Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2129–2137
Marseille, 11–16 May 2020

Towards Entity Spaces

Marieke van Erp[*], **Paul Groth**[†]
[*]KNAW Humanities Cluster - DHLab, Amsterdam, NL
[†]University of Amsterdam, Amsterdam, NL
marieke.van.erp@dh.huc.knaw.nl, p.groth@uva.nl

Abstract

Entities are a central element of knowledge bases and are important input to many knowledge-centric tasks including text analysis. For example, they allow us to find documents relevant to a specific entity irrespective of the underlying syntactic expression within a document. However, the entities that are commonly represented in knowledge bases are often a simplification of what is truly being referred to in text. For example, in a knowledge base, we may have an entity for Germany as a country but not for the more fuzzy concept of Germany that covers notions of German Population, German Drivers, and the German Government. Inspired by recent advances in contextual word embeddings, we introduce the concept of entity spaces - specific representations of a set of associated entities with near-identity. Thus, these entity spaces provide a handle to an amorphous grouping of entities. We developed a proof-of-concept for English showing how, through the introduction of entity spaces in the form of disambiguation pages, the recall of entity linking can be improved.

Keywords: entity, identity, knowledge representation, entity linking

1. Introduction

Entities are a central element for knowledge bases and text analysis tasks (Balog, 2018). However, the way in which entities are represented in knowledge bases and how subsequent tools use these representations are a simplification of the complexity of many entities. For example, the entity *Germany* in Wikidata as represented by wikidata:Q183 focuses on its properties as a location and geopolitical entity due to its membership as an instance of sovereign state, country, federal state, republic, social state, legal state, and administrative territorial entity. Similarly, in DBpedia (version 2016-10), Germany is represented as entity of type populated place and some subtypes such as yago:WikicatFederalCountries and yago:WikicatMemberStatesOfTheEuropeanUnion.[1]

However, when the term *Germany* is used in text, it can take on many meanings that all have 'something to do' with Germany as it is represented in knowledge bases, but are all not quite the same:

(1) Germany imported 47,600 sheep from Britain last year, nearly half of total imports.

(2) German July car registrations up 14.2 pct yr / yr.

(3) Australia last won the Davis Cup in 1986, but they were beaten finalists against Germany three years ago under Fraser's guidance.

In Example (1), *Germany* refers partly to the location, but a location usually cannot take on an active role, such that the entity 'importing' the sheep is most likely a referent to the German meat industry. *Germany* in Example (2), refers to the German population buying and registering more cars than a year before. Finally, in Example (3), *Germany* refers to the German Davis Cup team from 1993 (the news article is from 1996). In the AIDA-YAGO dataset, this entity is tagged as dbp:Germany_Davis_Cup_team but this presents us with another layer of identities, namely that every year, or every couple of years, the German Davis cup team consists of different players. In 1993, the German Davis cup team consisted of Michael Stich and Marc-Kevin Goellner, in 1996 of David Prinosil and Hendrik Dreekmann and at the time of writing this article in 2019 of Alexander Zverev and Philipp Kohlschreiber. Both MAG (Moussallem et al., 2017) and DBpedia spotlight (Daiber et al., 2013a) annotate *Australia* and *Germany* in Example (3) as dbp:Australia and dbp:Germany respectively. While both the annotations and automatic linkages are close to the identity of the entity in resolving these referents to dbp:Germany, we argue this is an underspecification and highlights a larger problem with identity representation in knowledge bases.

Collapsing of identities has been a frequent topic within Semantic Web discourse. However, most discussions have focused on issues with *owl:sameAs* links (McCusker and McGuinness, 2010; Raad et al., 2018). However, the problem of simplified entity representations (e.g. the collapsing of identities) also occurs before the creation of such *owl:sameAs* links. Specifically, with the fact that most knowledge bases represent a single or limited number of an entity's facets. In this paper, we analyse the extent of the problem by connecting Semantic Web representations of identity to linguistic representations of entities, namely coreference and near-identity. To overcome this identity problem, we argue for the introduction of explicit representations of near-identity within knowledge bases. We term these explicit representations - entity spaces. We illustrate how the introduction of entity spaces can boost the performance of state-of-the-art entity linking pipelines.

Our contributions are: 1) the definition of *entity spaces*; 2) a prototype showing the use of entity spaces over multiple entity linking pipelines; and 3) experiments on 13 English entity linking datasets showing the impact of a more tolerant approach to entity linking made possible through entity spaces.

Our code and experimental results are available via https://github.com/MvanErp/entity-spaces.

[1] Germany also has `rdf:type` dbo:Person but we assume this is a glitch.

The remainder of this paper is organised as follows. In Section 2, we present the background and related work for this research. In Section 3, we provide a definition of entity spaces. To showcase the use of entity spaces, we present our proof-of-concept for a tolerant linker (Section 4). We evaluate our pipelines on thirteen entity linking datasets in Section 5 and follow this by a discussion and conclusion in Sections 6 and 7.

2. Background and Related Work

This research is related to prior work in linguistics concerning coreference resolution, in the Semantic Web concerning entity representation and entity linking as an application. In this section, we discuss each of these topics in turn.

