similar technologies (e.g., RDF slicing), it is important to assess its impact on scalability before deploying it into large production systems.

3.3.2 Name variance through algorithmic name generation and assessment

The second method draws upon an algorithm $\mathcal{A}$ that splits a literal $o_k \in V'$ from the RDF triple (s, p, o_k) into tokens $t_i = \{t_1, ...t_n\}$ that are then used to generate name variances $o_k^1...o_k^m$ and the corresponding RDF triples $(s, p, o_k^1), ..., (s, p, o_k^m)$ to be later integrated in the knowledge graph G'.

A simple variance of $\mathcal{A}$ obtains $(n-1)$ name variances by providing substrings $o_k^1 = t_1$, $o_k^2 = t_1 t_2, ..., o_k^{n-1} = t_1 t_2 ... t_{n-1}$ of the original name. The more advanced algorithm $\mathcal{A}'$ also (i) considers synonyms by generating name variances that replace tokens t_i with synonyms $t_i^1, t_i^2, ...t_i^m$, and (ii) uses heuristics encoded in regular expressions to create name variances by modifying and reordering tokens t_i. Applying $\mathcal{A}'$ to the name "United States Department of State", for example, yields the additional name variances "U.S. Department of State" and "US Department of State". The pattern `{Department of (\w+)/\$1 Department}`, for instance, generates the name variance "Commerce Department" from the initial name "Department of Commerce". Since in many cases the abbreviations are not necessarily available in the KBs, a dedicated component is used for extracting such abbreviations directly from text such as DBpedia abstracts, if they are available.

Some preprocessing steps that are typically applied include the following: i) noise - removal of dashes, white spaces, parentheses, etc.; ii) abbreviation - for extracting abbreviations from abstracts or long texts; iii) normalization - for normalizing the entity names; iv) tickers - for detecting the company stock ticker symbols; or v) URL - removal of URLs.

3.4 Mitigating name variance's impact on precision

Name variance per se tends to improve recall at the cost of precision. We, therefore, introduce *name analyzers*, i.e. components that identify name variances which might be particularly harmful to precision.

Name analyzers aim to balance the improved recall with precision by marking ambiguous name variances, i.e. names that

1. have a high probability of clashing with common terms (e.g. *Reading*, *Turkey*, etc.) and/or
2. may clash with terms from other entity classes (e.g. *Carolina/LOC* versus *Carolina/PER*).

More formally, a name analyzer for an entity type T is considered a function $\mathcal{N}_T : o_i \rightarrow b$ that provides a mapping of name variances o_i to a binary value b indicating whether the name is considered ambiguous or not. The disambiguation process uses this information and may, for instance, require additional evidence prior to the grounding of ambiguous name variances.

Since the evaluations discussed in Section 4 are focused on news articles, we assess name variances for PER with a simple heuristic that requires at least one common English first- or surname to be present within a candidate name. For GEO we employ a simple dictionary-based list that removes names that clash with standard vocabulary.

The most challenging entity type in terms of assessing name variances are organizations for which Recognyze Lite uses an entropy-based name analyzer, as well as a machine learning approach.

The next subsections introduce these two name analyzer implementations.

3.4.1 Entropy-based name analyzer

The entropy-based name analyzer has been inspired by research from [31] and computes a heuristic entropy score that is used for assessing whether a generated name variances is considered ambiguous or not.

In information theory the entropy H specifies the minimum number of bits needed to encode sequences of random variables X produced by a probability distribution p. High entropy values, therefore, also correspond to a high diversity of values $x_i \in X$ obtained from p.

The entropy-score heuristic presented in this paper draws upon these concepts by assessing the degrees of freedom in creating valid organization names from the computed name variances (i.e. answers the question of how many *valid* organization names can be created from the available tokens). A high entropy score indicates that the name variance is very likely unambiguous, a low score, in contrast, refers to ambiguous name variances.

