



UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

TOWARDS A FINE-GRAINED ENTITY LINKING APPROACH

TESIS PARA OPTAR AL GRADO DE
DOCTOR EN COMPUTACIÓN

HENRY ROSALES MÉNDEZ

PROFESORES GUÍAS:
AIDAN HOGAN
BARBARA POBLETE LABRA

MIEMBROS DE LA COMISIÓN:
FELIPE BRAVO MÁRQUEZ
CLAUDIO GUTIÉRREZ GALLARDO
GERHARD WEIKUM

Este trabajo fue financiado por CONICYT-PCHA/Doctorado Nacional/2016-21160017

SANTIAGO DE CHILE
2021

Resumen

La tarea Entity Linking (EL) implica vincular menciones de entidades en un texto con su identificador correspondiente en una base de conocimiento (KB) como Wikipedia, Babel-Net, DBpedia, Freebase, Wikidata, YAGO, etc. Se han propuesto numerosas técnicas para abordar esta tarea a lo largo de los años. Sin embargo, no todos los trabajos adoptan la misma convención con respecto a las entidades a las que debe desambiguar la tarea EL; por ejemplo, mientras que algunos trabajos EL apuntan a entidades comunes como “entrevista” que aparece en la base de conocimientos, otros solo apuntan a entidades nombradas como “Michael Jackson”. La falta de consenso sobre este tema (y otros) complica la investigación sobre la tarea EL; por ejemplo, ¿cómo se puede evaluar y comparar el rendimiento de los sistemas EL cuando los sistemas pueden apuntar a diferentes tipos de entidades? Si bien los enfoques tradicionales de EL se han centrado principalmente en textos en inglés, este problema no afecta solo al inglés, sino también a cada idioma.

En esta tesis, primero destacamos la importancia de formalizar el concepto de “entidad” y los beneficios que traería a la comunidad de Entity Linking, en particular, los relacionados con la construcción y evaluación de *gold standards* con fines de evaluación. Motivados por la escasez de datasets anotados, incluso más en escenarios multilingües, proponemos VOXEL: un *gold standard* anotado manualmente para EL multilingüe con el mismo texto en cinco idiomas europeos. Se seleccionaron cinco sistemas multilingües para comparar sus comportamientos. En general, nuestros resultados identifican cómo se comparan los resultados de diferentes idiomas y, además, sugieren que la traducción automática es ahora una alternativa competitiva al EL multilingüe.

El evidente desacuerdo sobre “¿*Cuáles entidades se deben enlazar?*” es también consecuencia de las diferentes aplicaciones que existen de EL. En lugar de proponer soluciones aisladas, nuestra posición es crear una definición más granular que cubra la mayoría de las necesidades actuales. En esta línea, proponemos un esquema de categorización detallado para EL que distingue diferentes tipos de menciones y enlaces. Proponemos una extensión del vocabulario actual que permite expresar tales categorías en conjuntos de datos de referencia de EL. Luego volvemos a etiquetar (subconjuntos de) tres conjuntos de datos EL populares de acuerdo con nuestro novedoso esquema de categorización, donde además discutimos una herramienta utilizada para semi-automatizar el proceso de etiquetado. A continuación, presentamos los resultados de desempeño de cinco sistemas EL para categorías individuales. Ampliamos aún más los sistemas EL con componentes Word Sense Disambiguation y Coreference Resolution, creando versiones iniciales de lo que llamamos sistemas *Fine-Grained Entity Linking (FEL)*, midiendo el impacto en el rendimiento por categoría. Finalmente, proponemos una medida de rendimiento configurable basada en conjuntos difusos que se pueden adaptar a diferentes escenarios de aplicación. Nuestros resultados destacan una falta de consenso sobre los objetivos de la tarea EL, muestran que los sistemas evaluados efectivamente se dirigen a diferentes entidades y revelan además algunos desafíos abiertos para la tarea (F) EL con respecto a formas más complejas de referencia para entidades.

Abstract

The Entity Linking (EL) task involves linking mentions of entities in a text with their corresponding identifier in a Knowledge Base (KB) such as Wikipedia, BabelNet, DBpedia, Freebase, Wikidata, YAGO, etc. Numerous techniques have been proposed to address this task down through the years. However, not all works adopt the same convention regarding the entities that the EL task should target; for example, while some EL works target common entities like “interview” appearing in the KB, others only target named entities like “Michael Jackson”. The lack of consensus on this issue (and others) complicates research on the EL task; for example, how can the performance of EL systems be evaluated and compared when systems may target different types of entities? While traditional EL approaches have largely focused on English texts, this problem does not affect only English, but also each language.

In this thesis, we first highlight the importance of formalizing the concept of “entity” and the benefits it would bring to the Entity Linking community, in particular, relating to the construction and evaluation of gold standards for evaluation purposes. Motivated by the scarcity of annotated datasets – even more in multilingual scenarios – we propose VOXEL: a manually-annotated gold standard for multilingual EL featuring the same text expressed in five European languages. We compare the behavior of state of the art EL (multilingual) systems for five different languages. Overall, our results identify how the results of different languages compare and suggest that machine translation is now a competitive alternative to dedicated multilingual EL configurations.

The evident disagreement about “*What should entity linking link?*” is also a consequence of the different applications of EL. Rather than proposing isolated solutions, our position is to create a more granular definition that meets the majority of current needs. In this line, we propose a fine-grained categorization scheme for EL that distinguishes different types of mentions and links. We propose a vocabulary extension that expresses such categories in EL benchmark datasets. We then relabel (subsets of) three popular EL datasets according to our novel categorization scheme, where we additionally discuss a tool used to semi-automate the labeling process. We next present the performance results of five EL systems for individual categories. We further extend EL systems with Word Sense Disambiguation and Coreference Resolution components, creating initial versions of what we call *Fine-Grained Entity Linking (FEL)* systems, measuring the impact on performance per category. Finally, we propose a configurable performance measure based on fuzzy sets that can be adapted for different application scenarios. Our results highlight a lack of consensus on the goals of the EL task, show that the evaluated systems do indeed target different entities, and further reveal some open challenges for the (F)EL task regarding more complex forms of reference for entities.

to my dear mom, my greatest treasure

Acknowledgements

I would like to take this opportunity to sincerely thank all DCC professors for their wisdom, teaching ability, and kindness. Especially to my advisors Aidan Hogan and Barbara Poblete, who are amazing human beings and extraordinary researchers. I couldn't have had better guidance and support.

Thank you to the committee members Claudio Gutiérrez, Felipe Bravo, and Gerhard Weikum for their valuable suggestions and thoughtful comments, many of which have been incorporated into the final manuscript.

I would not have been able to get that far if it weren't for my family. Although far away, they were always with me in every stumble, in every achievement. Every step I take is always thinking of them.

I could not be more proud of my friends. I want to thank them for being available and forgiving the time that I spend without being in contact. Thanks to my dear Jacqueline, who by my side has witnessed all this journey.

Contents

1	Introduction	1
1.1	Problem Statement	2
1.2	Hypothesis	5
1.3	Research Goals	6
1.3.1	General Goals	6
1.3.2	Specific Goals	6
1.4	Contributions	6
1.5	Publications	8
1.6	Thesis outline	9
2	Preliminaries	11
2.1	Definition of “entity”	12
2.2	EL Formalisms	13
2.3	Resource Description Framework	15
2.4	RDF query languages	18
3	Related Work	20
3.1	Entity Recognition	20
3.1.1	NERL strategy	21
3.1.2	End-to-End strategy	22
3.2	Entity Disambiguation	23
3.2.1	Candidate Entity Generation	23
3.2.2	Candidate Entity Ranking	25
3.2.3	Unlinkable Mention Prediction	27
3.3	Entity Linking Evaluation Measures	28
3.4	EL Benchmark Datasets	30
3.5	EL Formats	31
3.6	EL Design Issues	34
4	Consensus about Entity Linking	38
4.1	Questionnaire on the Goals of EL	39
4.2	Proposed Solution	43

5	Multilingualism	45
5.1	Multilingual EL Systems	46
5.1.1	Multilingual EL Datasets	47
5.1.2	Non-English Entity Linking: Spanish use-case	49
5.2	The VoxEL Dataset	54
5.3	Multilingual EL performance	56
5.3.1	Why not translate to English?	58
5.4	Translating across languages	61
6	Fine-Grained Entity Linking	64
6.1	Fine-Grained Categories	64
6.1.1	Base Form	65
6.1.2	Part of Speech	66
6.1.3	Overlap	67
6.1.4	Reference	67
6.2	Fine-Grained EL Format	68
6.2.1	Vocabulary	68
6.2.2	Extending NIF	69
6.3	Fine-Grained Entity Annotation	73
6.4	NIFify	75
6.4.1	NIF Construction	76
6.4.2	Validation	77
6.4.3	Result Visualization	79
6.5	Relabeling KORE50, ACE2004 and VoxEL	81
6.6	Fine-Grained Evaluation	83
6.7	Fuzzy Recall and F_1 Measures	86
6.7.1	Fuzzy Framework	86
6.7.2	Fuzzy Evaluation	88
7	Fine-Grained Entity Linking Systems	91
7.1	Adding Coreference Resolution	91
7.2	Adding Word Sense Disambiguation	92
7.3	Combined CR and WSD Results	94
8	Conclusions and Future Work	97
8.1	Contributions and results	98
8.2	Limitations and Future Work	99
8.3	Outlook	102
	Bibliography	104

List of Figures

1.1	Annotations of Babelfy (b), DBpedia Spotlight (d), FRENTE (f) and TagME (t) on the same sentence	5
2.1	RDF graph extracted from DBpedia	16
2.2	Serialization of the same RDF graph using the (a) RDF/XML and (b) Turtle formats	17
2.3	Example of an SPARQL query	19
3.1	Example of Part-of-Speech tags	21
3.2	SPARQL query to obtain the number of triples in a RDF KB.	24
3.3	MSNBC format for EL annotations	32
3.4	IITB format for EL annotations	32
3.5	AIDA/CoNLL format for EL annotations	33
3.6	SemEval format for EL annotations	34
3.7	CAT format for EL annotations	35
3.8	NIF format for EL annotations (in Turtle syntax)	36
4.1	The two sentences used for the questionnaire annotated with the ratio of respondents who suggested to annotate the corresponding mentions with some link; in the case of underlined mentions, multiple links were proposed, as presented in Table 4.1.	39
5.1	Summary of the Micro- F_1 results over VOXEL Relaxed/Strict for the translation experiments, comparing mean values for setting the EL system to the language of the text (<i>Calibrated</i>), translating the text to English first (<i>Translation</i>), and the corresponding F_1 score for EL over the original English text (<i>English</i>)	61
6.1	EL categorization scheme with concrete alternatives (leaf-nodes) shaded for each dimension	65
6.2	Hierarchy of classes belonging to the Fine-Grained Entity Linking vocabulary and its links to external vocabularies.	70
6.3	NIF triples to specify the annotation of “Moscow” from sentence S3 ; we use multiple annotations to denote an <i>OR</i> over the links	72

6.4	NIF triples to specify the annotation of “them” from sentence S4 ; we use multiple <code>itsrdf:taIdentRef</code> values to denote an <i>AND</i> over the links	73
6.5	The main view of NIFify showing: (a) the class-reference input to filter annotations; (b) the document text input; (c) the mention identification field; and (d) the annotation visualization.	76
6.6	NIFify’s validation tree view for the mention “Tehran” in the ACE2004 dataset	80
6.7	Cumulative best-first progression of precision, recall and F_1 scores for Babelify (relaxed/strict), TagME, DBpedia Spotlight, AIDA and F _{REME} for the unified dataset considering combinations of categories [138]	85
6.8	α -based fuzzy F_1 scores for off-the-shelf systems	89

Chapter 1

Introduction

With the invention of the World Wide Web (WWW) in 1989, Tim Berners-Lee brought the world closer to the digital age, providing mechanisms to share documents through the Internet. More than five billion¹ websites have been published to date, being a source of valuable information that is continuously expanding. However, the initial proposal of WWW focused on serving information to people, not machines. Hence, the majority of web pages provide natural language content that is easy to understand by humans, but its automatic processing remains a challenge. In this context the concept of Semantic Web [72] emerged as an extension of the current document-based Web to a “Web of Data” with the goal of allowing the integration and understanding of heterogeneous sources. A key role in this ecosystem is played by RDF [19], a new standard data model that improves interoperability on the Web and allows the implementation of named relationships as well as hyperlinks. In this new scenario, the linking structure behind RDF defines a directed and labeled graph, which provides a better machine-readable representation of data on the Web.

The Semantic Web was warmly welcomed since its inception, and many projects were dedicated to the automatic generation of RDF resources from encyclopedias [142, 155, 156]. Another branch of researchers formalized a query language model for RDF, among them, RQL [80], OQL², DQL³, and SquishQL [105]. Although there was rapid adoption of the Semantic Web, pieces needed to achieve an interlinked Web of Data were still missing. To address this gap, the Linked Open Data (LOD)⁴ [12] principles were proposed, along with the 5 Stars of Linking Open Data, which outlined five steps (or practices) advocating for the release of data under open licenses, in a structured format, based on a non-proprietary format, following open standards (e.g., RDF) and having links to other datasets.

This new ecosystem forms the basis for enabling many new tasks; therefore, The Semantic Web is continuously growing and covering data in several domains. Along these lines,

¹<https://www.worldwidewebsite.com>

²<https://www.w3.org/RDF/Metalog/paper980828.html>

³<http://www.daml.org/dql/>

⁴<https://www.w3.org/DesignIssues/LinkedData.html>

Knowledge Bases (KB) [73] have gained the attention of many communities, helping to address problems from a semantic point of view. Numerous Knowledge Bases (KB) are now available online, including semi-structured KBs such as Wikipedia, and structured KBs such as BabelNet [110], DBpedia [86], Freebase [14], Wikidata [167], YAGO [133], etc. These KBs provide detailed descriptions of millions of entities – spanning multiple domains and languages – where each such entity is associated with a unique KB identifier. Often these KBs are made openly available on the web using the aforementioned Semantic Web standards. KBs are considered a valuable source of structured knowledge that facilitates data readability and expressiveness, often stored as RDF triples. These desirable properties are not explicitly present in natural language data, which is the most popular way to make claims on the Web, such as in social media posts, online newsletters, online books, scientific papers, and others.

A foundational task that makes a bridge between unstructured sources of data (text) and (semi-)structured sources of data (KBs) is Entity Linking (EL), which involves identifying entity mentions in a text (or potentially a semi-structured source [93]) and associating them with their corresponding unambiguous identifier in a KB. For example, given the input text “Michael Jackson was managed by his father Joseph Jackson” and DBpedia as a reference KB, an EL tool may identify “Michael Jackson” and “Joseph Jackson” as entity mentions, linking them to the DBpedia entities “`dbr:Michael_Jackson`” and “`dbr:Joe_Jackson_(manager)`”, respectively.⁵ Associating entity mentions with KB identifiers in this manner not only disambiguates the entities that the text speaks of, but also provides access to background knowledge from the KB about the entity, such as to know that “Michael Jackson” refers to a pop singer born in Gary, Indiana. In other words, EL allows one to take advantage of the full potential that is already implemented in the Semantic Web. EL can further form the basis for techniques performing more complex tasks, such as Semantic Search (e.g., to find documents about U.S. pop singers), Relation Extraction (e.g., to extract the binary relation `dbo:father(dbr:Michael_Jackson,dbr:Joe_Jackson)` from the previous text), Question Answering (e.g., to answer “who was Michael Jackson’s manger?”), among others [172, 93].

This thesis focuses on Entity Linking. Though it belongs to the Information Extraction field, it can be applied in the majority of those scenarios where one needs links, for example, in order to leverage the structured content in KBs for enriching or understanding text.