Identity in Linguistics

In (Recasens et al., 2011), an account of the flexibility of language use and how it affects the resolution of entities is provided. They also describe a model for different degrees of near-identity relationships. As the work is focused on coreference, which also includes anaphoric referents such as *he* it moves a bit out of the realm of entity representation in knowledge bases, but the issues and model largely translate to this setting too as the author state that they set out to "develop a more encompassing theoretical account of coreference phenomena that explains under what circumstances linguistic expressions are interpreted as coreferent, or quasi-coreferent." They distinguish coreference where entities with the same feature values corefer, which can be interpreted as entities in knowledge bases that can be linked through *owl:sameAs* relations. There are also coreferents that definitely do not refer to the same entity, in which case there is a non-identity relationship. However, in between identity and non-identity (Recasens et al., 2011) describe a space where entities share most but not all feature values and hence some sort of *near-identity* can be established. Language users perform refocusing and neutralisation operations that highlight differences in feature values leading to greater granularity or neutralise differences leading to lesser granularity respectively.

The phenomenon of near-identity is present in *multifaceted entities* which are entities that can belong to more than one taxonomy. This phenomenon is related to polysemy, in which a word or phrase can have multiple meanings. However, polysemy is generally found to concern regular content words (as found in a dictionary) and not names. Furthermore, besides polysemy, where words can have very different meanings (e.g. battery meaning a container for power and battery meaning the unlawful infliction of personal violence), words can often have slightly different, but related meanings in different contexts. In (Pustejovsky, 1995), seven aspects are discerned that explain different uses and interpretations of noun phrases. Switching between product/producer and container/containee aspects, can for example explain the different interpretations of *newspaper* as organisation, physical object, and the information contained in the newspaper in:[2]

(4) a. The newspapers attacked the President for raising taxes.
b. Mary spilled coffee on the newspaper.
c. John got angry at the newspaper.

The focus in (Pustejovsky, 1995) is on verbs, nouns and adjectives, for which lexicons such as WordNet (Miller, 1995) have definitions. However, named entities such as *Germany* (which can be a location as well as a more abstract administrative entity) are also used ambiguously. Current knowledge bases do not distinguish these facets explicitly.

Near-identity also occurs in *sets*. For example, in discourse, language users seamlessly zoom in and out to express referents to smaller and larger groups or when a set is mentioned and the sum of its members is also mentioned.

Identity in the Semantic Web

Research on identity in the Semantic Web has mostly focused on logical equality - where one thing which has two (or more) names (Halpin et al., 2015). Analysis of *owl:sameAs* relationships has shown that, in many cases, the identity criterion of logical equality is not abided by (De Melo, 2013; Halpin et al., 2015; Raad et al., 2018). The way *owl:sameAs* is used, in certain cases, seems to correspond with the linguistic concepts of coreference and near-identity. Essentially, creators of these links consider two entities logically identical within a given context. In these cases, the creators had a subset of properties of the resources in mind on the basis of which an identity link was established. For example, user might link the ruler of Spain to Franco but that was only true in the context of the time period 1939 - 1975. Automatic detection of erroneous *owl:sameAs* links is presented in (De Melo, 2013; Raad et al., 2018). However, this only shows the extent of the problem, as most knowledge bases (i.e. knowledge graphs) do not have explicit semantics to deal with near-identity or identity in particular contexts. To resolve some of these issues, (McCusker and McGuinness, 2010) have proposed an identity ontology. (Beek et al., 2016) further build on this and compute identity relationships over sets of properties instead of all properties. This allows for the automatic definition of different identity relationships in different contexts.

Our work does not resolve the semantics of expressing various identity relationships, instead we focus on linking to near-identity representations within a knowledge base.

Entity Linking and Contextual Word Embeddings

Entity linking has received much attention in both the computational linguistics as well as semantic web communities since one of the first approaches to link entities to Wikipedia pages was presented in (Milne and Witten, 2008) in 2008. In computational linguistics venues, the named entity linking or named entity disambiguation task is therefore also referred to as 'wikification'.

Many approaches such as DBpedia Spotlight (Daiber et al., 2013b), first focus on identifying phrases in text that may refer to an entity, then try to disambiguate the phrase (as *John Smith* may refer to many different people) by comparing the context of the newly spotted entity to contextual information of DBpedia resources based on abstracts

[2] (Pustejovsky, 1995, pages 91–92)

describing the resources and category information. A critical challenge within entity linking as described in (Rosales-Méndez et al., 2018b) is "What should entity linking link?". Through an analysis of gold standards, (Rosales-Méndez et al., 2018b) shows the difficult nature of defining direct mappings to entities.

To deal with the fuzziness of language, natural language processing has turned towards deep learning. Many named entity linkers now incorporate it in their pipelines (cf. (Gupta et al., 2017; Kolitsas et al., 2018; Raiman and Raiman, 2018)). In particular, deep learning approaches for natural language processing use contextual word embeddings (Devlin et al., 2018; Peters et al., 2018). These embedding approaches assign different vector representations for words depending on their context within a larger sentence or corpus. These sub-symbolic representations are useful in many downstream tasks. However, in entity linking and knowledge base population, an explicit representation that reifies implicit contextual identity is needed. One cannot link to or construct structured queries (e.g. SPARQL) without a representation in the knowledge base. This is where the notion of entity spaces that we introduce plays a role.

3. Defining Entity Spaces

The first contribution of this work is the definition of entity spaces. In the concept of entity spaces, we aim to make the implicit sub-identity of entities, as expressed in human language, explicit.