Tokens that are known to be used in organization names, contribute a higher entropy $H_{\text{token}}(t_j)$ (e.g. *Inc.*, *Plc.*, *AG* etc.) than tokens that are not specific to company or organization names. The heuristic also considers the number of token classes H_{classes} (i.e. abbreviation, name, legal form, etc.) used in the name variance. We compute the entropy of a name variance $\{t_i\}$ that comprises n tokens $\{t_1, t_2, ...t_n\}$ as follows:

$$H(\{t_i\}) \quad = \quad f_{\text{constr}}(\{t_i\}) \cdot \Big[H_{\text{case}}(\{t_i\}) + H_{\text{classes}}(\{t_i\}) + \sum_{t_j \in \{t_i\}} H_{\text{token}}(t_j) \Big] \tag{1}$$

The initial entropy H_{case} discounts case insensitive name variance, and the factor f_{constr} eliminates name variances that violate syntactic rules.

$$H_{\text{case}}(\{t_i\}) \quad = \quad \begin{cases} 0.0 & \text{if caseSens}(\{t_i\}) \\ -0.5 & \textit{else.} \end{cases} \tag{2}$$

$$f_{\text{constr}}(\{t_i\}) \quad = \quad \begin{cases} 0.0 & \text{if } \neg\text{constr}(\{t_i\}) \\ 1.0 & \textit{else.} \end{cases} \tag{3}$$

These constraints enforce that name variants (i) contain at least two characters and (ii) do not end with a connector or possessive form. This rule prevents broken names such as "Zingg &" or "Society of".

The obtained entropy measure ensures that names are unique enough to prevent ambiguities with common terminology and phrases specific to the text's language. A comprehensive corpus of ambiguous and unambiguous name variances has been used to experimentally determine suitable values for $H_{\text{case}}(\{t_i\})$, $H_{\text{classes}}(\{t_i\})$ and $H_{\text{token}}(t_j)$, to fine tune the heuristics for generating the entropy scores, and to determine the optimal threshold below which name variances should be considered ambiguous.

3.4.2 Machine learning name analyzer

We use the Java implementation of libSVM[2] to create a name analyzer that draws upon machine learning rather than heuristics for classifying name variants into ambiguous and unambiguous ones. The machine learning component considers a total of 81 features such as

[2] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

morphological features (whether tokens are case sensitive, capitalized, all uppercase, contain letters, punctuation, etc.), syntactical features (pronouns, prepositions etc.) and semantic features (number of words mentions that refer to popular fist names, given names, trades, locations, common dictionary terms in English, French or German, etc.). Since dictionaries often also contain popular company names, a preprocessing step removes abbreviations (e.g. BBC, CNN, etc.) and the names of Forbes 2000 companies to improve their usefulness for distinguishing between common terms and potential company names.

The language-specific training corpus has been composed of (i) manually curated language-specific lists of Fortune 1000 companies, and the largest Austrian, German and Swiss companies that have been retrieved from Wikipedia, and (ii) additional 539 gold-standard entries that have been automatically derived from unit test cases used in the development of the name analyzer heuristic. A cross-validation and grid-search procedure yielded the best results for a radial basis function kernel with C=8 and $\gamma=2^{-5}$.

4 Experiments

The following section elaborates on datasets and tools used for the evaluations, the chosen evaluation settings and the evaluation results.

4.1 Datasets and Evaluation Tools

Evaluations were performed with the Orbis scorer [20], because GERBIL [30] and the neleval scorer [10] do not provide means for visually debugging results. The evaluation datasets have been selected based on the following criteria: (i) they should be available in the format, and (ii) (where possible) have been use in recent evaluation tools or challenges such as GERBIL [30] and TAC-KBP [14]. We have used two datasets included in GERBIL: N3 Reuters128 (news, multiple domains) [23] and OKE2015 (abstracts, biographies) [19].

Evaluations were performed on four state-of-the-art NEL systems which also provide REST endpoints that allow the use of sophisticated evaluation frameworks such as GERBIL and Orbis: DBpedia Spotlight [4], Babelfy [17], AIDA [11], and Recognyze Lite.

While we have tested different builds of the Knowledge Bases, the experiments described in this section used DBpedia 2015-10, Wikidata 2016-08-01 and GeoNames 2016-02-26, we preferred to use an older DBpedia version (2015-10) for the Reuters128 evaluation presented in Table 1, since the data set itself was not updated since 2014 (one year before the respective DBpedia version). This version or the one from 2014 are closer to the date when the data set was created, therefore ensuring that we are not delivering any entities that were marked as NIL (or not linked to the target KB) in the original data set, since they were not available in DBpedia at that time.