1.1 Problem Statement

Given the central importance of the EL task, a broad number of EL techniques and systems have been proposed in recent years [172]. The EL task can generally be sub-divided into two high-level sub-tasks [172, 93]. The first sub-task is *recognition*, where entity mentions in the text – e.g., “Michael Jackson” and “Joseph Jackson” – are identified. The second is

⁵We use prefixes as denoted in <http://prefix.cc/>.

disambiguation, where these entity mentions are associated with candidate entities in the KB, the candidates are ranked, and a single unambiguous identifier is chosen; for example, candidates selected for “Michael Jackson” in DBpedia might include:

- `dbr:Michael_Jackson`
- `dbr:Michael_Jackson_(radio_commentator)`
- `dbr:Michael.A._Jackson`
- `dbr:Michael_Jackson_(bishop)`
- ...

and so forth; the EL system must then rank these candidates and select the one it deems most likely to have been referred to by the text based on information available in the surrounding text, the KB, and potentially other reference sources. The main challenges of this task include the presence of multiple names for the same entity (e.g., “Joseph Jackson” vs. “Joe Jackson” vs. “Joe” referring to `dbr:Joe_Jackson_(manager)`) and multiple KB candidates for mentions (as seen for “Michael Jackson”).

While the previous challenges for EL are well-known, another more fundamental issue is often overlooked by the community: the question of *what is an “entity”*? Though several definitions have emerged about what an entity should be [58, 42, 162, 122], there is, as of yet, no clear consensus [15, 88].

This question has a major impact on EL research, leaving unclear which entity mentions in a text should be linked by EL systems or annotated by gold standards for evaluation purposes. To illustrate, Figure 1.1 shows an example text and the annotations produced by popular EL approaches: Babelfy [110], DBpedia Spotlight [101], FREDER [143], and Tagme [50]. Here we can see how these systems differ in their recognition of entities. Although most systems correctly recognize and link popular entity mentions like *Michael Jackson*, but for no entity mention do all systems agree. The fundamental question then is: *which annotations are “correct”*? The answer depends on how “entity” is defined. The notion of an entity may even vary across languages and cultures.

There is a distinction between two separate issues in the definition of an entity that can lead to misunderstanding:

- What entities in the KB should EL consider? There is agreement that it should include `wiki:Michael_Jackson`, but what about `wiki:Living_with_Michael_Jackson`, which is a documentary? Our understanding of EL places no restrictions on which KB entities we should link to. Therefore, there would be no reason not to consider `wiki:Living_with_Michael_Jackson` as a target KB entity.
- What mentions in a text should EL consider? There is agreement that it should include “Michael Jackson”, or “M.J.” or “King of Pop”. But should it include coreferences such

“he” , or descriptions such as “the inventor of the moonwalk”, or inner mentions such “Living with {Michael Jackson}”? According to our understanding, this is one of the reasons for the lack of consensus in the EL community, namely whether or not to mark a chunk of text as an entity mention for a further linking process.

There is arguably a third issue that combines both KB entities and mentions in the text:

- What types of reference EL should consider? For example, should “the Russian President” refer to `wiki:Vladimir_Putin`, or `wiki:President_of_Russia`, or both? Does it depend on the context?

This ambiguity affects further processing of EL systems’ outputs since different application scenarios have different requirements on what mentions should be involved. Furthermore, this problem affects various stages in the EL process. One such stage is the benchmark dataset selection process where, some works conservatively include only mentions of entities referring to fixed types such as person, organization and location as entities (similar to the traditional NER/TAC consensus on an entity), while other authors note that a much more diverse set of entities are available in Wikipedia and related KBs for linking, and thus consider any noun-phrase mentioning an entity in Wikipedia to be a valid target for linking [122]. Hence, applications that do not fit with one of these two main branches will have no suitable benchmark dataset, either for training, or evaluation. This problem also complicates EL assessment because we do not know how we can define the ideal result that such a system should achieve.

Some efforts have been made to standardize which mentions we should identify for annotation, as is the case of the work by Jha et al. [78], who propose a set of rules to serve as guidelines for benchmark creation. However, these rules force the adoption of some considerations that may not suit certain applications and on which there is thus no consensus. For instance, Jha et al., advocate for the omission of overlapping mentions like “{Michael Jackson}”, but authors such as Ling et al. [88] disagree. In a semantic search scenario, for example, looking at Figure 1.1, should such a document be considered relevant for a user interested in texts about Michael Jackson, or more generally, texts about American pop singers?

In this thesis, we pay special attention to efforts made to achieve multilingual EL approaches, their cross-lingual performance, and the impact when dealing with different types of entities. One of the obstacles to ongoing research on multilingual EL is a scarcity of annotated datasets with the same text in different languages. Multilingualism has been applied in EL, mainly supported by multilingual resources like Wikipedia that contain data in different languages. However, with the advance achieved in machine translation, its usage to address multilingualism in EL is promising and it is not much explored. Could we not simply focus on supporting one language in the EL system and translate the input text to that language?

In an [interview]^{td} with [Martin Bashir]^{btf} for the 2003 [documentary]^{td} [Living with {Michael Jackson}^{bd}]^{btf}, the King of [Pop]^d recalled that [Joe]^t often sat with a white belt at hand as he and his four [siblings]^{td} rehearsed.

Figure 1.1: Annotations of Babelify (b), DBpedia Spotlight (d), FREDER (f) and TagME (t) on the same sentence

In this direction, the following research questions are being addressed:

1. What should Entity Linking link?

RQ1a How can we define the goal of the EL task?

RQ1b Is consensus possible on the definition of an “entity”?

RQ1c If not, how can we define benchmark EL datasets and what metrics can we use to reflect the lack of consensus?

2. How well do EL systems perform in multilingual settings?

RQ2a How well do available EL systems do for languages other than English (as the most common primary language)?

RQ2b How does the performance of systems compare for multilingual EL?

RQ2c Why do results differ across languages?

RQ2d How would a method based on machine translation to English compare with directly configuring the system for a particular language?

1.2 Hypothesis

We believe that addressing these questions will help to unlock the full potential of EL tools for diverse applications, with diverse languages, diverse notions of entity, etc. The general hypothesis in this Ph.D. work is that when it comes to EL systems, one size does not fit all: different scenarios and different applications may have different requirements for an EL system, including, but not limited to, the types of entities targeted, the languages supported, etc. Along these lines, we define the following specific hypotheses:

1. Different EL systems consider different “entity” definitions, and thus target different sets of KB entities.
2. Current EL quality measures are not suitable for the evaluation of approaches that consider different “entity” definitions.
3. The majority of multilingual EL approaches behave in different ways for different languages.

4. Machine translation could be used in multilingual EL scenarios and reach/improve state-of-the-art multilingual EL approaches.

1.3 Research Goals

To answer the research questions and validate the hypothesis, we define here the focus of this Ph.D. work.

1.3.1 General Goals

The goal of this thesis is to perform finer-grained evaluation of EL systems under different requirements and different assumptions.

1.3.2 Specific Goals

To reach our general goal, we have identified the following specific research goals:

1. Understand how the goal of EL systems may vary across different applications and how that affects the consensus of what an “entity” is.
2. Consider how different EL systems perform for different languages, where we have published some results and proposed a novel dataset along these lines.
3. Compare the behavior of multilingual EL approaches when they are performed using their own multilingual model, and when they are set to work with English over an input text translated to English.
4. Propose a way to include/exclude entities per the application requirements. Our thought aims to propose a categorization of entity mentions which allows their separations.
5. Propose finer-grained evaluation protocols for EL according to previous findings, which address the lack of consensus.

1.4 Contributions

The standard approach to tackle EL has been to make certain design choices explicit, such as to enforce a particular policy with respect to overlapping mentions, or common entities, etc., when labeling an EL dataset or performing evaluation. However, the appropriate policy may depend on the particular application, setting, etc. This thesis pursues an alternative approach, which embraces different perspectives of the EL task, investigating ways to evaluate and support EL for multiple languages, further proposing a fine-grained categorization of different types of EL mentions and links. Specifically, our contributions are as follows:

- We design and build NIFify: a tool that simultaneously supports the creation, visualization, and validation of EL benchmark datasets.
- We publish the VOXEL dataset: a manually-annotated gold standard for EL considering five European languages, namely German, English, Spanish, French and Italian.
- We use VoxEL to study the EL performance using machine translation of the input to languages other than English.
- We design and present the results of a questionnaire addressed to authors of EL papers intended to understand the consensus (or lack thereof) regarding the goals of the EL task.
- We propose a fine-grained categorization scheme for the EL task covering details regarding base form, part of speech, overlap, and reference type.
- We relabel and publish three existing EL datasets – ACE2004 (subset), KORE50 and VoxEL – per our novel categorization scheme, extending the set of annotations as appropriate.
- We present the results of a fine-grained evaluation of the performance of five EL systems with respect to individual categories of EL annotations.
- To address the lack of consensus, we propose a fuzzy recall and F_1 measure based on a configurable membership function, presenting results for the five EL systems.
- We present conclusions about the performance of the EL systems surveyed for different types of entity mentions/links and highlight open challenges for the EL task.

As part of the contribution of this thesis, we make available source codes and different kinds of resources that are listed next and will be described in more detail later in the thesis:

VoxEL <<https://users.dcc.uchile.cl/~hrosales/VoxEL.html>>

A multilingual manually-annotated gold standard for EL. For each language, we create two versions of the annotations, one of them only with annotations of *person*, *organization*, and *places*; and another version with any possible annotation to Wikipedia, including concepts such a belt, chair, table, eyes, etc.

NIFify <https://github.com/henryrosalesmendez/NIFify_v4>

A standalone JavaScript application to manually create benchmark datasets for EL focused on the NIF format. Additionally, this tool supports the curation, visualization, and validation of EL annotations.

fel **vocabulary** <<https://users.dcc.uchile.cl/~hrosales/fel.html>>

An RDF-based vocabulary that gathers entity mentions into categories, and subcategories. In this line, the FEL vocabulary allows for making decisions about which entities to consider when evaluating EL systems, thus adjusting the evaluation to the corresponding scenarios.

CR Wrapper <<https://pypi.org/project/wrapperCoreference/>>

A Python package that provides a common platform for using Coreference Resolution (CR) tools. At the time of writing this thesis, it includes only models provided by Stanford CoreNLP [91].

WSD Wrapper <<https://pypi.org/project/wrapperWSD/>>

A Python package that provides a common platform for using Word Sense Disambiguation (WSD) tools.

NIF Wrapper <<https://pypi.org/project/nifwrapper/>>

A Python package for parsing and handling NIF annotations. This package includes all the novelties proposed in this thesis concerning the NIF format and provides methods for safely ingesting outputs from the CR Wrapper and WSD Wrapper packages.

FEL Benchmark <https://github.com/henryrosalesmendez/EL_exp>

All Python scripts created for fine-grained EL experiments in papers [138] and [137].

Survey and responses <<https://users.dcc.uchile.cl/~hrosales/questionnaire>>

A questionnaire to gain insights into what the community considers to be the goal of Entity Linking, along with the responses received. We present two sentences as examples which consider English Wikipedia as the target of Entity Linking. We then ask participants which annotations they think an Entity Linking system should output.

Reannotation of VoxEL, Kore50 and ACE04 <https://github.com/henryrosalesmendez/categorized.EMNLP_datasets>

This GitHub repository contains the reannotation of the datasets VoxEL, Kore50, and ACE04 following the FEL vocabulary.

1.5 Publications

The main contributions of this thesis have been presented in the following publications:

Journal paper:

- **Henry Rosales-Méndez**, Aidan Hogan and Barbara Poblete. Fine-Grained Entity Linking. *Journal of Web Semantics*, 2020.

Conference papers:

- **Henry Rosales-Méndez**, Aidan Hogan and Barbara Poblete. Fine-Grained Evaluation for Entity Linking. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019* pp 718–727

- **Henry Rosales-Méndez**, Aidan Hogan and Barbara Poblete. NIFify: Towards Better Quality Entity Linking Datasets. *Companion of The 2019 World Wide Web Conference, WWW 2019*. pp 815–818
- **Henry Rosales-Méndez**, Aidan Hogan and Barbara Poblete. VoxEL: A Benchmark Dataset for Multilingual Entity Linking. *International Semantic Web Conference, ISWC 2018*. pp 170–186

Workshop papers:

- **Henry Rosales-Méndez**, Aidan Hogan and Barbara Poblete. What should Entity Linking link? *Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management, AMW 2018*.
- **Henry Rosales-Méndez**, Barbara Poblete and Aidan Hogan. Multilingual Entity Linking: Comparing English and Spanish. *In the Proceedings of the Linked Data for Information Extraction, LD4IE 2017*. pp 62–73

Other papers:

- **Henry Rosales-Méndez**. Towards Better Entity Linking Evaluation. *Companion of The 2019 World Wide Web Conference, WWW 2019*. pp 50–55
- **Henry Rosales-Méndez**, Aidan Hogan and Barbara Poblete. Machine Translation vs. Multilingual Approaches for Entity Linking. *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference, ISWC 2018*

1.6 Thesis outline

This thesis is organized as follows:

Chapter 2 describes the preliminary concepts needed to better understand the content of this thesis. We start this chapter by highlighting some of the main definitions of “entity”, which is an essential concept for EL. Additionally, we formalize each phase that composes EL and the RDF representation often used to define KBs.

Chapter 3 includes the main approaches proposed to address each phase in EL. We discuss both NERL and End-to-End strategies separately. Finally, we review the main quality measures proposed to evaluate EL systems.

Chapter 4 highlights the elephant in the room, the lack of consensus about what Entity Linking should link. We provide clear evidence about the lack of consensus in response to this question, the possible negative consequences of such a lack of consensus, and outline our position on how to potentially address this issue. This chapter also describes the current benchmark datasets for EL and the main formats used for their definitions.

Chapter 5 is focused on EL approaches that deal with more than one language. We survey the main approaches and the EL datasets that cover more than one language. We propose a new parallel corpus called VoxEL and use it to evaluate existing EL systems. Additionally, we explore the use of machine translation in multilingual settings, translating from unsupported languages to English and then performing EL.

Chapter 6 proposes a set of fine-grained categories for benchmark annotation that allows the inclusion or exclusion of different types of entities as may be appropriate for different domains of application. We use this categorization in order to relabel some current benchmark datasets, and also, we propose a quality measure that considers the categories in order to achieve a fine-grained score according to the domain of application. With these re-labeled datasets and quality measures, we study the behavior of EL systems in fine-grained scenarios. On the other hand, given that NIF is now the most used format, we highlight some design issues of NIF, and we propose an extension of NIF to address them. Finally, we describe NIFify, our proposal for benchmark creation, visualization, and validation.

Chapter 7 explores how EL systems behave when their outputs are enriched with Coreference Resolution and Word Sense Disambiguation. We run experiments related to each category and measure the impact of including both techniques in an isolated and joint way.

Chapter 2

Preliminaries

As part of the continuous advances achieved in NLP, the identification of entities in unstructured textual data becomes more critical every day. Entities have a key role in sentences since they constitute a unit of information that serve several functions and are usually related to each other. Many tasks of NLP are dedicated to working with entities. Possibly one of the most popular is the Named Entity Recognition (ER) task, with the goal of identifying entities in unstructured textual data. On the other hand, the Relation Extraction (RE) task focuses on discovering their relationships. Coreference Resolution (CR) is also a widely explored task of NLP that searches for those words or phrases that refer to the same entity. Other NLP tasks concern all the words in sentences, for instance, being dedicated to identifying their part of speech (PoS) or disambiguating the sense of the words (Word Sense Disambiguation). Entities are also explored in many other areas, such as Information Extraction (IE), which focuses on the extraction of structured information from unstructured and/or semi-structured data.

In this context, the Entity Linking (EL) task emerges as a consequence of two main factors: the explosive increase in the amount of unstructured textual data, and simultaneously, the availability of large KBs, which describe billions of real-world entities. One of the first steps in this direction was the KIM Platform [127] in 2004, where researchers from the Ontotext Lab provided a semantic annotator to links words in a text to entities in the KIM Ontology (KIMO) [126]. In 2006, Bunescu et al. [22] target Wikipedia for EL for the first time, followed by many other researchers in a task coined *Wikification* in 2007 by Mihalcea et al. [102]. Later works, such as [125, 101, 109, 163, 71], would begin to target structured KBs described in the RDF format.

In this chapter, we begin by establishing some preliminaries that are central to the topic of this thesis. In particular, we first discuss definitions for “entity” as traditionally used in ER and EL tasks, and formalize the EL task itself. We further describe the RDF format and its query languages, as are commonly used to represent and answer questions over the structured KBs that modern EL tools commonly target.