An **entity space** is an explicit representation of a set of entities in a knowledge base that have a strong near-identity relationship and whose linguistic labels can be used interchangeably in certain contexts.

If we look at the example *Germany* again, from Section 1, we presented examples that refer to the German meat industry, the German population and the German Davis Cup team from 1993. Furthermore, we can imagine *Germany* referring to its government[3] or the actual location.[4]

Named entity recognisers such as (Ratinov et al., 2011), are able to discern between the different contexts in which phrases such as *Germany* are used, and named entity linkers to a certain extent as well (Gupta et al., 2017). However, in most general purpose knowledge bases to which entities are resolved in a named entity linking task, these contexts are not made explicit. While this need not necessarily be a problem when the entity linker correctly links *Germany* to a resource describing the Germany Davis Cup team in a sentence such as (3), when there is no suitable resource to link to, we propose to link to the entity space for *Germany* instead of the country, and it would be even better to link to a *GermanSports* entity space. This corresponds to the concept of 'good enough interpretation' (Poesio et al., 2006).

Thus, one answer, to the question of what should entity linker link to when a system is unsure, is an entity space.

[3] e.g. *"Germany said on Thursday that it had received assurances from the Russian government that its forces would observe the latest ceasefire in Chechnya."*

[4] e.g. *"Motor gasoline stocks dipped slightly as barges left for Germany."*

4. Tolerant Entity Linking

To experiment with this notion of linking to entity spaces, we construct a series of entity linking pipelines that include the capability to link to representations of entity spaces. These pipelines combine a state-of-the-art neural network based entity disambiguator, a state-of-the art flexible named entity recogniser and similarity search.

4.1. Creating entity spaces

We have conceptually defined a **entity spaces**, but to use them we need an explicit representation. Fortunately, we have a reasonable proxy for entity spaces, namely Wikipedia disambiguation pages. Whilst in future work, we would like to build even more explicit entity space representations in knowledge bases, the disambiguation pages are a widely available and useful representation of which entities are closely associated.

While Wikipedia disambiguation pages have been used within named entity disambiguation (Bunescu and Paşca, 2006; Chang et al., 2016), using these pages as explicit representation targets has not been done by these prior methods.

For this purpose, we use the DBpedia 2016-10 Disambiguations dump, which contains information on 269,228 disambiguation pages linking to 1,537,180 different resources. However, we found that this dump is incomplete, we therefore also scraped Wikipedia for Wikipedia Disambiguation pages, resulting in an additional set of 269,062 pages. We found that only 8.8% of the Wikipedia and DBpedia disambiguation pages overlap, indicating that the representation of such pages leaves room for improvement.

The DBpedia Disambiguations dump provides the dbo:wikiPageDisambiguates links indicating which resource they refer to. For the Wikipedia pages, we queried DBpedia to gather the dbo:wikiPageDisambiguates links, resulting in 1,367,929 links. Together, this added up to 2,602,628 page/resource disambiguates page/resource pairs, of which 302,481 (11.6%) occurred both in the DBpedia disambiguation dump and the Wikipedia pages we gathered.

4.2. Entity Linking Pipelines

We perform two series of experiments. In the first series, we use a recent state-of-the-art[5] end-to-end trained neural entity linking system presented in (Kolitsas et al., 2018) as a first baseline, henceforth referred to as NEURAL. The system can be run with an internal named entity recogniser or with pre-recognised entities.[6] To recognise entities separately, we use the FLAIR named entity recogniser (Akbik et al., 2018). FLAIR is useful as it offers state-of-the-art pre-trained models for a number of NER tasks. Here, we use their pre-trained CoNLL-03 four class model trained for CPU systems mode (ner-fast).[7]

In the second series, we use MAG (Moussallem et al., 2017) as an additional baseline system, upon which we build our entity space linking. MAG requires that entities

[5] http://nlpprogress.com/english/entity_linking.html

[6] https://github.com/dalab/end2end_neural_el

[7] https://github.com/zalandoresearch/flair

are pre-recognised, so here the FLAIR entity recogniser is used for all experiments.
If the MAG or NEURAL entity disambiguation system does not return a suitable candidate for linking, our entity space linker kicks in. For all datasets, we test two settings of the entity space linker: a strict setting that requires an exact match between the entity mention and the entity space vector, and a relaxed setting that biases more towards recall.
In summary, we run seven entity linking pipelines:

neural el - The NEURAL system (Kolitsas et al., 2018) run in entity linking setup.

neural ed - NER using FLAIR. The text with NER spans are provided to NEURAL for named entity disambiguation.

neural d1.0 - The NEURAL ED setting but if a no entity is returned the system searches for corresponding Wikipedia disambiguation using the SimString (Okazaki and Tsujii, 2010) approximate similarity search algorithm with an exact similarity as measured by the Jaccard distance.

neural d0.8 - The same setting as NEURAL D1.0 but with a threshold of 0.8 similarity.

mag ed - NER using FLAIR. The text with NER spans are provided to the MAG entity disambiguator.

mag d1.0 - The MAG ED pipeline but if no entity return we perform a similarity search for an Wikipedia disambiguation page as in NEURAL D1.0.

mag d0.8 - The same pipeline as MAG D1.0 but with a with a threshold of 0.8 similarity.