Roth et al. [24] use Wikipedia link anchor text such as *UNBRO* to expand queries for the corresponding entity (in this case *United Nations Border Relief Operation*). We apply this approach to extract additional name variances from the Wikipedia 2017-12-01 dump but only consider unambiguous link anchor text. The extracted name variances yield the Wikipedia dataset[3] used in the evaluations.

Since entity spans are to some extend dependent on a gold standard's annotation policy, we use Orbis' mention-based evaluation setting where a mention is considered correct if it (i) is found within a span that overlaps the gold standard, and (ii) refers to the same named entity as the overlapping gold standard annotation. For the gold standard sentence

[3] Available at `https://github.com/AlbertWeichselbraun/wikipedia-link-extractor`.

1. *“[Avco Corporation] has increased its profits by 10% in 2017.”* where *[Avco Corporation]* refers to `dbr:Avco` both the mention *[Avco]* and *[Avco Corporation]* would be considered correct, if they refer to `dbr:Avco`.
2. The same is true for the overlapping mention *[the Netherlands]* from the sentence *“... the [Netherlands] planted a record...”* if it refers to `dbr:Netherlands`.

4.2 Evaluation Settings

The first set of evaluations demonstrates the impact of different name variance settings on the NEL performance. The *baseline* setting does not consider any name variance, operates on DBpedia only and solely uses the *rdfs:label* field for generating entity names. Setting (a) is still limited to DBpedia but considers additional DBpedia properties such as *foaf:name* and *dbp:name*. The (b1-b4) settings, draw upon multiple KBs with the intention to improve recall.

Nevertheless, the results for both the (a) and the (b1-b4) settings (Table 1) indicate that just adding additional data fields and KBs without any evaluation of name variances might even be counter productive.

Setting (c) builds upon the baseline by adding algorithmic name generation which yields considerable improvements in terms of recall at the cost of precision. The (d1-d4) settings apply algorithmic name generation to the additional KB only. The (e1-e2) configurations extend the baseline by introducing name analyzers although they are not that effective without additional name variances and, therefore, only yield significant F1 improvements for the PER type. The best performing setting (f) combines the baseline with additional properties, algorithmic name generation and Wikidata as a supplemental KB for which algorithmic name generation has been enabled as well. The heuristic name analyzer ensures a good balance between precision and recall.

Table 1 summarizes the evaluation results. We have used the R implementation of the Wilcoxon rank sum test to verify whether a particular setting yields a significant improvement at the p=0.05 significance level. Bold values indicate significant improvements, all other values are either non-significant or losses.

The second evaluation serves to illustrate that considering name variance yields competitive results. Table 2, therefore, compares Recognyze Lite’s performance to three popular NEL services that offer publicly available APIs [4]. AIDA, Babelfy and Recognyze Lite use KG disambiguation techniques, while Spotlight uses statistical disambiguation. It has to be noted that each service builds its entity graph differently, therefore, not only the NEL algorithms, but also the differences between KGs can lead to variation in the results. AIDA is based on Wikipedia and, therefore, operates on a substantially different KG than the other tools. Babelfy uses the Babelnet KG and provides DBpedia links via the *owl:sameAs* property. Spotlight and Recognyze Lite both draw upon DBpedia, although Spotlight is fine-tuned for knowledge extraction tasks, whereas Recognyze Lite is optimized for NEL and various domain specific extraction tasks (e.g., Slot Filling for the recognized entities).

The Recognyze Lite baseline (Table 1) which does not consider name variance yields results that are on par with the other top systems in Table 2. Once the name variance strategies proposed in this paper are activated, the resulting system clearly outperforms all other approaches, as outlined in Table 2.

[4] Since no recommended settings for performing evaluations on Reuters128 and OKE2015 datasets have been published, we have dedicated approximately two days to experimental optimization of the evaluation settings of all evaluated third-party tools.

Table 1 Impact of name variance on the Recognyze Lite Named Entity Linking performance for the Reuter128 dataset. Bold figures indicate statistically significant improvements over the baseline.