While EL has a strong relation with Word Sense Disambiguation (WSD) where both

need to make the disambiguation process based on the context, the following differences make them two separate tasks:

- EL takes KBs as references, while WSD use lexicons such as WordNet.
- EL has to deal with the name variation of entities where several mentions could be linked to the same KB entry. In contrast, WSD supposes that all the synonyms are already contained in the lexicon.
- EL has to deal with multi-word mentions, while WSD focuses on the disambiguation of single words that appear in an input text.

2.1 Definition of “entity”

While the importance of entities are well-known, another more fundamental issue is often overlooked by the community: the question of *what is an “entity”*? Though several definitions have emerged about what an entity should be [58, 42, 162, 122], there is, as of yet, no clear consensus [15, 88].

For the 6th Message Understanding Conference [58] (MUC-6), the concept of “*named entity*” was defined as those terms that refer to instances of proper-name classes such as *person*, *location* and *organization*, and also, to numerical classes such as *temporal expressions* and *quantities*. Many *Named Entity Recognition* (NER) tools and training datasets/gold standards were developed to recognize and type entity mentions corresponding to these classes. However, researchers later became interested in Entity Linking (EL), where mentions were no longer simply recognized, but also linked to a reference KB (often using Wikipedia). Such KBs contain entities that do not correspond to traditional MUC-6 types so this definition was no longer exhaustive: in Figure 1.1, while the people and organizations would be covered under the MUC-6 consensus, the documentary “*Living with Michael Jackson*” would not; on the other hand, no system annotates “*2003*” from the MUC-6 class *Timex*.

Some authors have since defended the class-based proposal of MUC-6, incorporating new classes into the initial definition such as *products*, *financial entities* [108], *films*, *scientists* [44], etc. On the other hand, Fleischman [51] proposed to separate the classes into multiple specific subclasses (e.g., deriving *city*, *state*, *country* from the class *location*). Different processes and models can then be applied for different entity types. In general, however, such class-based definitions are inflexible, where at the time of writing, a KB such as Wikidata has entities from 50,000 unique classes, with more classes being added by users. Hence some authors have preferred more general definitions, but these often lack formality [42, 162].

Another point of view is to define an entity based on what is described by a knowledge-base; e.g., Perera et al. [122] define an entity as those described by Wikipedia pages with no ambiguity. While this avoids class-based restrictions and offers a practical, operational

definition for EL purposes, it too has issues. Entities are tied to a particular version of a KB, making it impossible to create general gold standards or to reflect *emerging entities* that may be added to the KB in future. Furthermore, Wikipedia has articles for general terms such as *documentary* and *belt*, though as per Figure 1.1, many tools and authors would not consider such terms as “entities”, but rather as being general words/concepts (and thus the subject of a different task: Word Sense Disambiguation (WSD)).

Even if we establish a clear definition for “*entity*”, we are still left to clarify what kinds of *entity mentions* should be recognized by EL. For example, all prior definitions agree that the singer *Michael Jackson* is an entity, but in the text of Figure 1.1, no definition clarifies whether or not an EL system should recognize and link the mentions *Jackson* (a *short mention*) and/or *he* (a *pronoun*) to the KB entity for *Michael Jackson* to which they refer; some authors, such as Jha et al. [78], consider this a task independent of EL called *Coreference Resolution* (CR), while others consider it part of EL to disambiguate entity types [41]. Furthermore, in the mention “*Living with Michael Jackson*”, some authors consider the inner *overlapping mention* of “*Michael Jackson*” as valid [110, 90]; others, such as Jha et al. [78], only consider the larger mention as valid.

2.2 EL Formalisms

Entity Linking is a task in Information Extraction that focuses on linking the entity mentions in a text collection with entity identifiers in a given knowledge base. Formally, let E be a set of entities in a KB and M the set of entity mentions in a given text collection. The EL process focuses on linking each entity mention $m \in M$ in a text collection with an entity identifier $e \in E$ in a given Knowledge Base (KB) [145]. Nowadays, there are large KBs that describe a huge list of entities (such as Wikipedia, DBpedia, Wikidata, etc.); furthermore, new entities emerge every day. Those mentions not (yet) included in the KB are labeled *NIL* (Not In Lexicon).

Generally speaking, EL models are commonly separated into two main phases, detailed below:

Entity Recognition (ER) This phase spots which phrases of the input text should be taken as mentions. This problem is also addressed by the Named Entity Recognition (NER) task, where a variety of techniques have been employed to this goal. On the other hand, some works regard ER itself as an independent task, out of the scope of EL [125].

Entity Disambiguation (ED) This phase decides which KB entities should be associated with the identified mentions. This phase is commonly divided into the following steps:

Candidate entity generation: For each entity mention $m \in M$ this stage selects E_m : a candidate set $E_m \subseteq E$ that represents entities with a high probability of cor-

responding to m is selected. Often this selection is based on matching m with entity labels for E in the knowledge base.

Candidate entity ranking: Each entity $e_m \in E_m$ is ranked according to an estimated confidence that it is the referent of the textual mention m . This can be performed considering a variety of features, such as the perceived “popularity” of e_m , its relation to candidates for nearby mentions, and so forth. The candidate in E_m with the best ranking may be selected as the link for m , possibly assuming it meets a certain threshold confidence (or other criteria).

Unlinkable mention prediction: Some tools consider unlinkable mentions, where no entity in the knowledge base meets the required confidence for a match to a given entity mention m . Depending on the application scenario, these mentions may be simply ignored, or may be proposed as “emerging entities” – annotated as NIL – that could be added to the knowledge base in the future.

Many techniques have then been proposed down through the years to address these sub-tasks [172, 93, 82, 25, 46]; we can distinguish two high-level strategies employed by different systems, which we term: *Named Entity Recognition & Linking (NERL)* systems [158] and *End-to-End Entity Linking (E2E)* systems [25, 94].¹

NERL systems decouple the recognition and disambiguation steps of the Entity Linking task [71, 70, 76, 59, 55]. Such systems apply recognition using an existing *Named Entity Recognition (NER)* system, the results of which are input into a separate disambiguation phase with respect to the KB. The NER task predates the EL task and involves identifying the named entities in a text (independently of a KB). A commonly-used convention for the entities targeted by NER systems, as previously discussed, was defined in the Message Understanding Conference 6 (MUC-6) [58], including those of type Person, Organization, Place, Numerical/Temporal and (sometimes) other Miscellaneous entities. NERL systems then typically apply existing NER tools (which have been developed over decades) to recognize entities in the text, feeding the results into a later disambiguation (ED) step.

Conversely, E2E systems apply recognition and disambiguation in a more unified manner. Rather than use an existing NER tool, a common E2E strategy is to attempt to directly match the labels of KB entities to substrings within the input text [50, 101, 110, 163], thus simultaneously recognizing entity mentions and KB candidates for disambiguation; mentions without confident KB candidates may further be filtered during disambiguation. In this way, the recognition and disambiguation sub-tasks can be combined and interleaved by E2E systems, further allowing – for example – for joint optimization models [90].

Both NERL and E2E systems present relative advantages and disadvantages. On one hand, NERL systems benefit from years of development on state-of-the-art NER tools, and

¹We remark, however, that the precise definitions vary from author to author, where we introduce the convention used here; e.g., Luo et al. [90] refer to E2E systems as *Joint Entity Recognition and Linking (JERL)*.

furthermore can identify *emerging entities* that do not (yet) appear in the KB. On the other hand, NER systems typically only identify mentions for a subset of entities that appear in KBs: returning to the sentence “Michael Jackson was managed by his father Joseph Jackson”, we find that DBpedia, Wikipedia, Wikidata, etc., have entities denoting “father” and “manager” that are not named entities and thus would be missed by NER tools; furthermore, in the sentence “Michael Jackson’s first studio album was Got to Be There.”, given the typical MUC-6 types targeted by NER tools, the album “Got to Be There” may not be detected although it is a named entity.² With a dataset such as Wikidata defining around fifty thousand entity classes, E2E systems will thus often detect a wider range of entities described by a KB than NERL systems [88]. Recognizing these relative strengths and weaknesses, hybrid [77] and ensemble [134] approaches propose to combine NERL and E2E results.

2.3 Resource Description Framework

Entity Linking has an inherent principle of providing a kind of structure to text where there is none. This notion of structure is bootstrapped by targeting a structured KB, leveraging all the relationships that the corresponding KB entities have. In this subsection we provide a better understanding of RDF which is a W3C³ standard that is used to represent some of the largest and most prominent structured KBs on the Web, such as YAGO, DBpedia and Wikidata, as well a many other small KBs⁴ that form part of the Semantic Web ecosystem.

The initial version of RDF was released in 1998 [19], followed by two updates: RDF 1.0 [18] in 2004, and RDF 1.1 [35] in 2014. Generally speaking, RDF structures data into triples `<subject,predicate,object>` meaning that the *subject* and *object* are two nodes of a directed graph related by the *predicate*. Both nodes are resources that are unambiguously identified with an URI (Uniform Resource Identifier) or can be left unidentified using blank nodes⁵; additionally, *object* nodes can be literals that store data-type information. The *predicate* – a.k.a *property* – is also a URI and denotes a relationship between the *subject* and *object*. Figure 2.1 shows a tiny RDF graph extracted from DBpedia, which models knowledge about Michael Jackson. We therefore know:

- who one of the parents of “Michael Jackson” is.
- “pop” format is one of the kinds of music he performed.
- that “Michael Jackson” stars in the documentary “Living with Michael Jackson”
- that “Martin Bashir” was the presenter of “Living with Michael Jackson”

²It is worth noting that there have been numerous proposals on how to diversify the entities recognized by NER tools, such as the proposal by Fleischman and Hovy [52] of a fine-grained classification of named entities; however, NER tools still predominantly follow MUC-6 definitions.

³<https://www.w3.org/>

⁴A list of RDF dumps is available in <https://www.w3.org/wiki/DataSetRDFDumps>

⁵For more information about blank nodes see <https://www.w3.org/TR/rdf-concepts/>

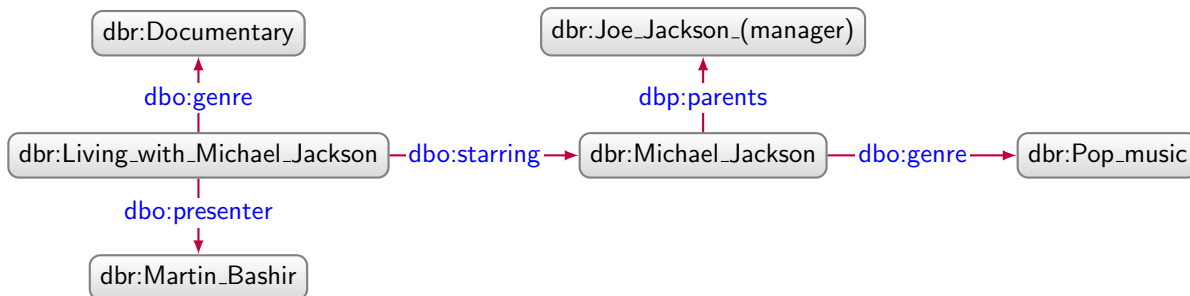


Figure 2.1: RDF graph extracted from DBpedia

RDF is not an isolated standard. In 2002, its extension, RDFS (RDF Schema) [17] was proposed: a data-modeling vocabulary for RDF data that allows the definition of custom properties, classes, and their relationships. RDFS also introduces **domain** and **range** to restrict the class of subjects and objects respectively. However, RDF/RDFS still lack expressive power. For instance, the conditions for membership/equivalence/disjointness of classes that can be expressed are limited. To fill these gaps, the Web Ontology Language [150] (OWL) – later updated to OWL 2 [2] – was standardized as a new layer on top of RDFS, and provides an extended set of classes and properties that enable more expressive reasoning.

With RDFS/RDFS/OWL, it is possible to create ad-hoc vocabularies that can model different domains, thus allowing the application of Semantic Web languages in real scenarios. For example, one popular vocabulary is FOAF⁶, which contains definitions to represent personal information in the Web, including links to other known people. Another example is SKOS [103], a vocabulary proposed for modeling the basic structure of concept schemes, e.g., thesauri, taxonomies, classification schemes, subject heading lists, and others. Both of these vocabularies are defined in terms of the RDFS and OWL standards.

There are various syntaxes proposed for the serialization of RDF graphs. Historically, RDF/XML [84] was the first W3C standard for serializing RDF graphs and is based on the well-known XML format. Nodes and predicates are represented in XML terms, starting with the root element `<rdf:RDF>`, followed by a recursive list of XML elements that store its triples. In the case of predicates, this serialization defines namespaces to organize and generate short definitions. In Figure 2.2 (a) we present an RDF/XML representation corresponding to the graph from Figure 2.1. In this short example, **Michael Jackson** is the first element, and it is associated with three predicate–object pairs. However, some researchers [74, 97, 23] have criticized RDF/XML mainly when large KBs are involved, stressing that RDF/XML is very verbose, not very easy to read by humans, and inherits all the disadvantages that come with trying to represent a graph in a way that is compatible with the hierarchical nature of XML.

Some new serialization formats have been proposed to overcome these RDF/XML drawbacks. One of them is the proposal of the non-XML serialization N3 [11], which goes

⁶<http://xmlns.com/foaf/spec/>

(a) RDF/XML

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:ns0="http://dbpedia.org/property/"
  xmlns:ns1="http://dbpedia.org/ontology/">

<rdf:Description
  rdf:about="http://dbpedia.org/resource/Living_with_Michael_Jackson">
  <ns0:genre rdf:resource="http://dbpedia.org/resource/Documentary"/>
  <ns0:presenter rdf:resource="http://dbpedia.org/resource/Martin_Bashir"/>
  <ns0:starring>
    <rdf:Description rdf:about="http://dbpedia.org/resource/Michael_Jackson">
      <ns1:parents
        rdf:resource="http://dbpedia.org/resource/Joe_Jackson_(manager)"/>
      <ns0:genre rdf:resource="http://dbpedia.org/resource/Pop_music"/>
    </rdf:Description>
  </ns0:starring>
</rdf:Description>

</rdf:RDF>
```

(b) Turtle

```
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/>

dbr:Michael_Jackson dbp:parents dbr:Joe_Jackson_(manager);
  dbo:genre dbr:Pop_music.

dbr:Living_with_Michael_Jackson dbo:genre dbr:Documentary;
  dbo:presenter dbr:Martin_Bashir;
  dbo:starring dbr:Michael_Jackson.
```

Figure 2.2: Serialization of the same RDF graph using the (a) RDF/XML and (b) Turtle formats

beyond RDF by including variables and nested graphs [10]. Among its advantages, N3 gains in expressiveness and achieves a readable notation. It also introduces URI abbreviations as prefixes, which contribute to a more compact serialization. For instance, let **dbr** be the prefix of URI `<http://dbpedia.org/resource/>`, and **dbo** be the prefix of URI

<<http://dbpedia.org/ontology/>>; we can represent the knowledge “The pop is one of the kind of music that Michael Jackson sang” using the following N3 statement:

```
dbr:Michael_Jackson dbo:genre dbr:Pop_music.
```

The Turtle serialization format [9] is a subset of N3 focusing only on RDF graphs but keeping all the syntactic conciseness that N3 provides. Although other variations of N3 have been proposed, such as N-Triples [8], Turtle is still widely used in the community. In Figure 2.2 we show how the serializations of the graph from Figure 2.1 look, in (a) RDF/XML and (b) Turtle where we can perhaps appreciate the benefits that Turtle provides in terms of its human-readability and compactness. However, various authors stressed that the aforementioned serializations are still verbose and propose serializations focused in compression techniques to reduce the storage space, such as HDT [49] and HDT-FoQ [92]. Another branch of authors has looked for convergence between RDF and HTML applications, promoting the development of new serializations, such as RDFa [1], JSON-LD [153], and others. RDFa is a semi-structured representation that has been recently proposed as a W3C standard, and consists of markup annotations embedded in HTML pages that allow for specifying semantic information. On the other hand, JSON-LD is proposed to handle data in web applications extending JSON with the incorporation of semantic information.

2.4 RDF query languages

When RDF was proposed, the need for a query language for RDF likewise arose. A plethora of proposals have been presented since RDF, where the QL'98 workshop⁷ hosted a hub of discussion on this topic. Some initial proposals followed a navigational approach, where languages exploit the XML structure of RDF/XML representations such as XPath [29], and XQuery [13]. Other approaches incorporate the XPath language as their foundation, adding new features on top to gain expressiveness in the queries.

Another group of proposals focused on the proposition of a more human-understandable query language that followed an SQL style. Many authors have summarized and compared the state-of-the-art of query languages for RDF [62, 63, 54]. We next present a brief exemplification of some popular query languages that can be used over RDF according to the categorization provided by Dave Beckett in [7]:

Using XML : XSLT [81], XPath [29], XQuery [13], XQueryX [100]

XPath-like : Versa [119], RPath⁸, FSL [123], RDF Twig⁹

SQL-like : RDQL/Squish [144], SeRQL [63], rdfDB QL [128], RQL [26], SPARQL [129]

⁷<https://www.w3.org/TandS/QL/QL98/>

⁸<http://www.xulplanet.com/ndeakin/arts/rpath-fns.txt>

⁹<https://norman.walsh.name/2004/projects/rdf Twig>

Rules-like : N3QL [54], Xcerpt [21], Triple [149], DQL, OWL-QL

Language-like : Algae2, Fabl¹⁰

Of all these proposals, SPARQL [129] is (in modern times) the most popular language for querying RDF, and became the official W3C recommendation for querying RDF. The SPARQL syntax is similar to SQL, which facilitates its comprehension to SQL users; on the other hand, it uses triple patterns to define graph conditions that make the language more natural for querying RDF data.