5. Evaluation

There are a number of entity linking datasets available for evaluation. In (Rosales-Méndez et al., 2018a), 15 multilingual datasets are mentioned, and GERBIL (Röder et al., 2017) contains 23 datasets for English. However, not all datasets represent ambiguous entities well, as analyses of a subset of some commonly used entity linking datasets in (Van Erp et al., 2016) and (Ilievski et al., 2018) indicate. In this paper, we investigate 13 datasets. These datasets were chosen based on their previous use in entity linking evaluations, their availability and breadth (news, microblogs and wikipedia pages). As the GERBIL framework does not provide access to the system output which we think is paramount to understanding the tested systems, we chose to do an offline analysis and system evaluation (Section 6).
The datasets we investigate present a flattened view of entities, which already starts with their named entity annotations. For example, datasets will consistently annotate *Germany* as an entity of type LOCATION even when the context indicates its meaning in the sense of a GEOPOLITICAL ENTITY. This extends to the entity links provided in the gold standards, for example for *Germany* in Example (1) is annotated with the link to the Wikipedia page on Germany.
In this section, we first describe the datasets characteristics. We then evaluate the performance of the 7 entity linking pipelines. We also give indications of the extent of multi-faceted entities contained in the datasets.

5.1. Evaluation Datasets

With many evaluation datasets being available in various formats in various places, it can be difficult to obtain the exact same version of a dataset as reported on in prior work. As the foundation for our entity linking experiments, we use the linker developed in (Kolitsas et al., 2018) and we thus wanted to first reproduce their results in order to have a solid basis for our experiments. This proved to be non-trivial as (Kolitsas et al., 2018) performed some their evaluations using Gerbil and on some additional datasets not in Gerbil. We attempted to obtain their versions of the datasets or get as close as possible to them. To illustrate the different outcomes of different versions of datasets, we sometimes test our approach on two different versions of a dataset.[8]
We used the following datasets:

ACE2004 (Mitchell et al., 2005) was created for the Automatic Content Extraction challenge that took place in 2004. For named entity linking, entity links were added via a crowdsourcing. We have two versions of this dataset, one provided in the data directory of (Kolitsas et al., 2018) and one provided by (Rosales-Méndez et al., 2019) containing the first 20 articles from the ACE evaluation.

AIDA-YAGO (Hoffart et al., 2011) enriches the original CoNLL 2003 entity recognition dataset[9] with entity links. We have two version of this dataset, one provided in the data directory of (Kolitsas et al., 2018) and the original version from (Hoffart et al., 2011).[10]

AQUAINT (Milne and Witten, 2008) This can be considered the first entity linking evaluation dataset. It consists of 50 randomly selected articles from the English portion of the AQUAINT text. We used the version provided in the data directory of (Kolitsas et al., 2018).[11] The system in (Milne and Witten, 2008) was run on the articles and through crowdsourcing, the correctness of each link was judged and corrected if necessary.

DBpedia Spotlight (Mendes et al., 2011) consists of 35 paragraphs from New York Times documents from 8 different categories. Multiple annotators independently annotated the corpus, after which disagreements were resolved through discussion. We used the NIF version provided in the Gerbil data directory and converted it to the AIDA-YAGO format using NiFify v2.

[8] We could not deduce whether the differences arose from different preprocessing or reformatting, but we leave this as an open discussion for future work.
[9] https://www.clips.uantwerpen.be/conll2003/ner/
[10] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/
[11] https://catalog.ldc.upenn.edu/LDC2002T31

Derczynski (Derczynski et al., 2015) consists of 182 microblog posts, each annotated by three NLP experts. We used the NIF version provided in the Gerbil data directory.

KORE50 (Hoffart et al., 2012) is a 50-sentence subset of the AIDA-YAGO dataset that aims to capture hard to disambiguate entity mentions. We used the NIF version provided in the Gerbil data directory and converted it to the AIDA-YAGO format using NiFify v2.

MSNBC (Cucerzan, 2007) consists of 20 news articles from MSNBC's ten news categories from 2 January 2007. The (Cucerzan, 2007) system was run over it and errors were corrected in a post-hoc annotation. We have two versions of this dataset, one provided in the data directory of (Kolitsas et al., 2018) and one provided by (Rosales-Méndez et al., 2019).

OKE2015 (Nuzzolese et al., 2015) consists of 196 manually annotated sentences from Wikipedia for the Open Knowledge Extraction Challenge held at ESWC 2015. We used the original data from the OKE Challenge page.

VoxEL (Rosales-Méndez et al., 2018a) consists of 15 annotated news articles from a Voxeurop[12] in 5 languages. The dataset contains two variants, a strict version only containing annotations of entity mentions of types *person*, *location*, and *organisation*, and a relaxed version, which also contains annotations of a miscellaneous or *other* entity mention type. This to account for the lack of consensus on what an entity is exactly. We obtained the data through the VoxEL dataset page and only use the English language portion.

In Table 1, we provide statistics on the datasets. Besides general statistics on the size and number of entities, we also provide a hint on the ambiguity contained in the dataset, namely the maximum number of links per surface form and the maximum number of surface forms per link. This reporting is inspired by (Van Erp et al., 2016).