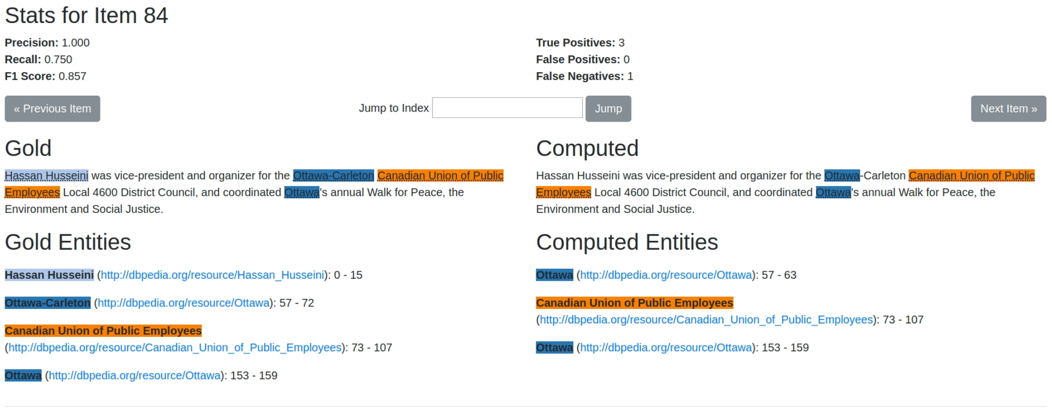
	Setting	LOC			ORG			PER			All		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
	baseline	63	54	58	72	34	46	57	23	33	66	39	49
(a)	additional properties	63	54	58	71	33	45	57	23	33	66	38	49
(b1)	Wikidata	14	41	20	40	**41**	40	12	**38**	19	21	41	28
(b2)	Wikipedia	61	54	57	69	33	45	58	25	35	64	39	48
(b3)	GeoNames	60	54	57	71	33	45	57	23	33	64	38	48
(b4)	baseline + (b1 + b2 + b3)	14	41	21	39	**41**	40	12	**38**	19	21	41	28
(c)	algorithmic name generation	54	**72**	62	35	**53**	42	68	**49**	**57**	43	**58**	50
(d1)	name generation on Wikidata	52	54	53	71	38	50	59	26	36	61	42	50
(d2)	name generation on Wikipedia	58	52	55	68	35	46	60	29	39	63	39	48
(d3)	name generation on GeoNames	48	53	51	70	33	45	57	23	33	58	38	46
(d4)	baseline + (d1 + d2 + d3)	46	53	50	70	38	50	61	30	40	58	42	49
(e1)	name analyzer (heuristic)	64	52	57	47	**44**	46	60	**56**	**58**	54	**48**	51
(e2)	name analyzer (machine learning)	65	51	57	33	**47**	39	55	**47**	**50**	42	**48**	45
(f)	baseline + (a, c, d1, e1)	53	**70**	61	61	**52**	**57**	60	**56**	**58**	58	**58**	**58**

4.3 Discussion

Many of the settings included in Table 1 shed light on pitfalls relevant to name variance for NEL. When we designed Recognyze Lite, we proceeded incrementally, therefore expecting better results for each setting. This has not always been the case. For instance, the setting (b1) *baseline+wikidata* yields considerably worse results than the baseline profile. Initially we suspected that this effect might have been caused by data quality issues within Wikidata which is considered a relatively novel data source [6]. An analysis of the issue uncovered that the quality of Wikidata is actually high and that it yields lot of name variants per entity. This in itself is a problem as (i) gold standards usually consider a limited number of name variants for each entity, and (ii) they rarely take into account partial matches [2].

Table 2 Comparison of the system performance on the Reuters 128 and OKE2015 corpora.

Corpus	System	LOC			ORG			PER			All		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Reuters 128	AIDA	44	64	52	76	29	42	50	49	50	53	43	47
	BabelNet	29	31	30	47	16	24	21	29	24	32	22	26
	Recognyze	**53**	**70**	**61**	61	52	57	60	56	58	58	58	58
	Spotlight	41	**70**	52	64	42	51	47	22	30	50	49	49
OKE 2015	AIDA	25	37	30	69	43	53	66	41	50	50	41	45
	BabelNet	21	35	26	67	40	50	55	14	22	40	26	32
	Recognyze	**62**	**73**	**67**	70	51	59	85	57	68	73	59	65
	Spotlight	50	72	59	81	50	62	56	11	18	61	36	45



Stats for Item 84

Precision: 1.000
Recall: 0.750
F1 Score: 0.857

True Positives: 3
False Positives: 0
False Negatives: 1

Gold

Hassan Husseini was vice-president and organizer for the Ottawa-Carleton Canadian Union of Public Employees Local 4600 District Council, and coordinated Ottawa's annual Walk for Peace, the Environment and Social Justice.