```
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/>

SELECT ?singer WHERE {
  ?singer dbo:genre dbr:Pop_music.
  ?documentary dbo:starring ?singer.
  ?documentary dbo:genre dbr:Documentary.
  ?documentary dbo:presenter dbr:Martin_Bashir.
}
```

Figure 2.3: Example of an SPARQL query

Figure 2.3 shows an SPARQL query that responds to the question “*what is the name of the pop singer who starred in a documentary presented by Martín Bashir?*” This example defines two variables: `?singer` and `?documentary`, which will match with those nodes that fulfill the triple patterns, but only the values of `?singer` is projected in the response.

¹⁰<http://fabl.net/>

Chapter 3

Related Work

Nowadays, there is a large variety of EL approaches. In this chapter, we gather many techniques that have been used to tackle each stage of the EL process: entity recognition, entity candidate generation, entity ranking, and unlinkable mention prediction. Additionally, we include the important stage of EL evaluation; while it does not belong to the core process of EL, it is a key factor to quantify and obtain progress in this task.

3.1 Entity Recognition

In the conception of EL as a task, one of the debates was concerned with the incorporation (or not) of ER as a part of the EL core. While some initial proposals advocate for keeping ER separated from EL’s formalism (with some authors [40, 66] arguing that EL should focus only on disambiguation and not recognition, which is addressed by NER), nowadays, there is no doubt about its inclusion. Indeed, ER phases are the entry point for NERL approaches, which directly affect the entire EL process’s quality.

The disagreement about what is an “*entity*” has a direct impact on the ER phase. Initial definitions (e.g. MUC-6 definition) advocate for gathering entities according to their entity type, and including only those entities that fall in the categories of *person*, *location*, and *organization*. However, authors have focused mainly on ranking stages and adopting different ER techniques that cover the group of entities they need. As a result of this disagreement, some authors released APIs where users could specify which entities they want to link. For instance, AGDISTIS¹ [163] proposed to enclose those selected entity mentions with brackets [...] in the input text. However, as aforementioned, most modern EL tools include an ER phase.

¹<https://agdistis.demos.dice-research.org/>

Michael	Jackson	was	managed	by	his	father	Joseph	Jackson
NNP	NNP	VBD	VBD	IN	PRP	NNP	NNP	NNP

Figure 3.1: Example of Part-of-Speech tags

3.1.1 NERL strategy

The majority of EL approaches follow a NERL strategy, where Entity Recognition is the first step to deal with. Commonly, authors choose to perform this phase with external NLP systems, leveraging the progress achieved in the Entity Recognition Task’s scope. One of the most used ER tools in EL environment is the well-known Stanford NER², which has been used by popular EL approaches such as AIDA [71], CohEEL [59], and Weasel [160]. Other approaches combine more than one external ER tool; this is the case of FOX [152], which is based on Stanford NER, Illinois NER [131], Balie [112], and OpenNLP³. Derczynski et al., in [38] study the behavior of these tools in social media scenarios and show that their accuracy is lower than 50% F_1 for Twitter messages.

On the other hand, lower-level Part-of-Speech (POS) tools have also been successfully employed to recognize entities. For a given text, these techniques identify the grammatical category of its words. For instance, we can extract with POS tools⁴ the POS information corresponding to each word of the sentence “Michael Jackson was managed by his father Joseph Jackson” as shown in Figure 3.1 where NNP means proper nouns, VBD means verbs, IN means prepositions, and PRP means pronouns. Many authors leverage POS techniques to identify mentions, commonly associated with a continuous sequence of noun words, or combined with prepositions. Babelfy [110] exploits POS – using the Stanford POS tagger [159] in some cases – to extract the mentions with at most five words where at least one of them should be tagged as a noun. Some POS tools are proposed to deal with noisy environments – such as social media messages – that usually include the processing of special characters such as hashtags (#) and the at symbol (@). One of the most used is the ARK Twitter Part-of-Speech Tagger [57], for example, which is the foundation of the ER phase proposed by Ghosh et al. [56].

A variety of other approaches have addressed the ER stage using machine learning techniques, as well as some external ER tools. The popular Stanford NER is a Conditional Random Field (CRF) classifier trained using CoNLL 2003 dataset. In this same aim, Zhang et al. [177], address the ER stage by applying an SVM classifier trained with the ACE 2005 dataset achieving 88.2% F_1 .

Within the ADEL system, Plu et al. [125] propose a different approach that consists of six different ER techniques wrapped in modules that can be included or excluded from the EL

²<https://nlp.stanford.edu/software/CRF-NER.shtml>

³<https://opennlp.apache.org/>

⁴For this example, we use the tagger available in <https://parts-of-speech.info/>

process. Among them, *Date Tagger* and *Number Tagger* deal with the recognition of number and temporal expressions; a *POS Tagger* extracts singular and plural proper nouns; an *NER Tagger* extractor is based on external ER tools; a *Gazetteer Tagger* incorporates external frameworks such as GATE⁵ and RegexNER⁶. ADEL’s approach not only contributes to the improvement of EL performance but also with achieving a model that is adaptable to the application environment.

3.1.2 End-to-End strategy

NERL approaches start with the ER stage; therefore, any mistakes made at this stage are carried throughout the EL process. This problem was first highlighted by Guo et al. [60] in the social media scenario, where the entity recognition is not accurate due to noise, typographical errors, and the inclusion of special characters (e.g., # and @) in social media messages. However, this same observation was made by some other authors outside the social media context [148, 5]. Therefore, the E2E strategy emerges to address this situation, with the idea of not using “off the shelf” NER but rather using custom techniques adapted for the targeted KB. In general, the main reasons to chose an E2E approach instead of a traditional NERL strategy are:

- Errors produced in ER phase will be propagated to the rest of the EL process, and are not recoverable [60, 148, 90].
- The ER stage does not benefit from the information coming from KBs used in the rest of the phases [90]. For instance, given a piece of text that includes “The New York Times”, ER tools could return “New York Times” as the entity mention without “The”; however, with the prior information of the KB, entities can be recognized with the full mention (including “The”) due to there being a Wikipedia page with this sequences of words in its title.
- ER and ED phases may yield inconsistent outputs [90]. With the information of KBs one can rather align patterns to obtain mentions accordingly. To exemplify this fact, Luo et al. [90] give the example of the occurrence of “George Washington” in a text, where ER tools recognize only “Washington”, but an ED phase could recognize the entire entity mention due to the context of the sentence and the appearance of an entity with that name in the KB.

Some E2E approaches behave similarly to how NERL approaches behave. For instance, Guo et al. [60], assume that entity mentions should match with anchor phrases in Wikipedia. Therefore, they extract all n-grams with size $\leq k$ and search for them using a gazetteer to generate the candidates. This procedure is similar to what NERL approaches do with a gazetter-based ER stage except that the gazetter is based on the KB itself, rather than (for example) a generic list of common first names and last names, organizations, places, etc.

⁵<https://gate.ac.uk/>

⁶<https://stanfordnlp.github.io/CoreNLP/regexner.html>

Other authors identify entity mentions using string matching techniques, commonly with the prior construction of gazetteers⁷ containing the target words to be searched. The majority of these approaches base the gazetteer generation on the KB, following a E2E strategy. DBpedia Spotlight [101] constructs a gazetteer with the labels from each DBpedia resource and its redirect and disambiguation pages. In this scenario, they perform an Aho-Corasick’ based method for ER. Similarly, ExPoSe [120] creates a gazetteer using the anchor text from the pages of Wikipedia, redirects, and disambiguation pages. They also opt for a case-insensitive variant of the Aho-Corasick algorithm to identify the entity mentions.

It is not clear the maximum number of words that compose an entity mention. While Babelfy handles a maximum number of five continuous words, each proposal could define a different quota, as is the case of the system developed by Yamada et al. [173] that considers mentions with up to 10 words. The construction of gazetteers also varies depending on each technique, i.e., NERFGUN [65] is based on one gazetteer constructed over DBpedia data and the Wikipedia anchor text.

Other approaches opt for jointly classifying mentions and links. Such an approach was proposed by Sil et al. [148], who create a classifier that includes in the link prediction the start and end position of entity mentions. The authors propose other features that help the classifier learn about when to consider a sequence of words as one mention. These include features to highlight when words are capitalized, how often an entity mention sequence links to the same entity, how many words match exactly with one of the names of a KB entity, etc. To include a high-recall set of entity mentions, they use a way of overgenerating them by combining more than one ER technique. In this same direction, Luo et al. [90] propose a classifier-based E2E approach, but base their approach on the inclusion of features for the entity type and *confidence information*.

3.2 Entity Disambiguation

3.2.1 Candidate Entity Generation

What would happen if, for each entity mention, we apply a ranking algorithm among all the entities in a KB? This is not a time-optimal scenario. EL commonly targets large KBs, where English DBpedia contains around 400 million⁸ and Wikidata more than 11 billion⁹ RDF triples. SPARQL endpoint implementations, such as Virtuoso, allow quick access to data in RDF format using SPARQL queries. For example, with the query shown in Figure 3.2, we obtained the information on how many triples DBpedia and Wikidata have in just 12.4s and 240ms, respectively. However, the retrieval of a list of triples also includes many other factors that considerably increase the time complexity, such as the size of data to transmit from its source to the users, the Internet bandwidth for remote endpoints, the

⁷<https://gate.ac.uk/sale/tao/splitch13.html>

⁸Specifically, DBpedia has 438,336,271 triples by September 19, 2020.

⁹Specifically, Wikidata has 11,534,038,938 triples by September 19, 2020.