The maximum number links per surface form indicates how many different resources the same string is linked to. For example *World Cup* in AIDA-YAGO can refer to FIFA World Cup, Rugby World Cup, FIS Freestyle Skiing World Cup, FIS Alpine Ski World Cup, Biathlon World Cup, FIS Ski Jumping World Cup and Speed Skating World Cup. If knowledge bases (and their associated evaluation datasets) would contain more complex entity representations, we would see higher numbers of links per surface form, for example to distinguish between the different meanings of the term *Germany* presented in Section 1. The Germany disambiguation page for example links 21 different pages, ranging from the country, its different names through time, people with 'Germany' in their names and a race horse name 'Germany'.

The maximum number of surface forms per link refers to how many different strings are used to refer to the same resource. For example in ACE2004-20 wikipedia:Florida can be referred to as *State of Florida*, *Florida*, or *Fla.*. As the relatively low numbers for these statistics indicate, these datasets do not represent extremely ambiguous entities and entity mentions.

[12] https://voxeurop.eu/en

5.2. Results

To evaluate the performance of the entity linking pipelines, we use the CoNLL NER evaluation script.[13] We first run the system from (Kolitsas et al., 2018) and MAG on the datasets without introducing entity spaces in order to check whether we can reproduce previously reported scores for these systems. For AIDA-YAGO (both versions), and msnbc (from the (Kolitsas et al., 2018) distribution), our results are similar to those reported in (Kolitsas et al., 2018). For KORE50 and OKE 2015, our scores using the NEURAL EL setup are more than 10 points lower.[14]

As four of our pipelines suggest disambiguation pages, we cannot directly match their output to the gold standard. We, therefore, check if the gold standard link is connected to the disambiguation page via a dbo:wikiPageDisambiguates link. If this is the case, we deem the suggestion as correct. We are aware of the fact that this suggestion is less precise in some cases than the gold standard, but in cases where the alternative is no information about an entity mention, we believe this is preferable. Furthermore, our pipelines *only* suggests a link when the baseline system does not, therefore previous correct system predictions are not harmed. Table 2, shows the microaveraged precision, recall and F_1 measure on the 13 datasets. The best F-measure is highlighted in bold.

We perform additional analyses of the results of the entity linking pipelines. In previous named entity linking evaluations, it seems that non-linkable entities are not taken into account (cf. (Ratinov et al., 2011; Kolitsas et al., 2018)). Non-linkable entities (often denoted NIL or NME) are entities which are marked up in the dataset as entities, but there is no known matching resource in the knowledge base. However, this makes the results look better than they are, as the most difficult cases are arguably ignored.

In Table 3, we show the results of our baselines and pipelines when non-linkable entities are taken into account. As not all datasets contain marked non-linkable entities, we only report results on the five datasets that do. Table 4 presents the number of non-linkable entities that were assigned a link using pipelines based on neural end-to-end system.

6. Analysis & Discussion

The results presented above suggest a clear benefit to taking a more tolerant approach to entity linking. In Table 2, we see that from the 13 datasets, a pipeline that uses an entity space performs better in 8 of the cases. This is by increasing recall. In some cases, providing a 12 percentage point boost in recall in the case of sVoxEL over the already more tolerant NEURAL ED setting. This is, in some sense to

[13] https://www.clips.uantwerpen.be/conll2002/ner/bin/conlleval.txt

[14] As the experiments in(Kolitsas et al., 2018) were not reproducible, a side-by-side comparison was not possible and this is based on Table 2 in (Kolitsas et al., 2018)

Dataset	Type	Tokens	Entities	Unique entities	Max. links	Max. surface
ACE2004 (Kolitsas et al., 2018)	news	15,100	255	197	2	3
ACE2004-20 (Rosales-Méndez et al., 2019)	news	25,645	285	233	2	3
AIDA-YAGO (Hoffart et al., 2011)	news	46,165	5,616	2,612	7	8
AIDA-YAGO (Kolitsas et al., 2018)	news	46,395	4,459	1,945	7	8
AQUAINT	news	11,972	711	587	2	3
DBpedia Spotlight	news	1,941	322	264	3	2
Derczynski	tweets	3925	289	271	1	1
KORE 50	news	722	141	127	2	2
MSNBC (Rosales-Méndez et al., 2019)	news	12,349	710	424	3	8
msnbc (Kolitsas et al., 2018)	news	12,501	650	391	2	8
OKE 2015	Wikipedia	3,481	664	440	1	2
rVoxEL	news	2,420	604	371	2	5
sVoxEL	news	2,420	204	107	1	5

Table 1: General entity statistics on evaluation datasets, "Max. links" denotes the maximum number of links per surface form and "Max. surface" denotes the maximum number of surface forms per link.

be expected, as the probability of connecting to the entity is built into the tolerance given by allowing for querying using the disambiguation page. However, by linking to an entity space, one at least has the opportunity for finding the identity connection in subsequent structured queries. For example, by performing a SPARQL query to find all pages where Germany as a country (i.e) is mentioned or could be mentioned (i.e .)

Interestingly, it is hard to achieve better results even with a more tolerant approach over the end-to-end trained entity linker when the linker (NEURAL EL) has been fitted on the dataset as is the case with AIDA-YAGO. The difference in performance on out-of-domain datasets such as Derczynski and OKE indicates that this system is highly tuned toward the data. However, given the diversity of text, a tolerant approach is beneficial. Another concern is the difference of almost 5 percentage points between the original AIDA-YAGO dataset and one supplied by the system modellers, indicating that replicability of systems is still non-trivial.