Computed

Hassan Husseini was vice-president and organizer for the Ottawa-Carleton Canadian Union of Public Employees Local 4600 District Council, and coordinated Ottawa's annual Walk for Peace, the Environment and Social Justice.

Gold Entities

Hassan Husseini (http://dbpedia.org/resource/Hassan_Husseini): 0 - 15

Ottawa-Carleton (http://dbpedia.org/resource/Ottawa): 57 - 72

Canadian Union of Public Employees (http://dbpedia.org/resource/Canadian_Union_of_Public_Employees): 73 - 107

Ottawa (http://dbpedia.org/resource/Ottawa): 153 - 159

Computed Entities

Ottawa (http://dbpedia.org/resource/Ottawa): 57 - 63

Canadian Union of Public Employees (http://dbpedia.org/resource/Canadian_Union_of_Public_Employees): 73 - 107

Ottawa (http://dbpedia.org/resource/Ottawa): 153 - 159

Figure 2 Debugging name variance with Orbis.

By far the most common problem was related to ambiguous name variances introduced by string splitting. Longer strings were often split into multiple entities (e.g., *Canadian Bashaw Leduc Oil and Gas Ltd* was split into *Canadian*, *Bashaw* and *Leduc*). This might not be an issue if the entity is a Person and some of the splits indicate actual roles, but if each token references a different entity (e.g. *West German Finance Minister Gerhard Stoltenberg* includes links to such ambiguous entities like `dbr:West,_Texas`, `dbr:German,_New_York` and `dbr:Minister_(Catholic_Church)`) or if there are any containment issues (e.g. *Texas Gulf Coast* is a part of *Texas*), this name variance generation strategy yields results that are similar to negative compounding. This observation triggered our research in Name Analyzer heuristics and machine learning algorithms which addresses this problem. When used in combination, both the algorithmic name generation and name analyzer components perform considerably better than the baseline+wikidata precisely because they delivered less ambiguous name variants.

DBpedia typing in itself can sometimes lead to issues, as often general terms like *stream* or *lake* might be tagged with the associated entity types, even though they are not entities. Another troubling case observed is the lack of a clear convention for embedded names (e.g., *Wells Fargo Alarm Services* embeds the name of geographical entity), geographical containment (e.g., *Texas Gulf Coast* is a part of *Texas*) or inclusion of titles in the name of entities (e.g., *chairman John Sandner* vs *John Sandner*). These problems have been especially relevant to the Recognyze Lite Wikidata and name generation evaluations (d1) presented in Table 1.

The comparison presented in Table 2 aims at providing insights into the competitiveness of the discussed name variance methods and an an assessment of whether other NEL systems could benefit from it as well. Each tool has committed a different set of errors, although the issue of ambiguous name variances due to the splitting of longer names was noticed in all tools to some degree. Most of the systems (e.g., AIDA, Babelnet) also failed to correctly identify all the name variants that belong to an entity (e.g., *Avco Financial Services*, *Avco Financial* or *Avco* can refer to the same entity). In addition, they either do not take into account abbreviations or they rarely get them correctly. In some cases, prefixes (e.g., country abbreviations – *U.S.*, *U.K.*) and suffixes (e.g., terminations like *and Co.*, *Ind.* or *GmbH*) have also created problems. Based on our analysis at least name analyzers and techniques for abbreviations would be beneficial for improving the performance of all analyzed systems.

It has to be noted that in some cases there might not be a correct way to annotate a certain entity as illustrated in Figure 2. In this example from the OKE2015 data set, the text *Ottawa-Carleton Canadian Union of Public Employees* can be annotated as (i) *Ottawa-Carleton*, (ii) *Canadian Union of Public Employees*, (iii) *Ottawa-Carleton Canadian Union of Public Employees*, or (iv) quite possibly with an even more expanded annotation that also includes *Local 4600 District Council*. Similarly it can be argued that *Ottawa's annual Walk for Peace* should be an annotation that identifies a single recurring event. Since the results also depend a lot on the annotation guidelines of each data set, we can argue that these annotation guidelines should be openly accessible in a machine readable format (e.g., NIF, Turtle) in order to standardize evaluations and provide better comparisons between tools. Nevertheless, it needs to be noted that name variance techniques will probably not always be sufficient to address these kinds of errors, since often assigning all name variants to the correct entities is also a coreference and clustering issue.