```
SELECT (count(*) as ?c) WHERE ?s ?p ?o
```

Figure 3.2: SPARQL query to obtain the number of triples in a RDF KB.

cost of applying query processing operators such as joins, and others. Hence, it is necessary to incorporate an intermediate step between the ER and the Entity Ranking. For each entity mention, this step should select – in an efficient way – which are the KB entities that could correspond to them. In this way, further stages can then focus in on a small number of KB entities instead of the entire set of KB entities.

Many KBs have already implemented this logic, providing a mechanism to search using Web APIs for KB entities (or pages) that corresponds to user-specified keywords. For instance, Wikipedia allows for searching pages that match with keywords from its user interface, as well as from the Wikipedia Search APIs¹⁰. These mechanisms return an already ranked list of pages. Therefore, taking the entity mentions as a keyword for searching, Wikipedia Search is aligned to what Candidate Generation and Ranking is looking for. Some authors – such as Dojchinovski et al. [39] – take these implemented search methods as its candidate generation stage. In this line, they select the top- k retrieved Wikipedia pages as the candidate set of entities for each entity mention, where the value of k is tuned to each approach.

Under the hood, the Wikipedia Search engine¹¹ uses ElasticSearch¹²: a search engine server based on the Lucene library. This same technique is re-implemented by other authors creating their own custom index and engine. In this line, Guo et al. [60] conduct one of the most straightforward approaches, first creating an index based on the anchor texts of Wikipedia. Mendes et al. [101] offer another approach, which takes all the links from Wikipedia pages and extracts the anchor text used to link their corresponding pages. Hence, they compute a conditional probability $P(r|m) = P(m,r)/P(m)$ to obtain a ranked list of candidates for each surface form, where $P(m,r)$ is the number of times that the mention m is an anchor text that targets the Wikipedia page r , and $P(m)$ is the number of all the occurrences of m . In this line, many other approaches use an index and a search engine to retrieve the candidate set of entities, including NERFGUN [65], ADEL [125], AGDISTIS [163], and DoSeR [180]. NERFGUN retrieves the top-10 candidates over an index constructed by Wikipedia anchor text and DBpedia. On the other hand, AGDISTIS first applies pre-processing to each entity mention, removing plural and genitive forms, common affixes, and mentions that contain numbers. Finally, AGDISTIS uses a 3-gram similarity to select the candidates over a custom index.

Previous approaches tackle general application scenarios; however, there is no silver bul-

¹⁰<https://www.mediawiki.org/wiki/API:Search>

¹¹<https://en.wikipedia.org/wiki/Help:Searching>

¹²<https://www.elastic.co>

let to generate candidates. Some works, as [47] and [178], stress this problem, proposing solutions to specific scenarios that require a different procedure. Zhou et al. [47], highlight the need for techniques that deal with scenarios with few resources available, for example, targeting versions of Wikipedias that correspond to languages such as Romani¹³ and Gothic¹⁴ that contain fewer than one thousand pages. On the other hand, Fang et al. [178], highlight that techniques based on sequence models ignore the relevance between the current mention and its subsequent entities. In this line, they propose a method to generate high recall candidate sets based on the following three strategies:

- *The surface form of mention*: The authors use a combination of two techniques: (a) an online gazetteer to find exact and partial entities that match with mentions; (b) requests to the Wikipedia Search API.
- *A semantic extension*: They use WordNet to search also for the synonyms of mentions.
- *Exception handling*: The authors assume that the text could be misspelled. Hence, they also use Google Search Engine when the mentions contain more than three words.

These three techniques over-generate the candidate sets of entities; therefore, a pruning step is proposed to remove unrelated candidates using a classifier.

3.2.2 Candidate Entity Ranking

This stage is where the disambiguation takes place. Taking the output of the ER and Candidate Generation phases, this stage decides which candidate KB entity most likely correspond to each mention. There is a plethora of techniques proposed to rank the candidate set of entities, where many of them can be grouped according to their behavior.

The first techniques used were mainly based on measuring the probability that a particular mention links to a page on Wikipedia, commonly using anchor text. For instance, here we can find popular approaches such as TagME [50] and DBpedia Spotlight [101]. In fact, many approaches generate initial rankings while generating candidates. EL approaches based on searching in documents are typically ordered according to Information Retrieval (IR) measures, such as TF-IDF, which can then be used, in turn, as an initial ranking of candidates.

Another group of approaches are founded on machine learning models and classify whether a mention should link a specific KB entity. One of the most popular classifiers used with this goal is SVM, being employed by approaches such as Guo’s proposal [60], OpenTapioca [37], and Weasel [160]. A last wave of approaches turns their focus to neural network models, taking advantage of the progress achieved with mention, entity, and graph embedding. For instance, E2E-NN [82] is an End-to-End approach based on a bidirectional LSTM.

¹³https://en.wikipedia.org/wiki/Romani_language

¹⁴https://en.wikipedia.org/wiki/Gothic_language

Many approaches take advantage of the graph structure of KBs. AIDA [71] searches for the densest subgraph that involves mentions and candidate entities. On the other hand, AGDISTIS uses a graph-based score to rank the candidates. Next, we review in detail the ranking stage of the most prominent EL approaches:

Wikify! (2007) proposes a voting mechanism between a machine learning model and a technique that measures the overlap between the contexts of mentions and candidates [102].

TagME (2010) uses analyses of anchor texts in Wikipedia pages to perform EL [50]. The ranking stage is based primarily on two measures: *commonness*, which describes how often an anchor text is associated with a particular Wikipedia entity; and *relatedness*, which is a co-citation measure indicating how frequently candidate entities for different mentions are linked from the same Wikipedia article. TagME is multilingual: it can take advantage of the Wikipedia Search API to apply the same conceptual process over different language versions of Wikipedia to support multilingual EL.

AIDA (2011) creates an undirected and weighted graph with mentions and candidates as nodes. This graph is weighted in two different ways: measuring the relation between mentions and candidates with a combination of popularity and similarity measures; and measuring the overlap between the Wikipedia links of two candidates. The final step searches for the densest subgraph that has only one edge for each mention, which corresponds to the disambiguation links.

THD (2012) is based on three measures [39]: *most frequent senses*, which ranks candidates for a mention based on the Wikipedia Search API results for that mention; *co-occurrence*, which is a co-citation measure looking at how often candidate entities for different mentions are linked from the same paragraphs in Wikipedia; and *explicit semantic analysis*, which uses keyword similarity measures to relate mentions with a concept. These methods are multilingual and applicable to different language versions of Wikipedia.

DBpedia Spotlight (2013) was first proposed to deal with English annotations [101], based on keyword and string matching functions ranked by a probabilistic model based on a variant of a TF-IDF measure. An extended version later proposed by Daiber et al. [36] leverages the multilingual information of the Wikipedia and DBpedia KBs to support multiple languages.

AGDISTIS (2014) bases its ranking stage on a disambiguation graph, which is initially created from candidate entities, and next, expanded with related KB entities in a fixed number of iterations [163]. The authoritative score of the HITS algorithm is used to rank this graph's nodes. Moussallem et al. [111] propose a multilingual extension of AGDISTIS incorporating language-independent features.

Babelfy (2014) performs EL with respect to a custom multilingual KB BabelNet¹⁵ constructed from Wikipedia and WordNet, using machine translation to bridge the gaps in information available for different language versions of Wikipedia [110]. Recognition is based on POS tagging for different languages, selecting candidate entities by string matching. Ranking is reduced to finding the densest subgraph that relates neighboring entities and mentions.

S-MART (2015) [174] is a tree-based structured learning approach based on multiple additive regression trees. This system is also applied to Social Media domains where positive results were obtained.

FREME (2016) delegates the recognition of entities to the Stanford-NER tool, which is trained over the anchor texts of Wikipedia corpora in different languages. Candidate entities are generated by keyword search over local indexes, which are then ranked based on the number of matching anchor texts in Wikipedia linking to the corresponding article of the candidate entity [143].

WikiME (2016) uses a model based on word embedding, which includes a final step for projecting each non-English language embeddings to the English one.

E2E-NN (2018) is an End-to-End neural network model that relies on word, entity and mention embeddings [82]. For each entity-candidate pair (m, e) , a context-aware compatibility is used to rank the set of candidate entities.

OpenTapioca (2019) uses an SVM classifier that predicts if a mention m should be linked to the KB entity e . For each pair (m, e) , this approach computes features that indicate how often both are related in Wikidata (local compatibility) and how similar their topics (semantic similarity) [37] are.

Martins et al. (2019), propose an EL approach based on an LSTM augmented with a stack pointer [94]. In this way, mentions are classified as soon as they are identified.

PNEL (2020) uses a single layer bi-LSTM pointer network model with pre-computed TransE entity embedding over Wikidata [5] which implicitly contains the entire KG structure information. The layer used is composed of 512 hidden units and an attention size of 128.

3.2.3 Unlinkable Mention Prediction

In ideal scenarios, the targeted KB will contain an entity for each identified mention; however, this is not possible due to the following factors:

- Entities are continuously emerging with books, films, organizations, etc; There is often a period of time between when an entity emerges until it is registered in a KB, which

¹⁵<http://babelnet.org/>; April 1st, 2018

may depend on how the KB is constructed and updated, and also what sorts of entities are involved.

- KBs typically contain only popular entities that might be of interest to some communities or large numbers of people. However, many entities that could be present in a document of interest do not meet this condition, such as secondary characters in books or ordinary people we cross daily.

For these reasons, many authors include a last step in the EL process to detect those entities that are not present in the targeted KB. These mentions are called by different ways, including “unlinkable mention” [146], “Not In Lexicon” (NIL)[120, 161, 130, 94], “emerging entities” [125, 69], and “out-of-KB entities” [170].

Unlinkable mention identification is constrained by the way that the different EL approaches perform the ER stage. Unfortunately, those approaches with a gazetteer-based ER stage are biased by the construction of the gazetteer. Generally, gazetteers are generated based on the entities from the target KB. Therefore, there is no way to spot any unlinkable mention.

While some authors do not include unlinkable mention identification, numerous EL approaches do. For instance, WikiME [161] identifies such mentions in the candidate generation stage; namely, they tag as NIL all mentions that do not have corresponding candidate entities. Other approaches combine the NIL prediction with the ranking of candidates. Among these approaches, Rao et al. [130] train an SVM classifier merging both ranking and NIL-prediction features.

3.3 Entity Linking Evaluation Measures

The evaluation of EL approaches also remains a challenge. It is commonly conducted by comparing the system output S of an EL system and the desired output G . The desired outputs are commonly built by expert humans and are known as gold standard, ground truth, and benchmark datasets. The construction of benchmark datasets are commonly conducted manually, as was done for AIDA/CoNLL [71], MEANTIME [108], and others. However, Ngonga Ngomo et al. propose with BENGAL [117] to build such datasets automatically from RDF data. Section 3.4 is dedicated to surveying the main benchmark datasets, as well as to detail some aspects of their construction.

While E2E approaches have to evaluate the entire EL process due to their nature of combining both ER and ED, some authors also propose to measure ER and ED using isolated assessments for NERL approaches [125, 132, 28, 161]. In this way, the ER assessment indicates how the entire EL process is affected by this stage. In both cases, the most popular measure is the traditional F_1 measure [179], which was initially proposed in the area of Information Retrieval, but applied to many other areas. As shown in Equation 3.1, F_1 is computed as the harmonic mean between two criteria: precision (P) and recall (R). One

performs analysis from system outputs to benchmark datasets, and the other one in inverse order.

$$F_1 = 2PR/(P + R) \quad (3.1)$$

Precision and recall – defined in Equations 3.2 and 3.3 – are based on a binary confusion matrix that reports the number of true positives (tp), false positives (fp), false negatives (fn), and true negatives (tn).

$$P = tp/(tp + fp) \quad (3.2)$$

$$R = tp/(tp + fn) \quad (3.3)$$

The above definition of F_1 is centered on binary measurements involving the evaluation of two sets of annotations; for instance, in this case, we wish to evaluate the annotations from one sentence versus the correct annotations provided by a benchmark dataset. However, there are various natural ways to organize natural language text: sentences, paragraphs, documents, etc. One of the most popular ways to organize annotations in EL evaluation is by documents. The micro- and macro-average F_1 are aggregations of the traditional F_1 proposed to deal with multi-class environments, and thus, with the EL evaluation per document. Micro- and macro-average F_1 are shown in Equation 3.6, 3.7 and 3.4, 3.5 respectively [32], where by D we denote the set of documents in the benchmark dataset, by mP and mR we denote the micro precision and recall, and by MP and MR we denote the macro average and precision. The corresponding micro and macro F_1 scores can then be computed from the corresponding precision and recall scores using Equation (3.1).

$$mP = \frac{\sum_{d \in D} tp_d}{\sum_{d \in D} tp_d + fp_d} \quad (3.4)$$

$$mR = \frac{\sum_{d \in D} tp_d}{\sum_{d \in D} tp_d + fn_d} \quad (3.5)$$

$$MP = \frac{\sum_{d \in D} P_d}{|D|} \quad (3.6)$$

$$MR = \frac{\sum_{d \in D} R_d}{|D|} \quad (3.7)$$

Another well-accepted measure in binary classification, but used in many other areas including EL, is Accuracy. It is computed as $A = (tp + tn)/(tp + fp + fn + tn)$, measuring the portion of cases that have been linked correctly. Bagga and Baldwin propose BCubed [3], another measure that has been inherited by EL from related areas. Let $S(e)$ be the documents in the system output to which the entity belongs, let $G(e)$ be the documents in the benchmark dataset (Gold Standard) to which the entity belongs, and let $E(e, H)$ be the set of entities

co-occurring in set H (i.e., S or G) with e in at least one candidate cluster. BCubed (B^3 for short) is defined replacing precision and recall measures of F_1 as shown in Equations 3.8 and 3.9 respectively.

$$PB^3 = \frac{1}{|U|} \sum_{e \in U} \frac{1}{|\bigcup_{g \in S(e)} g|} \sum_{e' \in E(e,S)} \psi(e, e') \quad (3.8)$$

$$RB^3 = \frac{1}{|U|} \sum_{e \in U} \frac{1}{|\bigcup_{g \in G(e)} g|} \sum_{e' \in E(e,G)} \psi(e, e') \quad (3.9)$$

Where $\psi(e, e')$ is the function called *correctness* that yields a score 1 if both entities belong to the same document in the system output, and at the same time, belong to the same document in the benchmark dataset. An extension of BCubed, referenced henceforth by B^3+ , also requires the inclusion of the correct entity links. Although F_1 and B^3+ are the most used, several other measures are still emerging for this task.

3.4 EL Benchmark Datasets

Benchmark datasets are a key factor for comparing different EL systems and for measuring incremental progress in terms of performance on the task. Numerous datasets have been proposed down through the years to evaluate EL systems. These datasets are often built by human experts who indicate the correct annotations from a text corpus that an EL system should obtain – i.e., who provide a gold standard for the EL task. EL systems can then be evaluated against these gold standards using metrics such as precision, recall, and F_1 ; such results can be presented separately for the recognition and disambiguation phase in NERL systems, as well as for macro (averaging results across different documents) as well as micro (concatenating all documents into one) variants (see Section 3.3). Evaluation benchmarks such as GERBIL [164] then allow for computing and visualizing such measures with respect to different EL datasets and systems. We now discuss benchmark datasets for EL, as well as the formats and criteria they use.

In Table 3.1, we provide a brief overview of existing EL datasets [140, 93]. We see that a selection of datasets have been proposed, where most have been manually labeled; note that most marked **X** were previously NER datasets to which KB links were added, with one exception being DBpedia Abstracts [20], which is based on Wikipedia hyperlinks and anchor text. We further see that relatively few systems provide details on the entity type. We also see that a selection of formats (described later) have been used to serialize these datasets. Of note is that many of these datasets were created with particular purposes in mind; for example, SemEval2015 Task 13 [109], DBpedia Abstracts [20], and MEANTIME [108] were designed specifically for evaluating multilingual EL systems, providing annotated texts in multiple languages. On the other hand, KORE50 [70] is intended as a succinct but challenging collection of highly-ambiguous entities in short sentences. Furthermore, DBpedia Abstracts [20] is intended for the purposes of training multilingual EL systems. Further

Table 3.1: Popular EL datasets (ordered in terms of recency) indicating whether or not all labels were manually annotated, whether or not entity types were provided, as well as the format used for representing the dataset

Dataset	Manual	Types	Format
MSNBC [33]	✗	✗	MSNBC
AQUAINT [107]	✗	✗	MSNBC
IITB [83]	✓	✗	IITB
ACE2004 [132]	✗	✗	MSNBC
AIDA/CoNLL [71]	✓	✗	AIDA
DBpedia Spotlight [101]	✓	✗	Lexvo
KORE50 [70]	✓	✗	AIDA
N ³ -RSS 500 [135]	✓	✗	NIF
Reuters 128 [135]	✓	✗	NIF
News-100 [135]	✓	✗	NIF
Wes2015 [168]	✓	✗	NIF
SemEval2015 Task 13 [109]	✓	✗	SemEval
Thibaudet [16]	✗	✓	REDEN
Bergson [16]	✗	✓	REDEN
DBpedia Abstracts [20]	✗	✗	NIF
MEANTIME [108]	✓	✓	CAT

details on these datasets can be found in the survey by Martinez-Rodriguez et al. [93] as well as in the discussions by Usbeck et al. [164], van Erp et al. [43] and Jha et al. [78] on EL evaluation.

3.5 EL Formats

As seen previously in Table 3.1, multiple formats have been used to serialize EL benchmark datasets. We will illustrate the most prominent such formats with the following sentence:

S1: “The singer Jackson is a best-selling music artist.”

One of the first formats proposed was the MSNBC dataset [34], which uses an XML-based format; we provide an example of the format in Figure 3.3, describing the mention “Jackson”

Figure 3.3: MSNBC format for EL annotations

```
<ReferenceInstance>
  <SurfaceForm>Jackson</SurfaceForm>
  <Offset>11</Offset>
  <Length>7</Length>
  <ChosenAnnotation>Michael_Jackson</ChosenAnnotation>
</ReferenceInstance>
```

Figure 3.4: IITB format for EL annotations

```
<annotation>
  <docName>doc1</docName>
  <userId>Jackson</userId>
  <wikiName>Michael_Jackson</wikiName>
  <offset>11</offset>
  <length>7</length>
</annotation>
```

in sentence **S1** (though not shown, MSNBC also includes tags to specify the number and names of the annotators). The IITB format is similar to MSNBC – being also based on XML – but rather using different tags; we provide an example in Figure 3.4 for the same sentence as shown before.

The AIDA/CoNLL dataset is an extension of the CoNLL dataset, and likewise the format is an extension of the CoNLL “IOB format”¹⁶ used for NER tasks where words are tagged with I/O/B to indicate inside/outside/begin named entities; AIDA/CoNLL extends the format to also include links in the case of B tags that indicate the beginning of a mention. We can see in Figure 3.5 that all words for sentence **S1** are tagged with 0, except “Jackson”, which is the only annotation in this example.

In 2015, SemEval competitions began including a track dedicated to Entity Linking, further introducing a new format for EL benchmark datasets [109]. In Figure 3.6 we provide an example of this format, which consists of two separate files: the first is an XML file for the input data indicating lemma and POS information for each word; the second is a file in TSV format that indicates identifiers from Wikipedia, WordNet and BabelNet (if they exist) for the given mention key in the XML file.

Another EL format is proposed for creating the MEANTIME [108] dataset, which consists

¹⁶<https://www.clips.uantwerpen.be/con112003/ner/>

Figure 3.5: AIDA/CoNLL format for EL annotations

```
-DOCSTART- doc1
The 0
singer 0
Jackson B Jackson wiki:Michael_Jackson
is 0
a 0
best 0
- 0
selling 0
music 0
artist 0
```

of 120 news articles from WikiNews11 with manual annotations of entities, events, temporal information and semantic roles. MEANTIME was built with the CAT¹⁷ tool, which exports annotations with an XML-based format that goes beyond the association of mentions to their correspondence KB resources, additionally including information associated to events that are described in the text. MEANTIME also includes information about the entity type and entity/event cross-document coreference. In Figure 3.7 we provide an example annotation serialized in the CAT format.

Along with increasing interest in the Semantic Web and Linked Data came new vocabularies for describing NLP resources. GOLD [48]¹⁸ was one of the first vocabularies proposed to specify linguistic descriptions in Semantic Web environments, allowing to analyze language data, such as paradigms, lexicons, and feature structures. Another initiative in this direction is *lemon* [96]¹⁹ – and its extensions *lemon-LexInfo*²⁰ and *ontolex-lemon* [95]²¹ – which allow for describing lexical information as RDF, including morphology, syntax, variation, and other descriptors. A number of NLP-related vocabularies further became used in the context of EL. Among these, Melo et al. [99, 98] proposed Lexvo as an RDF-based format and service that defines unique URIs for terms, languages, scripts, and characters from a text corpus; this format would become used in diverse applications, including the serialization of results from DBpedia Spotlight. Hellmann et al. [67] would later propose the NLP Interchange Format (NIF) as an RDF-based vocabulary for enabling interoperability of NLP tools, e.g., Part-Of-Speech, NER, and EL tools. An example of the NIF format is shown in Figure 3.8 for the running example.

¹⁷<https://dh.fbk.eu/resources/cat-content-annotation-tool>

¹⁸<http://linguistics-ontology.org/gold-2010.owl>

¹⁹<https://lemon-model.net/lemon>

²⁰<https://www.lexinfo.net/ontology/3.0/lexinfo.ttl>

²¹<https://www.w3.org/2016/05/ontolex/>

Figure 3.6: SemEval format for EL annotations

data.xml