We also observe that the use of entity spaces, at least using the representation we have chosen here, requires a strong baseline entity linker. Using MAG, only in 6/13 cases did the entity space approach improve the f_1 score and in most no cases was the improvement in recall over 1 percentage point.

A promising subset of entities for the use of entity spaces are non-linkable entities. For downstream tasks, it is beneficial to know some characteristics of the entity, even if the exact resource describing an entity is not present. As the results in Table 3 show, the impact of non-linkable entities can mean a drop in up to nearly 9 points in F_1 for the AIDA-YAGO dataset in the baseline setup (NEURAL EL). However, the entity spaces pipelines manage to recover some of this by providing links to supposed non-linkable entities, as is shown in Table 4.

To precisely assess the impact of entity spaces, richer evaluation datasets are necessary that contain annotation layers that express near-identity links. As this would be a paper on its own, we refrain from creating such annotations for the given datasets here and we inspect the returned results instead. The entity space suggestions are promising mostly for non-person entities. We see for example that for *ARAB*, which is marked as non-linkable in AIDA-YAGO, our pipeline suggests wikipedia:Arab_(disambiguation). It links *Prince Rupert* to wikipedia:Prince_Rupert_(disambiguation). There is more work to be done, as in many cases when a person shares a name with a more famous person (e.g. a journalist whose name is listed as the author of a news article sharing a name with an athlete or politician), our pipelines erroneously provide a link to the famous person. When not to link is still a difficult problem for many entity linkers. Smarter contextual knowledge may aid here.

7. Conclusions and Future Work

In this paper, we introduced entity spaces - an explicit representation of entities that have a strong near-identity relationship. To demonstrate the use of entity-spaces, we use Wikipedia disambiguation pages to experiment with this concept. We showed the need for entity spaces by documenting the negative effect of non-linkable entity mentions on entity linkers. Entity spaces provide a good default option for entity linkers when they have low confidence about what entity to link to.

There are a number of avenues for future work both in terms of representation of entity spaces and their use in text enrichment and knowledge base tasks. First, in terms of representation, Wikipedia disambiguation pages contain subsections, e.g. for Germany there are sections 'Other political entities', 'people' and 'other'. This information is not available in structured form. Furthermore, these categories are not complete. We believe that explicitly modelling different entity space contexts would better represent the nature of near-identity. In addition, investigating the overlap between entity space representations and web architecture concepts would be an interesting area of exploration.

In terms of downstream tasks, more explicit and better contextual knowledge in entity spaces, will enable systems to

	NEURAL EL	NEURAL ED	NEURAL ED D1.0	NEURAL ED D0.8	MAG ED	MAG ED D1.0	MAG ED D0.8
ACE2004	91.88 P	89.86 P	89.20 P	89.59 P	64.32 P	66.53 P	66.53 P
	70.98 R	76.47 R	74.51 R	77.65 R	60.78 R	61.57 R	61.57 R
	80.09 F	82.63 F	81.20 F	**83.19 F**	62.50 F	63.95 F	63.95 F
ACE2004-20 (Rosales-Méndez et al., 2019)	84.21 P	81.55 P	78.14 P	78.70 P	51.69 P	51.31 P	51.13 P
	56.14 R	58.95 R	58.95 R	59.65 R	48.42 R	48.07 R	47.72 R
	67.37 F	**68.43 F**	67.20 F	67.86 F	50.00 F	49.64 F	49.36 F
AIDA-YAGO* (Hoffart et al., 2011)	79.82 P	75.03 P	72.46 P	72.45 P	42.94 P	42.82 P	43.06 P
	77.08 R	78.39 R	79.09 R	78.97 R	52.58 R	52.40 R	52.66 R
	78.43 F	76.68 F	75.63 F	75.57 F	47.27 F	47.13 F	47.38 F
AIDA-YAGO* (Kolitsas et al., 2018)	87.71 P	85.75 P	83.67 P	83.82 P	54.98 P	54.72 P	54.62 P
	79.19 R	80.62 R	81.33 R	81.58 R	54.00 R	53.82 R	53.62 R
	83.23 F	83.10 F	82.48 F	82.68 F	54.48 F	54.27 F	54.11 F
AQUAINT	76.55 P	76.25 P	74.74 P	75.46 P	53.88 P	54.48 P	54.77 P
	38.12 R	40.65 R	40.79 R	40.65 R	31.22 R	31.65 R	31.50 R
	50.89 F	**53.03 F**	52.78 F	52.83 F	39.54 F	40.04 F	40.00 F
DBpedia Spotlight	71.43 P	66.15 P	67.69 P	69.70 P	39.13 P	39.13 P	40.00 P
	10.87 R	13.35 R	13.66 R	14.29 R	8.39 R	8.39 R	8.70 R
	18.87 F	22.22 F	22.74 F	**23.71 F**	13.81 F	13.81 F	14.29 F
Derczynski	65.38 P	47.02 P	46.24 P	45.50 P	29.44 P	30.17 P	30.36 P
	22.37 R	25.99 R	26.32 R	28.29 R	24.01 R	24.01 R	24.67 R
	33.33 F	33.47 F	33.54 F	**34.89 F**	26.45 F	26.74 F	27.22 F
KORE 50*	68.09 P	66.32 P	54.62 P	55.30 P	26.28 P	25.90 P	25.74 P
	22.70 R	44.68 R	50.35 R	51.77 R	25.53 R	25.53 R	24.82 R
	34.04 F	53.39 F	52.40 F	**53.48 F**	25.90 F	25.71 F	25.27 F
MSNBC (Rosales-Méndez et al., 2019)	52.96 P	55.61 P	52.73 P	52.06 P	44.55 P	42.33 P	41.42 P
	30.28 R	32.82 R	39.44 R	39.15 R	44.37 R	41.97 R	41.13 R
	38.53 F	41.28 F	**45.12 F**	44.69 F	44.46 F	42.15 F	41.27 F
msnbc* (Kolitsas et al., 2018)	85.79 P	85.88 P	85.31 P	85.52 P	67.15 P	67.69 P	67.79 P
	72.46 R	78.62 R	78.62 R	78.15 R	63.85 R	64.15 R	64.77 R
	78.57 F	**82.09 F**	81.83 F	81.67 F	65.46 F	65.88 F	66.25 F
OKE 2015	55.14 P	50.68 P	49.23 P	49.61 P	42.12 P	41.63 P	41.39 P
	40.50 R	42.79 R	43.94 R	44.16 R	42.79 R	42.11 R	42.33 R
	46.70 F	46.40 F	46.43 F	**46.73 F**	42.45 F	41.87 F	41.86 F
rVoxEL	80.12 P	77.61 P	77.68 P	78.79 P	59.92 P	59.54 P	59.16 P
	22.68 R	25.83 R	29.97 R	30.13 R	25.99 R	25.83 R	25.66 R
	35.35 F	38.76 F	43.25 F	**43.59 F**	36.26 F	36.03 F	35.80 F
sVoxEL	92.19 P	87.01 P	87.29 P	87.85 P	74.75 P	72.77 P	72.45 P
	57.84 R	65.69 R	77.45 R	77.94 R	74.02 R	72.06 R	72.08 R
	71.08 F	74.86 F	82.08 F	**82.60 F**	74.38 F	72.41 F	72.26 F