5 Outlook and Conclusion

Considering name variances in NEL tasks significantly improves system performance. The research presented in this paper introduced three strategies for generating name variances from linked data: (i) combining knowledge repositories, (ii) algorithmic name variance generation, and (iii) name analyzers for identifying ambiguous name variances. As outlined and discussed in Section 4 these three strategies need to be deployed in concert to be effective. The use of multiple knowledge repositories or algorithmic name variance on their own does not yield significant improvements since higher recall is usually offset by lower precision or by negative effects on other entity types. Rigorous evaluations and drill-down analyses allowed understanding these issues which in turn paved the way for the development of the entropy-based name analyzer and the machine learning based name analyzer presented in this paper. These name analyzers identify and handle ambiguous name variances, substantially improving system performance. Since name variance and name analyzers can be deployed on top of existing NEL systems, the presented approach can be considered a blueprint for considerably improving the accuracy of such systems.

Future work will focus on (i) developing additional methods for identifying name variances based on deep learning, (ii) studying the effect of co-reference and clustering issues related to name variance, and (iii) better leveraging the potential of ambiguous name variances which is particularly challenging since these name variances have a high likelihood of reducing precision due to collisions with terminology used in the text that does not refer to a named entity.

References

1 Gabor Angeli, Victor Zhong, Danqi Chen, Arun Tejasvi Chaganty, Jason Bolton, Melvin Jose Johnson Premkumar, Panupong Pasupat, Sonal Gupta, and Christopher D. Manning. Bootstrapped Self Training for Knowledge Base Population. In *Proceedings of the 2015 Text Analysis Conference, TAC 2015, Gaithersburg, Maryland, USA, November 16-17, 2015, 2015*. NIST, 2015. URL: `https://tac.nist.gov/publications/2015/participant.papers/TAC2015.Stanford.proceedings.pdf`.

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

Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2018/summaries/612.html.

3 Lorenzo Canale, Pasquale Lisena, and Raphaël Troncy. A Novel Ensemble Method for Named Entity Recognition and Disambiguation Based on Neural Network. In Denny Vrandecic, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I*, volume 11136 of *Lecture Notes in Computer Science*, pages 91–107. Springer, 2018. doi:10.1007/978-3-030-00671-6_6.

4 Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 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, 2013. doi:10.1145/2506182.2506198.

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

6 Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandecic. Introducing Wikidata to the Linked Data Web. 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 50–65. Springer, 2014. doi:10.1007/978-3-319-11964-9_4.

7 Roxana Girju, Adriana Badulescu, and Dan I. Moldovan. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics, 2003. URL: http://aclweb.org/anthology/N/N03/N03-1011.pdf.

8 Swapna Gottipati and Jing Jiang. Linking Entities to a Knowledge Base with Query Expansion. 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 804–813, 2011. URL: http://www.aclweb.org/anthology/D11-1074.

9 Yuhang Guo, Wanxiang Che, Ting Liu, and Sheng Li. A Graph-based Method for Entity Linking. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*, pages 1010–1018. The Association for Computer Linguistics, 2011. URL: http://aclweb.org/anthology/I/I11/I11-1113.pdf.

10 Ben Hachey, Joel Nothman, and Will Radford. 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, 2014. URL: http://aclweb.org/anthology/P/P14/P14-2076.pdf.

11 Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 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, 2011. URL: http://www.aclweb.org/anthology/D11-1072.

12 Lifu Huang, Avirup Sil, Heng Ji, and Radu Florian. Improving Slot Filling Performance with Attentive Neural Networks on Dependency Structures. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2588–2597. Association for Computational Linguistics, 2017. URL: https://aclanthology.info/papers/D17-1274/d17-1274.

13 Filip Ilievski, Giuseppe Rizzo, Marieke van Erp, Julien Plu, and Raphaël Troncy. Context-enhanced Adaptive Entity Linking. 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), 2016. URL: `http://www.lrec-conf.org/proceedings/lrec2016/summaries/852.html`.

14 Heng Ji and Joel Nothman. Overview of TAC-KBP2016 Tri-lingual EDL and Its Impact on End-to-End KBP. In *Eighth Text Analysis Conference (TAC)*. NIST, 2016.