```
<?xml version="1.0" encoding="UTF-8" ?>
<corpus lang="en">
  <text id="d001">
    <sentence id="d001.s001">
      <wf id="d001.s001.t001"
        pos="X">The</wf>
      <wf id="d001.s001.t002"
        lemma="singer" pos="N">singer</wf>
      <wf id="d001.s001.t003"
        lemma="jackson" pos="N">Jackson</wf>
      ...
    </sentence>
  </text>
</corpus>
```

data.key

```
d001.s001.t002    d001.s001.t003
bn:00047836n    wiki:Michael_Jackson
```

Recalling Table 3.1, we see how the aforementioned EL datasets use these formats. Different formats support different features; for example, early formats did not provide tags to indicate the entity type; on the other hand, the AIDA/CoNLL format does not support overlapping mentions. Noting that Table 3.1 is ordered by recency – with more recent datasets appearing lower in the table – we see that NIF has gained the attention of the EL community: datasets such as N3-RSS 500, Reuters 128, News-100, and Wes2015 were created with NIF, where others have further been transcribed from their own formats to NIF (e.g., ACE04, DBpedia Spotlight and KORE50). Due to the advantages and popularity of NIF, benchmark tools – such as GERBIL [164]²² and NIFify [139]²³ – are based on the NIF format, and support converting other EL formats to NIF.

3.6 EL Design Issues

The goals of the EL task were preceded by those defined for the related NER task. As discussed in the introduction, for the 6th Message Understanding Conference (MUC-6) [58], the concept of a “named entity” was defined as those phrases in a text that refer to instances

²²<http://aksw.org/Projects/GERBIL.html>

²³https://github.com/henryrosalesmendez/NIFify_v3

Figure 3.7: CAT format for EL annotations

```

<?xml version="1.0" ?>
<Document doc_id="1" doc_name="doc1"
  lang="en" url="http://ex.org">
  ...
  <token number="2"
    sentence="0" t_id="3">Jackson</token>
  ...
  <Markables>
    <ENTITY_MENTION m_id="1">
      <token_anchor t_id="3"/>
    </ENTITY_MENTION>
    <ENTITY_TAG_DESCRIPTOR="Jackson"
      ent_type="PER" m_id="101"/>
  </Markables>
  <Relations>
    <REFERS_TO r_id="1">
      <source m_id="1"/>
      <target m_id="101"/>
    </REFERS_TO>
  </Relations>
</Document>

```

of proper name classes such as Person, Location and Organization, and also to numerical classes such as Temporal Expressions & Quantities. Many NER tools were later developed following these guidelines. However, authors such as Fleischman and Hovy [52] remarked that the MUC-6 categories were too coarse for many applications, proposing a finer-grained categorization for people according to their occupation (Athlete, Politician, etc.). Other works rather developed NER systems that could adapt to arbitrary types of entities, where, for example, the work by Etzioni et al. [44] proposed to use Hearst patterns (e.g., “[pop singers] such as [Michael Jackson]”) to identify entities of discovered types.

Turning to EL, while approaches adopting an NERL strategy were based on established NER tools, and thus inherited MUC-6 conventions, there was growing awareness that such types are limited for the purposes of EL when considering diverse KBs like Wikipedia, DBpedia, Freebase, Wikidata, YAGO, etc.; for example, Wikidata contains around fifty thousand entity types. The types typically missed by NER tools include not only common entities in the KB (e.g., “father”, “interview”), which are *arguably* part of a separate Word Sense Disambiguation (WSD) task [116], but also named entities referring to albums (e.g., “Got to Be There”), movies (e.g., “The Godfather”), laws (e.g., “Hooke’s Law”), diseases (e.g., “Ebola”) and so forth.

Figure 3.8: NIF format for EL annotations (in Turtle syntax)