Table 2: Micro-averaged Precision (P), Recall (R) and F_1 measure on baseline system (NEURAL EL), FLAIR-ner + neural entity disambiguation (NEURAL ED) and FLAIR ner + entity space linking at 10 and 0.8 matching (NEURAL ED D1.0 and NEURAL ED D08.) and MAG with FLAIR ner (MAG ED) and MAG with FLAIR ner + entity space linking at 10 and 0.8 matching (MAG ED D1.0 and MAG ED D0.8). The * symbol behind the dataset name indicates that our results are similar to those reported in (Kolitsas et al., 2018).

avoid erroneous links, for example when a non-famous person shares a name with a famous person. Alternative links such as links to categories (e.g. journalists) could be an option here to provide some information for non-linkable entities that may aid analyses in downstream tasks. We believe that directly training models using entity spaces is also an interesting avenue for future work. Our entity spaces linker was set up to be quite conservative and not touch links suggested by the baseline systems, but with more contextual knowledge, pro-active attempts to fix potential erroneous links could be explored.

In this paper, we have connected theoretical frameworks on entity and identity from linguistics to identity on the Semantic Web by introducing the concept of *entity spaces*. The introduction of explicit representations of near-identity can provide an important foundation for creating and using complex contextual entity representations in knowledge bases.

	NEURAL EL	NEURAL ED	NEURAL ED D1.0	NEURAL ED D0.8	MAG ED	MAG ED D1.0	MAG ED D0.8
ACE2004-20	84.21 P	81.55 P	77.67 P	78.24 P	51.69 P	51.31 P	51.13 P
(Rosales-Méndez et al., 2019)	56.14 R	58.95 R	58.60 R	59.30 R	48.42 R	48.07 R	47.72 R
	67.37 F	68.43 F	66.80 F	67.47 F	50.00 F	49.64 F	49.36 F
AIDA-YAGO	79.82 P	75.03 P	71.73 P	71.71 P	42.94 P	42.82 P	43.04 P
(Hoffart et al., 2011)	61.56 R	62.61 R	62.52 R	62.43 R	41.99 R	41.84 R	42.04 R
	69.51 F	68.26 F	66.81 F	66.75 F	42.46 F	42.33 F	42.54 F
Derczynski	65.38 P	47.02 P	45.66 P	44.97 P	29.44 P	30.17 P	30.36 P
	22.37 R	25.99 R	25.99 R	27.96 R	24.01 R	24.01 R	24.67 R
	33.33 F	33.47 F	33.12 F	34.48 F	26.45 F	26.74 F	27.22 F
rVoxEL	80.12 P	77.61 P	66.95 P	67.97 P	59.92 P	59.54 P	59.16 P
	22.68 R	25.83 R	25.83 R	25.99 R	25.99 R	25.83 R	25.66 R
	35.35 F	38.76 F	37.28 F	37.60 F	36.26 F	36.03 F	35.80 F
sVoxEL	92.19 P	87.01 P	74.03 P	74.59 P	74.75 P	72.77 P	75.51 P
	57.84 R	65.69 R	65.69 R	66.18 R	74.02 R	72.06 R	75.13 R
	71.08 F	74.86 F	69.61 F	70.13 F	74.38 F	72.41 F	75.32 F

Table 3: Micro-averaged Precision (P), Recall (R) and F_1 measure on datasets including non-linkable entities. Evaluated settings are the baseline system (NEURAL EL), FLAIR-ner + neural entity disambiguation (NEURAL ED) and FLAIR ner + entity space linking at 10 and 0.8 matching (NEURAL ED D1.0 and NEURAL ED D08.) and MAG with FLAIR ner (MAG ED) and MAG with FLAIR ner + entity space linking at 10 and 0.8 matching (MAG ED D1.0 and MAG ED D0.8).