15 Tomás Kliegr. Linked hypernyms: Enriching DBpedia with Targeted Hypernym Discovery. *J. Web Sem.*, 31:59–69, 2015. `doi:10.1016/j.websem.2014.11.001`.

16 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. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015. `doi:10.3233/SW-140134`.

17 Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244, 2014. URL: `https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/291`.

18 Diego Moussallem, Ricardo Usbeck, Michael Röder, and Axel-Cyrille Ngonga Ngomo. MAG: A multilingual, knowledge-based agnostic and deterministic entity linking approach. *CoRR*, abs/1707.05288, 2017. `arXiv:1707.05288`.

19 Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Darío Garigliotti, and Roberto Navigli. 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*, volume 548 of *Communications in Computer and Information Science*, pages 3–15. Springer, 2015. `doi:10.1007/978-3-319-25518-7_1`.

20 Fabian Odoni, Philipp Kuntschik, Adrian M. P. Braşoveanu, and Albert Weichselbraun. On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance. In Anna Fensel, Victor de Boer, Tassilo Pellegrini, Elmar Kiesling, Bernhard Haslhofer, Laura Hollink, and Alexander Schindler, editors, *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, 2018. `doi:10.1016/j.procs.2018.09.004`.

21 Julien Plu, Giuseppe Rizzo, and Raphaël Troncy. Enhancing Entity Linking by Combining NER Models. In *Semantic Web Challenges - Third SemWebEval Challenge at ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected Papers*, volume 641 of *Communications in Computer and Information Science*, pages 17–32. Springer, 2016. `doi:10.1007/978-3-319-46565-4_2`.

22 Livy Real and Alexandre Rademaker. HAREM and Klue: how to compare two tagsets for named entities. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 43, 2015.

23 Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 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*, pages 3529–3533, 2014. URL: `http://www.lrec-conf.org/proceedings/lrec2014/summaries/856.html`.

24 Benjamin Roth, Tassilo Barth, Michael Wiegand, Mittul Singh, and Dietrich Klakow. Effective Slot Filling Based on Shallow Distant Supervision Methods. *arXiv:1401.1158 [cs]*, January 2014. `arXiv:1401.1158`.

25 Arno Scharl, Albert Weichselbraun, Max C. Göbel, Walter Rafelsberger, and Ruslan Kamolov. Scalable Knowledge Extraction and Visualization for Web Intelligence. In *49th Hawaii International Conference on System Sciences, HICSS 2016, Koloa, HI, USA, January 5-8, 2016*, pages 3749–3757. IEEE Computer Society, 2016. `doi:10.1109/HICSS.2016.467`.

26 Avirup Sil and Radu Florian. One for All: Towards Language Independent Named Entity Linking. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. URL: `http://aclweb.org/anthology/P/P16/P16-1213.pdf`.

27 Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. LinkedGeoData: A core for a web of spatial open data. *Semantic Web*, 3(4):333–354, 2012. `doi:10.3233/SW-2011-0052`.

28 Zequn Sun, Wei Hu, and Chengkai Li. Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I*, volume 10587 of *Lecture Notes in Computer Science*, pages 628–644. Springer, 2017. `doi:10.1007/978-3-319-68288-4_37`.

29 Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. 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)*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 1113–1114. IOS Press, 2014. `doi:10.3233/978-1-61499-419-0-1113`.

30 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. GERBIL: General entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015*, pages 1133–1143, 2015. `doi:10.1145/2736277.2741626`.

31 Albert Weichselbraun, Philipp Kuntschik, and Adrian M. P. Braşoveanu. Mining and Leveraging Background Knowledge for Improving Named Entity Linking. In *Proceedins of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS 2018)*, Novi Sad, Serbia, 2018. ACM. `doi:10.1145/3227609.3227670`.

32 Ganggao Zhu and Carlos Angel Iglesias. Sematch: Semantic Entity Search from Knowledge Graph. In *Joint Proceedings of the 1st International Workshop on Summarizing and Presenting Entities and Ontologies and the 3rd International Workshop on Human Semantic Web Interfaces (SumPre 2015, HSWI 2015) co-located with the 12th Extended Semantic Web Conference (ESWC 2015), Portoroz, Slovenia, June 1, 2015.*, volume 1556 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015. URL: `http://ceur-ws.org/Vol-1556/paper2.pdf`.