```
<http://example.org#char=11,18> a nif:String,  
  nif:Context, nif:Phrase, nif:RFC5147String;  
  nif:anchorOf "Jackson"^^xsd:string;  
  nif:beginIndex "11"^^xsd:nonNegativeInteger;  
  nif:endIndex "18"^^xsd:nonNegativeInteger;  
  itsrdf:taIdentRef </wiki/Michael_Jackson>.
```

Hence authors began to propose more general definitions for “entity” in the context of the EL task. Rather than use a class-based definition, for example, Ling et al. [88] define that entities mentions are “*substrings corresponding to world entities*”, which though providing a more general perspective, is problematic in the cyclical use of the term “*entity*”; they acknowledge that “*there is no standard definition of the [EL] problem*”, proposing that EL target both named and common entities while NEL target only common entities. Guo et al. [60] rather define an entity as: “*a nonambiguous, terminal page (e.g., The Town (the film)) in Wikipedia (i.e., a Wikipedia page that is not a category, disambiguation, list, or redirect page)*”; while again providing a more general perspective on the types of entities that EL should link, the definition depends on a particular KB and, indeed, a particular version of that KB; furthermore, this definition includes various types of entities that EL systems typically will not link, such as names (e.g., `wiki:Jackson_(name)`), numbers (e.g., `wiki:4`), years (e.g., `wiki:1984`), units (e.g., `wiki:Kilometre`), symbols (e.g., `wiki:Exclamation_mark`), and so forth; should EL also link mentions of such entities?

Even assuming we settle on a particular definition for “entity”, authors have raised further issues relating to the EL task in terms of what kinds of mentions should be considered. With respect to Figure 1.1, for example, while Michael Jackson is clearly an entity of interest, should we link the mention “[he] and his four siblings” to his KB identifier? Though the pronoun is a mention of an entity of interest, some would rather consider this as part of a separate Coreference/Anaphor Resolution task [157]. Consider, then the case of “Living with [Michael Jackson]”, where the entity mention is contained inside another mention: should this be considered a mention of the singer? Overlapping mentions are discussed by, for example, Guo et al. [60], Ling et al. [88], van Erp et al. [43]²⁴, Jha et al. [78], and more besides, with differing opinions; for example, Ling et al. [88] consider overlapping mentions to be useful to include, while Jha et al. [78] consider overlapping mentions to be an error.

Ling et al. [88] further raise two other issues regarding EL, both of them related to the issue of *reference*. Consider for example the sentence “Portugal drew with Spain in their opening game of the World Cup.” The first issue relates to how specific a link should be offered

²⁴This paper refers to overlapping entities across datasets, which is in fact a different issue referring to dataset homogeneity; however, they also mention inner vs. outer entities and nested entities.

by an EL system or dataset; for example, should “World Cup” be linked to `wiki:World_Cup`, `wiki:2018_FIFA_World_Cup`, or maybe even `wiki:2018_FIFA_World_Cup_Group_B`? The second issue relates to indirect type of reference, where they note that “Portugal” should not be linked to `wiki:Portugal` (the country) but rather to `wiki:Portugal_national_football_team` given that countries cannot play football: rather “Portugal” is a meronymic reference to the national football team.

In summary, numerous authors have highlighted a number of difficult issues that complicate research on the EL task. We believe that differing design choices regarding such issues explain some (though not all) of the differences that we have been seeing since Chapter 1 with respect to the results of four EL systems. We can also see evidence of these differences of opinion in different EL datasets, where the SemEval 2015 Task 13 [109] and DBpedia Spotlight [101] datasets allow overlapping entities, while datasets such as ACE2004 [132] and AIDA/CoNLL [71] do not; in fact, Jha et al. [78] consider the overlapping mentions in DBpedia Spotlight to be errors and remove them. We also note that MEANTIME [108] provides coreference annotations. Comparing the performance of EL systems is then complicated by the varying design decisions adopted by the systems and the datasets considered for evaluation.

One of our goals in this thesis is to highlight, understand and address these design issues regarding EL, where we begin in the section that follows with a questionnaire to first understand what consensus (or lack thereof) exists regarding the goals of the task.

Chapter 4

Consensus about Entity Linking

As seen in the previous chapter, a wide variety of techniques have been brought to bear on the EL task. Perhaps as a result, a number of authors have noted a lack of consensus on the precise goals of the task, particularly in terms of what kinds of mentions in an input text an EL system should link to which identifiers in the KB; this issue affects not only EL systems, but also the definition of benchmark datasets [88, 169, 43, 78, 141]. This lack of consensus on EL’s goals presents complications for the EL research community, particularly when it comes to evaluating and comparing different systems making different assumptions.

Anecdotally, Figure 1.1 presents the entity mentions recognized by a selection of popular online EL systems – Babelfy (strict configuration) [110], DBpedia Spotlight [101], FRED [55] and TagME [50] – for an example input sentence. We see that no entity is recognized by all four systems. While some of the differences can be attributed to varying performance by the system – e.g., DBpedia Spotlight misses the **Martin Bashir** mention, though it is a named entity appearing in the DBpedia KB – we argue that other differences are due to the systems targeting different types of entity. For example, while all systems target named entities based on proper nouns like “**Michael Jackson**”, behavior differs across EL systems for *common entities* based on common noun phrases like “**interview**” [88]; in particular, TagME and DBpedia Spotlight recognize common entities, while Babelfy and FRED exclusively label named entities. Other differences may be explained by varying policies regarding *overlapping entities* – entity mentions with overlapping text – where Babelfy identifies both “**Living with Michael Jackson**” and the inner mention “**Michael Jackson**”, while the other three systems identify one or the other, but not both.

So which system is “correct”? We argue that the types of entities that an EL system should target depends on the application, and hence there is no correct answer to questions such as the types of entities that should be targeted, whether or not overlapping entities should be allowed, and so forth. More specifically, different EL applications may have different requirements. At the same time, however, with these varying perspectives on the EL task, it is not clear how we should define gold standards that offer a fair comparison of tools [88, 169, 43, 78, 141]. A typical approach to address this issue has been to make cer-

tain design choices explicit, such as to enforce a particular policy with respect to overlapping mentions, or common entities, etc., when designing an EL system, labeling an EL dataset, or performing evaluation. In this thesis, we rather consider that one size does not fit all, and pursue a different direction, which is to better understand the goals of the EL task, and to subsequently propose a fine-grained categorization of different types of entity mentions and links, allowing us to compare the performance of different EL systems for different categories of entity mentions and links.

4.1 Questionnaire on the Goals of EL

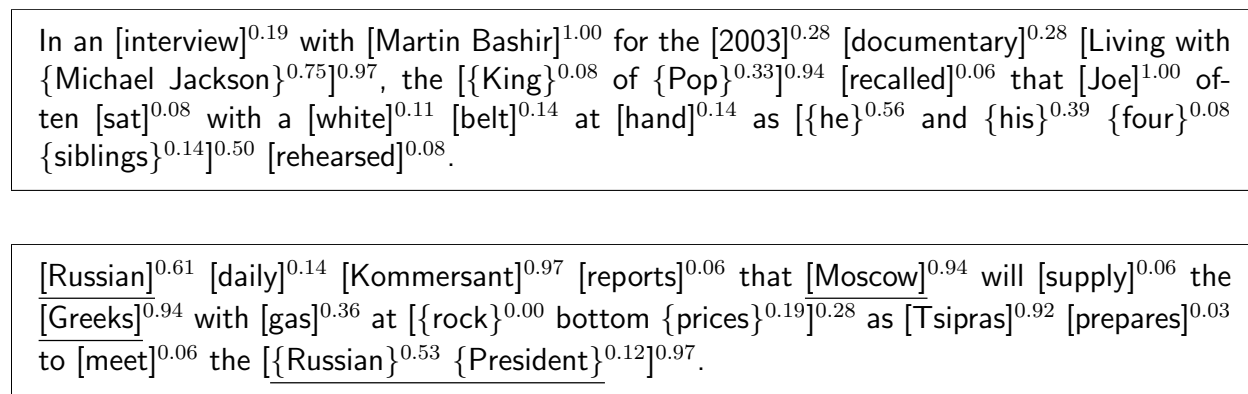


Figure 4.1: The two sentences used for the questionnaire annotated with the ratio of respondents who suggested to annotate the corresponding mentions with some link; in the case of underlined mentions, multiple links were proposed, as presented in Table 4.1.

Based on the previous discussion, we see that there are often diverging perspectives with respect to the EL task. This raises a key question: *what are the goals of the EL task?* We believe that the answer to this question is a matter of convention, and we wish to understand what consensus exists within the EL research community itself. Along these lines, we created a short questionnaire with two sentences that contain concrete examples for the issues discussed. We show the sentences in Figure 4.1 (along with results that will be discussed presently). Subsequently addressing the questionnaire to the EL research community, we aim to gain insights into the varying perspectives regarding the following questions on the goals of EL (referring to **RQ1a**):

1. *KB types*: should types of entities not typically considered under MUC-6 definitions be targeted (e.g., linking the documentary “Living with Michael Jackson” to the KB)?
2. *Overlapping mentions*: should mentions whose text overlaps with other mentions be allowed (e.g., should “Michael Jackson” be annotated inside the “Living with Michael Jackson” mention)?

3. *Common entities*: should common entities be annotated in cases where the KB provides a corresponding identifier for that entity (e.g., “documentary”)?
4. *Parts of speech*: should EL only target mentions that are noun phrases or should mentions using other parts of speech also be linked (e.g., “Russian” or “reports”)?
5. *Indirect mentions*: should pronouns (e.g., “he”) and descriptive noun phrases (e.g., linking “he and his four siblings” to `wiki:The_Jackson_5`) be targeted?
6. *Complex reference*: should EL only link mentions to the entity being explicitly named (e.g., linking “Moscow” to `wiki:Moscow`), or should EL resolve more complex forms of references, such as *metonymy* (e.g., linking “Moscow” to `wiki:Government_of_Russia`), *hypernymy* (e.g., linking “daily” to `wiki:Newspaper` with it being the closest entity in the KB, or linking Russian President to `wiki:Vladimir_Putin`), or *metaphor* (e.g., linking King to `wiki:King`) be considered?

For each of the two sentences in Figure 4.1, the respondent was provided a list of questions. Each question proposed a mention – in sequential order of the text – along with a list of one or more possible KB links, or the option not to annotate the mention at all (with any link). We chose Wikipedia as the target KB where we assume that it is the most likely KB for most respondents to be familiar with. A total of 38 questions were asked, corresponding to 38 potential mentions in the two sentences. Each question was optional. Respondents were asked at the start of the questionnaire to select the mentions and links that they believe an EL system should *ideally* target in each case presented; we also highlighted that there was no “correct” answer and that we rather sought their opinions on the annotations.¹

We wished to use this questionnaire to ascertain the perspectives on the goals of the EL task among members of the EL research community. Along these lines, taking the recent EL survey paper of Wu et al. [172], we manually extracted the emails of all authors of papers referenced by the survey that are directly related to the EL task. We successfully extracted the emails of 321 authors. Sending a link to the questionnaire to all authors, 232 individual mails were delivered without an error message. From these mails, we received a total of 36 responses. Detailed responses are available online², where in Figure 4.1 we summarize the results, indicating in superscript the ratio of respondents who agreed to some link being provided for the given mention.

Regarding initial high-level conclusions, of the 36 respondents, all agree that “Martin Bashir” and “Joe” – corresponding to named entities included in the MUC-6 definitions with non-overlapping, direct mentions – should be linked to their corresponding KB identifiers. Conversely, the respondents also unanimously agreed that “rock” – corresponding to a common entity with a potentially overlapping mention making a metaphorical reference – should

¹The questionnaire design can be reviewed online: <https://users.dcc.uchile.cl/~hrosales/questionnaire>

²<https://users.dcc.uchile.cl/~hrosales/questionnaire>

not be linked to the KB. All of the other mentions – 35/38 of the cases – exhibited some level of (varying) disagreement among the respondents.

1. *KB types*: Per the response for “Living with Michael Jackson” (0.97), which refers to a documentary in the KB, the vast majority of respondents believe that entities other than traditional MUC-6 types should be considered.
2. *Overlapping mentions*: Per the response for “Michael Jackson” (0.75) – combined with the positive response for “Living with Michael Jackson” (0.97) – most respondents believe that mentions contained within other mentions should be considered.
3. *Common entities*: Most respondents do not believe that common entities in the KB should be considered, where the mention of a common entity with the highest positive response was “gas” (0.36). Of note is that more than double the respondents agree with annotating “gas” (0.36) when compared with “belt” (0.14); our results are inconclusive as to why this might be the case.
4. *Parts of speech*: Most respondents believe that mentions other than noun phrases should be considered, where the non-noun mention with the highest positive response was the (first appearance of the) adjective “Russian” (0.67).
5. *Indirection mentions*: There was considerable disagreement on whether or not indirect forms of reference should be considered, with “he” (0.56) and “he and his four siblings” (0.5)³ being considered by roughly half of the respondents; fewer supported the possessive adjective “his” (0.39) being linked to Michael Jackson.
6. *Complex reference*: We offered multiple links on the mentions underlined in Figure 4.1 to determine if respondents prefer to consider direct forms of reference or to resolve more complex forms of reference (or both: the questions were multiple choice). The results are shown in Table 4.1, where of particular interest are the results for “Moscow”, which indicate that most respondents prefer to resolve the metonymic reference to the Government of Russia rather than directly linking to the city of that name; and the results for “Russian President”, which indicate that respondents preferred to link to the person indirectly referred to rather than the office directly named. These results indicate that respondents prefer to resolve complex forms of reference rather than merely linking mentions to entities with corresponding labels. Finally, returning to Figure 4.1, we note that metaphorical references such as “King” (0.08) and “rock” (0.00) received little support.

Overall, we see support by the majority of participants for considering named entities of any KB type in the EL task, including those not considered by MUC-6 definitions and those involved in overlapping mentions. On the other hand, a minority of respondents consider

³One respondent commented that, from the given context, they were not certain that the mention “he and his four siblings” referred to The Jackson 5, which was the KB link suggested for the question.

Table 4.1: The ratio of respondents choosing particular links for mentions with multiple choices (underlined) in Figure 4.1; the questions were multiple choice, so respondents could choose multiple possibilities

Link	Ratio
[Russian] daily Kommersant ...	
wiki:Russia	0.61
wiki:Russians	0.11
wiki:Russian_language	0.08
... that [Moscow] will supply ...	
wiki:Government_of_Russia	0.77
wiki:Moscow	0.36
... supply the [Greeks] with gas ...	
wiki:Greece	0.77
wiki:Greeks	0.36
... the [Russian] President.	
wiki:Russia	0.42
wiki:Russians	0.19
... the [Russian President].	
wiki:Vladimir_Putin	0.77
wiki:President_of_Russia	0.61

common entities as part of the EL task. Most respondents agree that some non-noun phrases can be considered as mentions. Opinions are more divided regarding pro-forms and other forms of descriptive mentions. There was also a clear preference for resolving complex forms of reference, i.e., that EL should ideally link to the entity being talked about rather than the entity explicitly named by the mention.

We reiterate that we do not interpret any “correct” answer here, and that the goal of the questionnaire is to collect data about the perspectives that exist, potentially informing conventions for the EL task. In general, however, we see considerable disagreement, suggesting that it would be premature to propose a rigid definition of the goals of EL from this questionnaire; for example, while only a minority of respondents consider common entities – and thus we might consider excluding such entities from the EL task, concluding perhaps that they are rather part of a separate Word Sense Disambiguation (WSD) task [116] – still, the 36% of respondents including “gas” is not an inconsiderable number. Likewise, while we might exclude pro-forms from consideration by the EL task – considering them part of a separate Coreference/Anaphora Resolution (CR) task [157] – again, mentions such as “he” received majority support.

More generally, we believe that the appropriate definition of the goals of the EL task depend on the particular setting. For example, if EL is to be incorporated as part of a Relation Extraction framework, then having links for pronouns such as “he” is important to find additional relations and improve recall. On the other hand, if EL is to be used for the purposes of Semantic Search, then it may suffice to have a subset of named mentions for an entity to know that the document speaks of that entity. Along these lines, we propose that no one definition of the goals of the EL task fits all such settings. Rather than pursue a universal definition of the task, we thus instead propose to be more explicit about these different types of mentions and links, reflecting the diversity of perspectives seen in this questionnaire, and allowing to understand the performance of EL systems under different assumptions. Along these lines, in the next section we propose a fine-grained categorization scheme for EL annotations that encapsulates these varying perspectives.

4.2 Proposed Solution

We are not the first to identify such issues: Ling et al. [88] provide similar examples of the lack of consensus for EL, while Jha et al. [78] also identify this problem and propose a set of rules to serve as best practices for benchmark creation. While standardizing the creation of EL benchmarks and making explicit the assumptions under which they are generated is a step in the right direction, as previously discussed, it is not clear what assumptions should, in reality, be adopted. Jha et al. [78] propose, for example, that overlapping mentions be omitted (and, in fact, refer to their inclusion as “errors”) but as discussed, other authors (including Ling et al. [88]) disagree on this specific issue. These facts answer **RQ1b**.

Our position is that the more fundamental question needing to be resolved in the context

of EL is not the semantic question of “*what is an ‘entity’?*”, but rather the practical question of “*what should Entity Linking link?*”. The answer to this latter question, we argue, depends heavily on the application. For the purposes of semantic search – for example, finding all documents about US singers – coreference is not so important since one mention of *Michael Jackson* in a document may be enough to establish relevance. On the other hand, for extracting relations between entities, many such relations may be expressed in text with pronouns. Likewise an EL process may choose to recognise and link mentions of terms such as “*singer*” to the KB to help to apply a more accurate (collective) disambiguation of neighbouring mentions such as “*Michael Jackson*” (as proposed by Babely). Any single set of rules or definitions by which EL should be conducted is, we thus argue, exclusionary and an oversimplification.

Hence our proposed solution is not to provide another unilateral definition of what EL should consider as an “*entity*” or an “*entity mention*”, but rather to be explicit on the different forms of entities and entity mentions that a particular EL system may wish to recognize and link. This would involve creating labeled texts – for training and benchmarking – that make explicit the different forms of entity mentions present, be they proper names, other terms present in the KB, overlapping entities, or coreferences. Tools and evaluators may then choose to explicitly include/exclude whichever entity (mentions) they consider relevant for their application. Much like the original MUC-6 definitions, we propose that such labels should be established through consensus in the community and included in standards such as the NLP Interchange Format (NIF) [67]. While this would add some additional complexity to the generation of labeled datasets and the processes of evaluation (when compared with, e.g., the proposals of Jha et al. [78]), we argue that such additional effort is no more than what the EL community will *require* as it matures. We would thus like to propose a metric that takes into account the ambiguity of what is an entity, and that measures the capacity of an EL system to link different types of entities.

In the next chapter, we propose a new multilingual dataset that addresses a key limitation of existing multilingual datasets: that the entities differ across different languages, making comparison across languages imprecise. When labeling the dataset, we are faced with his lack of consensus. Our initial solution to take a “strict” and “relaxed” notion of an entity. Later we rather propose a finer-grained categorizations scheme and associated metrics.

Chapter 5

Multilingualism

The majority of approaches have focused on EL over English text collections, leaving out the need to link entities over non-English text. However, building language-agnostic approaches arose to respond to this need. Initial multilingual approaches – such as TagME that initially supported English and Italian – were based on Wikipedia. At the time of writing this thesis, Wikipedia provides 303 active versions¹ where each of them corresponds to one language. In this context, Multilingual EL can generate two types of links:

direct-lingual links where the given text corpora and the targeted encyclopedia are in the same language.

cross-lingual links where the given text corpora and the target encyclopedia are written in different languages. For instance, given a Spanish text mentioning MICHAEL JACKSON, return a link to the English version of Wikipedia, i.e., https://en.wikipedia.org/wiki/Michael_Jackson.

Although the Text Analysis Conference² (TAC) in 2009 was one of the bootstrappers of EL, it was in 2011 when it included multilingualism. TAC 2011³ introduced a multilingual dataset based on English and Chinese languages, and their successive editions until 2016. In 2017, it included Spanish as well.

Since TAC 2016, multilingualism in EL has been addressed more and more, taking Wikipedia as a target resource of disambiguation [24, 45, 147, 176]. However, some authors have stressed the intrinsic characteristics of Wikipedia that do not favor this subtask. Among these perhaps the most important is the inequality in the amount of document across its different language versions; while the English version is the largest one with more than 6 million articles⁴, languages such as Spanish, Italian, Polish, and others, have fewer 2 million. On the other hand, the content may differ for each article in different languages. For

¹https://en.wikipedia.org/wiki/List_of_Wikipedias

²<https://tac.nist.gov/2009/>

³<https://tac.nist.gov/2011/>

⁴https://en.wikipedia.org/wiki/List_of_Wikipedias

instance, the article `LIVING WITH MICHAEL JACKSON` contains⁵ 1454 words on the English version, but only 184 in the Spanish one. RDF KBs can help to partially address these non-desirable properties. RDF KBs – such as Wikidata, DBpedia, and Yago – are by definition multilingual resources, where the multilingual information is stored on the same nodes rather than on separate nodes, per Wikipedia. Still, labels for nodes and edges are specific to a language, where certain languages may have more labels available than others.

5.1 Multilingual EL Systems

In order to help address this gap in terms of the availability of resources for different languages, in this chapter we propose a new EL dataset called VoxEL, which allows for comparison in performance across different languages, while also offering strict and relaxed annotations to address the lack of consensus on the goals of the EL task.

In this section, we survey those EL approaches that handle more than one language. We thus focus on EL systems that have published evaluation results over texts from multiple languages⁶, thus demonstrating proven multilingual capabilities. We summarise such systems in Table 5.1, where we provide details on the year of the main publication, the languages evaluated, as well as denoting whether or not entity recognition is supported⁷, and whether or not a demo, source code or API is currently available. As expected, a high-level inspection of the table shows that English is the most popularly-evaluated (and thus we surmise supported) language, followed by European languages such as German, Spanish, French, Dutch and Italian. We also highlight that most of the multilingual EL approaches included in the table have emerged since 2010.