		NEURAL EL		NEURAL ED		NEURAL ED D1.0		NEURAL ED D0.8	
Dataset	NIL	no link	link	no link	link	no link	link	no link	link
ACE2004-first20	40	36	4	32	8	31	9	30	10
AIDA-YAGO (Hoffart et al., 2011)	1,131	849	282	672	459	598	533	596	535
Derczynski	80	74	6	56	24	53	27	50	30
rVoxEL	3	2	1	2	1	2	1	2	1
sVoxEL	3	2	1	2	1	2	1	2	1

Table 4: Number of non-linkable entity mentions in total (NIL) and number of unlinked (no link) and linked entity mentions (link) after each experiment.

Global namespaces and identity are a key innovation of the Semantic Web, we believe that entity spaces can help bring near-identity to structured data on the web as well.

Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Balog, K. (2018). Entity linking. In *Entity-Oriented Search*, pages 147–188. Springer.

Beek, W., Schlobach, S., and van Harmelen, F. (2016). A contextualised semantics for owl: sameas. In *European Semantic Web Conference*, pages 405–419. Springer.

Bunescu, R. and Paşca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *11th conference of the European Chapter of the Association for Computational Linguistics*.

Chang, A. X., Spitkovsky, V. I., Manning, C. D., and Agirre, E. (2016). Evaluating the word-expert approach for named-entity disambiguation. *arXiv preprint arXiv:1603.04767*.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013a). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013b). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM.

De Melo, G. (2013). Not quite the same: Identity constraints for the web of linked data. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2-7 Jun. preprint at arXiv:1810.04805.

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

Halpin, H., Hayes, P. J., and Thompson, H. S. (2015). When owl: sameas isn?t the same redux: towards a theory of identity, context, and inference on the semantic web. In *International and Interdisciplinary Conference on Modeling and Using Context*, pages 47–60. Springer.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

Hoffart, J., Seufert, S., Nguyen, D. B., Theobald, M., and Weikum, G. (2012). KORE: keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 545–554. ACM.

Ilievski, F., Vossen, P., and Schlobach, S. (2018). Systematic study of long tail phenomena in entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 664–674.

Kolitsas, N., Ganea, O.-E., and Hofmann, T. (2018). End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529.

McCusker, J. P. and McGuinness, D. L. (2010). Towards identity in linked data. In *OWLED*.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

Mitchell, A., Strassel, S., Huang, S., and Zakhary, R. (2005). Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 1:1–1.

Moussallem, D., Usbeck, R., Röeder, M., and Ngomo, A.-C. N. (2017). MAG: A multilingual, knowledge-base agnostic and deterministic entity linking approach. In *Proceedings of the Knowledge Capture Conference (K-CAP 2017)*, Austin, TX, USA, 4 -6 Dec. ACM.

Nuzzolese, A. G., Gentile, A. L., Presutti, V., Gangemi, A., Garigliotti, D., and Navigli, R. (2015). Open knowledge extraction challenge. In *Semantic Web Evaluation Challenges*, pages 3–15. Springer.

Okazaki, N. and Tsujii, J. (2010). Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 851–859, Beijing, China, August.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Poesio, M., Sturt, P., Artstein, R., and Filik, R. (2006). Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse processes*, 42(2):157–175.

Pustejovsky, J. (1995). The generative lexicon: A theory of computational lexical semantics.

Raad, J., Beek, W., van Harmelen, F., Pernelle, N., and Saïs, F. (2018). Detecting erroneous identity links on the web using network metrics. In Denny Vrandečič, et al., editors, *Proceedings of the 17th International Semantic Web Conference (ISWC 2018)*.

Raiman, J. R. and Raiman, O. M. (2018). Deeptype: multilingual entity linking by neural type system evolution. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ratinov, L., Roth, D., Downey, D., and Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics.

Recasens, M., Hovy, E., and Martí, M. A. (2011). Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Röder, M., Usbeck, R., and Ngonga Ngomo, A.-C. (2017). Gerbil–benchmarking named entity recognition and linking consistently. *Semantic Web*, pages 1–21.

Rosales-Méndez, H., Hogan, A., and Poblete, B. (2018a). Voxel: A benchmark dataset for multilingual entity linking. In *International Semantic Web Conference*, pages 170–186. Springer.

Rosales-Méndez, H., Poblete, B., and Hogan, A. (2018b). What should entity linking link? In *AMW*.

Rosales-Méndez, H., Hogan, A., and Poblete, B. (2019). NIFify: Towards Better Quality Entity Linking Datasets. In *WWW'19 Companion*.

Van Erp, M., Mendes, P. N., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., and Waitelonis, J. (2016). Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *LREC*, volume 5, page 2016.