We will later conduct experiments using the GERBIL evaluation framework [164], which allows for invoking and integrating the results of a variety of public APIs for EL, generating results according to standard metrics in a consistent manner. Hence, in our later experiments, we shall only consider those systems with a working REST-API made available by the authors of the system. In addition, we will manually label our VOXEL system according to Wikipedia, with which other important KBs such as DBpedia, YAGO, Freebase, Wikidata, etc., can be linked; hence we only include systems that support such a KB linked with Wikipedia. Note that GERBIL automatically takes care of mapping coreferent identifiers across KBs (and even across languages in cases such as DBpedia with different KB identifiers for different languages and cross-language links).

⁵By the time of writing this thesis, September 1st, 2020. We use <https://wordcounter.net/> for counting.

⁶This excludes systems such as Apache Stanbol, OpenCalais, PoolParty, etc.

⁷Some systems assume that mentions have previously been extracted from the text and are given as input, thereafter focusing only on the disambiguation process.

Table 5.1: Overview of multilingual EL approaches; the italicised approaches will be incorporated as part of our experiments.

Name	Year	Evaluated Languages	ER	Demo	Src	API
KIM [127]	2004	EN,FR,ES	✓	✓	✗	✓
<i>TagME</i> [50]	2010	DE,EN,NL	✓	✓	✗	✓
SDA [27]	2011	EN,FR	✓	✗	✗	✗
ualberta [61]	2012	EN,ZH	✓	✗	✗	✗
HITS [45]	2012	EN,ES,ZH	✓	✗	✗	✗
<i>THD</i> [39]	2012	DE,EN,NL	✓	✓	✓	✓
<i>DBpedia Spotlight</i> [101, 36]	2013	DA,DE,EN,ES,FR,HU,IT,NL,RU	✓	✓	✓	✓
Wang-Tang [171]	2013	EN,ZH	✓	✗	✗	✗
AGDISTIS [163]	2014	DE,EN	✗	✓	✓	✓
<i>Babelfy</i> [110]	2014	DE,EN,ES,FR,IT	✓	✓	✗	✓
<i>FREME</i> [143]	2016	DE,EN	✓	✗	✓	✓
WikiME [161]	2016	AR,DE,EN,ES,FR,HE,IT,TA,TH,TL,TR,UR,ZH	✓	✓	✗	✗
FEL [121]	2017	EN,ES,ZH	✓	✗	✓	✗
FOX [151]	2017	DE,EN,ES,FR,NL	✓	✓	✓	✓
MAG [111]	2017	DE,EN,ES,FR,IT,JA,NL	✗	✓	✓	✓

5.1.1 Multilingual EL Datasets

In order to train and evaluate EL approaches, labelled datasets – annotated with the correct entity mentions and their respective KB links – are essential. In some cases these datasets are labelled manually, while in other cases labels can be derived from existing information, such as anchor texts. In Table 5.2 we survey the labelled datasets most frequently used by EL approaches (note that sentence counts were not available for some datasets).

We can see that the majority of datasets provide text in one language only – predominantly English – with the exceptions being as follows:

SemEval 2015 Task 13: is built over the biomedical, math, computer and social domains and is designed to support EL and WSD at the same time, containing annotations to Wikipedia, BabelNet and WordNet [109].

DBpedia Abstracts: provides a large-scale training and evaluation corpora based on the anchor texts extracted from the abstracts (first paragraph) of Wikipedia pages in seven languages [20].⁸

⁸<http://wiki-link.nlp2rdf.org/abstracts/>; April 1st, 2018

Table 5.2: Survey of datasets for EL task. For multilingual datasets, the quantities shown refer to the English data available. We present metadata about the relaxed and strict version of our dataset by VoxEL_R and VoxEL_S respectively. (Abbreviations: $|D|$ number of documents, $|S|$ number of sentences, $|E|$ number of entities, **Mn** denotes that all entities were manually annotated.)

Dataset	$ D $	$ S $	$ E $	Mn	Languages
AIDA/CoNLL-Complete [71]	1393	22,137	34,929	✓	EN
HoffartSNTW12 [70]	50	50	144	✓	EN
IITB [83]	103	1,781	18,308	✓	EN
ACE2004 [132]	57	-	306	✗	EN
RatinovRDA11 [132]	50	533	727	✗	EN
MSNBC [34]	20	668	747	✗	EN
DBpedia Spotlight [101]	10	58	331	✓	EN
N3-RSS 500 [136]	1	500	1000	✓	EN
Reuters 128 [136]	128	-	881	✓	EN
Wes2015 [168]	331	-	28,586	✓	EN
News-100 [136]	100	-	1656	✓	DE
Thibaudet [16]	1	3,807	2,980	✗	FR
Bergson [16]	1	4,280	380	✗	FR
SemEval 2015 Task 13 [109]	4	137	769	✓	EN,ES,IT
DBpedia Abstracts [20]	39,132	-	505,033	✗	DE,EN,ES,FR,IT,JA,NL
MEANTIME [20]	120	597	2,790	✓	EN,ES,IT,NL
VoxEL_R	15	94	674	✓	DE,EN,ES,FR,IT
VoxEL_S	15	94	204	✓	DE,EN,ES,FR,IT

MEANTIME: consists of 120 news articles from WikiNews⁹ with manual annotations of entities, events, temporal information and semantic roles [108].¹⁰

With respect to DBpedia Abstracts, while offering a very large multilingual corpus, the texts across different languages vary, as do the documents available; while such a dataset could be used to compare different systems for the same languages, it could not be used to compare the same systems for different languages. Furthermore, there are no guarantees for the completeness of the annotations since they are anchor texts/links extracted

⁹<https://en.wikinews.org/>; April 1st, 2018

¹⁰<http://www.newsreader-project.eu/results/data/wikinews/>; April 1st, 2018

from Wikipedia; hence the dataset is best suited as a large collection of positive (training) examples, in a similar manner to how TagME [50] and FRED [143] use anchor texts.

Unlike DBpedia Abstracts, the SemEval and MEANTIME datasets contain analogous documents translated to different languages (also known as *parallel corpora* [109]). Our VOXEL dataset complements these previous resources but with some added benefits. Primarily, both the SemEval and MEANTIME datasets exhibit slight variations in the annotations across languages, leading to (e.g.) a different number of entity annotations in the text for different languages; for example SemEval [109] reports 1,261 annotations for English, 1,239 for Spanish, and 1,225 for Italian, while MEANTIME [20] reports 2,790 entity mentions for English, 2,729 for Dutch, 2,709 for Italian and 2,704 for Spanish. On the other hand, VOXEL has precisely the same annotations across languages aligned at the sentence level, and also features datasets labelled under two definitions of entity. More generally, we see VOXEL as complementing these other datasets.

5.1.2 Non-English Entity Linking: Spanish use-case

In theory, any EL system can be applied to any language and can be expected to produce some partial results; even a system supporting only English may still be able to correctly recognise and link the name of a person such as *Michael Jackson* in the text of another language, assuming the alphabet remains the same. Hence, the notion of a multilingual EL system can become blurred. For example language-agnostic systems – systems that require no linguistic components or resources specific to a language – can become multilingual simply by virtue of having a reference KB with labels in a different – or multiple different – language(s).

A high-level inspection of the Table 5.1 will reveal that English is by far the most popular language. Beyond that, most languages tackled are European languages, with Spanish, French, German and Dutch appearing frequently. Outside of these European languages, Chinese is the most commonly encountered, with other languages appearing only in one or two tools. As an informal but intuitive observation, it would appear that the languages evaluated for a tool often relate to the language(s) spoken by the authors.

Some of the approaches mentioned in this table do not actually address the multilingual problem directly. Rather they are developed as language-agnostic EL systems that rely on generic processing methods that can perform EL over a broad range of languages assuming a suitable knowledge-base with lexical forms (i.e., entity labels and aliases) in that language. Such systems include KIM [127], SDA [27], THD [39], TAGME [50, 115] and AGDISTIS [163].

An example of a (largely) language-agnostic approach is DBpedia Spotlight. The first version of DBpedia Spotlight [101] only supports English language. However, a recent extension of DBpedia Spotlight was introduced in [36] which addresses multilingual EL using the variety of language versions now available for DBpedia. In this multilingual version,

DBpedia Spotlight identifies the entity mentions using Apache OpenNLP¹¹ and from the sequences of capitalized words. To perform ranking, the authors consider various (standard) features, including, for example, the probability that a mention could be a text anchor in Wikipedia.

While the previous systems assume a knowledge base in the same language as the text to analyze, a variety of tools rather support *cross-lingual EL*, where the goal is to link text in a language different to that from the given knowledge-base. Often the goal is to match text in a language other than English to a knowledge base with labels in English. This helps to address the aforementioned asymmetry in the structured information available in English versus other languages. Such cross-lingual approaches include ulberta [61], HITS [45], Babelify [110], and those proposed by Wang and Tang [171], and Tsai and Roth [161].

Next, we thus present some experiments to gain insights into research question **RQ2a**. In particular, we perform experiments comparing EL over the same text expressed in English and Spanish for a variety of systems.

A number of benchmarking frameworks have been proposed for Entity Linking systems, the most recent and comprehensive of which is GERBIL [164]; however, the system does not explicitly offer multi-lingual datasets. Other comparative evaluations have looked at multiple languages. For example, Narducci et al. [115] perform comparison of a variety of approaches – TAGME, WikiMiner and DBpedia Spotlight – for German and Dutch text collections. Still, many of these evaluations use different texts in different languages with the goal of comparing across systems; our emphasis is rather on understanding how systems perform across languages. Hence, to facilitate such comparison, we wish to perform evaluation over the same text in multiple languages.

For this reason, our initial experiments are based on the SemEval 2015 Task 13 [109] dataset, which is divided into four documents with the same content in English, Italian and Spanish. In total, there are 137 sentences. For the moment, we focus on the English and Spanish languages. The goal is then to perform linking to BabelNet. In fact, a number of tools responded to the call for Task 13, and have reported results in these languages. To validate our evaluation process, we first reevaluate the annotations performed by the participants shown in Table 5.3 for which we could locate source code. In all the cases, we obtain the same results as reported in the contests except in the case of SUDOKU-Run1 for English, which was scored with 0.534 in SemEval 2015 Task 13, versus our result of 0.494. We also include in Table 5.3 some new results for Babelify, which is the only other approach that links to BabelNet entries; in terms of F_1 , the system falls behind the SUDOKU configurations but tends to fare better than other systems. We also note that with the exception of SUDOKU-Run1 and EBL-Hope, systems perform better for English than Spanish (and sometimes markedly so).

¹¹<http://opennlp.apache.org/>

Table 5.3: Replicating results of available systems for the Spanish and English texts of the SemEval 2015 Task 13 (also adding novel Babelify results)

System	Spanish			English		
	P	R	F1	P	R	F1
SUDOKU-Run2	0.607	0.525	0.563	0.640	0.609	0.625
SUDOKU-Run3	0.592	0.512	0.549	0.644	0.612	0.627
SUDOKU-Run1	0.601	0.490	0.540	0.501	0.488	0.494
LIMSI	0.535	0.440	0.483	0.694	0.608	0.648
EBL-Hope	0.525	0.446	0.482	0.490	0.429	0.457
Babelify	0.586	0.427	0.493	0.642	0.574	0.606

As a first experiment approach, we extend the systems for which results are available on the selected SemEval 2015 Task 13 dataset. Given the wealth of EL systems proposed, in order to facilitate testing, we select systems based on four criteria: (1) details of the system must be published; (2) a public demo or API must be available for the system; (3) the system must be a complete EL system including both ER and ED phases; (4) the system must perform linking to Wikipedia or a related resource, such as DBpedia, YAGO or BabelNet. Hence, from the multilingual EL approaches selected in Table 5.1, these criteria mean we will test with THD, DBpedia Spotlight, TAGME, Babelify and WikiME. KIM is excluded since it does not link to a Wikipedia resource; AGDISTIS is excluded since their APIs do not perform ER; other systems are excluded for not having a demo/API.

One may note that Spanish is not listed for THD and TAGME. We are still interested to see to what extent having explicit multilingual support is really important for EL systems, and to compare systems that allow for explicitly selecting a given language such as Spanish versus those that do not allow for selecting a language and thus presume (e.g.) English text. We may consider, e.g., that *Michael Jackson* or *Chile* would be recognized/disambiguated by both systems, while *Irlanda* might not be recognized/disambiguated by systems not configured for Spanish [113]. For the purposes of comparison, we thus test not only the THD and TAGME systems – multilingual systems without explicit support for Spanish – but also the AIDA system [71] – a monolingual system that does not allow for selecting a language, but that meets the other criteria.

Hence the final list of systems selected for evaluation are: *configurable for Spanish*: Babelify, DBpedia Spotlight and WikiMe; *multilingual but not configurable for Spanish*: TAGME and THD; *monolingual*: AIDA. All systems are run with default configurations, except DBpedia Spotlight, which does not directly suggest defaults; we configured the system with *support* equal to 0 and *confidence* equal to 0.25 based on some initial experiments.

Table 5.4: ER-level evaluation of selected approaches for the SemEval 2015 Task 13 in Spanish and English. Approaches configured for Spanish are italicized.

System	Spanish			English			%
	P	R	F1	P	R	F1	
<i>Sentence level</i>							
<i>Babelfy</i>	0.727	0.540	0.620	0.820	0.644	0.721	85.99
<i>DBpedia-Spotlight</i>	0.298	0.607	0.400	0.556	0.554	0.555	72.07
<i>WikiMe</i>	0.737	0.018	0.036	0.656	0.028	0.053	67.92
TAGME	0.240	0.319	0.274	0.583	0.687	0.631	43.42
THD	0.281	0.061	0.100	0.587	0.080	0.142	70.42
AIDA	0.750	0.008	0.015	0.688	0.029	0.057	26.32
<i>Document level</i>							
<i>Babelfy</i>	0.765	0.581	0.661	0.864	0.704	0.776	85.18
<i>DBpedia-Spotlight</i>	0.300	0.612	0.403	0.555	0.549	0.552	73.01
<i>WikiMe</i>	0.783	0.023	0.045	0.621	0.024	0.046	97.83
TAGME	0.256	0.255	0.255	0.557	0.551	0.554	46.02
THD	0.277	0.060	0.098	0.587	0.080	0.142	69.01
AIDA	0.857	0.008	0.016	0.667	0.026	0.051	31.37

Next, we conduct a more deeper experimentation to better respond to the research question **RQ2a**. We evaluate approaches separately for ER and EL phases and for sentence-level and document-level texts. The evaluation results for ER and EL phases are presented in Table 5.4 and Table 5.5 respectively. Note that for quick reference, the % column presents the ratio of the F_1 measure for Spanish vs. English, directly comparing the performance for both languages.

From both tables, we can see that results for both EL and ER can vary significantly for Spanish and English, even for systems configurable for both languages. However, in general, those systems configurable for Spanish experienced much less of a gap across both languages when compared with the analogous results for systems not configured for that language.

The gap between Spanish and English performance is hardly surprising for tools not configured for Spanish: TAGME is based on the analysis of anchor text of the English Wikipedia pages; THD selects candidates using the Search API of English Wikipedia; AIDA is based on an English part-of-speech tagger. Clearly these approaches will not perform well for Spanish. The language gap for THD is not so pronounced; however, the F_1 scores

Table 5.5: Overall EL evaluation of selected approaches for the SemEval 2015 Task 13 in Spanish and English. Approaches configured for Spanish are italicized.

System	Spanish			English			%
	P	R	F1	P	R	F1	
<i>Sentence level</i>							
<i>Babelfy</i>	0.599	0.324	0.420	0.725	0.467	0.568	73.94
<i>DBpedia-Spotlight</i>	0.482	0.293	0.364	0.581	0.322	0.414	87.92
<i>WikiMe</i>	0.929	0.017	0.033	0.952	0.026	0.051	64.71
TAGME	0.371	0.118	0.179	0.568	0.391	0.463	38.66
THD	0.596	0.036	0.069	0.738	0.059	0.120	57.50
AIDA	0.667	0.005	0.010	0.773	0.022	0.044	22.73
<i>Document level</i>							
<i>Babelfy</i>	0.597	0.347	0.439	0.729	0.513	0.602	72.92
<i>DBpedia-Spotlight</i>	0.444	0.272	0.337	0.584	0.321	0.414	81.40
<i>WikiMe</i>	0.944	0.022	0.043	0.944	0.022	0.043	100.00
TAGME	0.327	0.083	0.133	0.555	0.306	0.395	33.67
THD	0.609	0.036	0.069	0.738	0.059	0.110	62.73
AIDA	0.667	0.005	0.010	0.900	0.023	0.046	21.74

in general are quite low, making it hard to draw conclusions: the performance for both languages is quite poor. In summary, such systems are likely to only be able to recognize/link entities that are also “valid” in English, such as *Michael Jackson* or *Chile*.

What is perhaps more interesting then, is the consistent gap between both languages for the three systems specifically configured for those languages. In particular, we propose that this result may be due to one (or more) of the following issues faced by multilingual systems:

- *The knowledge base contains different information for both languages.* In Wikipedia anyone can create or edit articles, but this is done separately for each language; thus, equivalent pages in both languages store different content; e.g., even though the label *Michael Jackson* does not change across languages, the content and links in the Spanish and English edition of Wikipedia involving that entity will change. Thus, the use of different editions of Wikipedia to handle multilingual EL can introduce a gap in the performance for both languages. This issue may in particular affect DBpedia Spotlight, which performs the ranking stage of ER based on the occurrence of the text anchors for each language-specific Wikipedia page. On the other hand, the EL model of WikiMe

uses a transliteration model to avoid this issue. Likewise, Babelfy should not be as affected by this issue since BabelNet includes a Machine Translation process in its construction.

- *The models/techniques change according to the target language.* Although using language-specific components will improve results for that specific language, it can also introduce another possible gap when considering performance across languages. For example, DBpedia-Spotlight’s ER could be affected by this issue since the model to perform ER is selected according to the targeted language, where some models may be better than others; for example, for English and German, they use OpenNLP models, whereas for Dutch, they used a corpus of manually corrected entities. As another example, Babelfy bases the detection of candidate mentions during the ER phase on part-of-speech tagging, which requires language-specific knowledge, but such components may vary in performance across languages.
- *Variations in the languages themselves.* We must also consider that some languages are inherently more difficult for an EL process than others. As a simple example, many tools rely on capitalization as a feature for detecting entities, where Spanish tends to use less capitalization than English, including for months, languages, religions, personal titles, and titles of works. Likewise some tools consider a fixed-length window of words/tokens as potential candidate mentions, as well as simple noun phrases, whereas Spanish works tend to have longer titles with non-noun tokens, especially when translated from English (e.g., combining both issues, *Star Wars* translates as *La guerra de las galaxias*, which would be *much* more challenging for ER to recognize).

Due to such issues, even the approaches configured for Spanish do not perform as well as for English. Only in the EL/document-level experiment does WikiMe perform equivalently for Spanish and English, though it should be noted that again, the F_1 measure is quite low in both cases (due to low recall).

Summarizing other aspects of the experiments, in general, there are no substantial differences between the performance of the approaches for the document-level and sentence-level experiments (even though systems such as TAGME are specifically designed for short text collections). The gap between languages is not specifically a factor of precision or recall: the gap is implicit in both aspects of performance. The best system for both the ER and EL stages and for both the English and Spanish languages is consistently Babelfy.

5.2 The VoxEL Dataset

In this section, we describe the VOXEL Dataset that we propose as a gold standard for EL involving five languages: German, English, Spanish, French and Italian. VOXEL is based on 15 news articles sourced from the VoxEurop¹² web-site: a European newsletter with the

¹²<http://www.voxeurop.eu/>; April 1st, 2018

same news articles professionally translated to different languages. This source of text thus obviates the need for translation of texts to different languages, and facilitates the consistent identification and annotation of mentions (and their Wikipedia links) across languages. With VOXEL, we thus provide a high-quality resource with which to evaluate the behavior of EL systems across a variety of European languages.

While the VoxEurop newsletter is a valuable source of professionally translated text in several European languages, there are sometimes natural variations across languages that – although they preserve meaning – may change how the entities are mentioned. A common example is the use of pronouns rather than repeating a person’s name to make the text more readable in a given language. Such variations would then lead to different entity annotations across languages, hindering comparability. Hence, in order to achieve the same number of sentences and annotations for each new (document), we applied small manual edits to homogenize the text (e.g., replacing a pronoun by a person’s name). On the other hand, sentences that introduce new entities in one particular language, or that deviate too significantly across all languages, are eliminated; fewer than 10% of the sentences from the original source were eliminated.

When labelling entities, we take into consideration the lack of consensus about what is an “*entity*” [78, 88, 141]: some works conservatively consider only mentions of entities referring to fixed types such as person, organization and location as entities (similar to the traditional NER/TAC consensus on an entity), while other authors note that a much more diverse set of entities are available in Wikipedia and related KBs for linking, and thus consider any noun-phrase mentioning an entity in Wikipedia to be a valid target for linking [122]. Furthermore, there is a lack of consensus on how overlapping entities – like *New York City Fire Department* – should be treated [78, 88]; should *New York City* be annotated as a separate entity or should we only cover maximal entities? Rather than take a stance on such questions – which appear application dependant – we instead create two versions of the data: a *strict* version that considers only maximal entity mentions referring to persons, organizations and locations; and a *relaxed* version that considers any noun phrase mentioning a Wikipedia entity as a mention, including overlapping mentions where applicable. For example, in the sentence “*The European Central Bank released new inflation figures today*” the strict version would only include “*European Central Bank*”, while the relaxed version would also include “*Central Bank*” and “*inflation*”.

To create the annotation of mentions with corresponding KB identifiers, we implemented a Web tool¹³ (described in Section 6.4) that allows a user to annotate a text, producing output in the NLP Interchange Format (NIF) [67], as well as offering visualisations of the annotations that facilitate, e.g., revision. For each language, we provide annotated links targeting the English Wikipedia entry, as well as that language’s version of Wikipedia (if different from English). In case there was no appropriate Wikipedia entry for a mention of a person, organization or place, we annotate the mention with a `NotInLexicon` marker. These

¹³<https://github.com/henryrosalesmendez/NIFify>; April 1st, 2018

annotations were created by the first author in English, which were then revised by the other authors according to the two labelling guidelines (*strict* and *relaxed*). The author then extended these annotations to the other languages using the sentence-level correspondence, thereafter verifying that each language has the same number of annotations and the same set of English Wikipedia identifiers for each sentence.

In summary, VOXEL consists of 15 news articles (documents) from the multilingual newsletter VoxEurop, totalling 94 sentences; the central topic of these documents is politics, particularly at a European level. This text is annotated five times for each language, and two times for the strict and relaxed versions, giving a total of 150 annotated documents and 940 sentences. The same number of annotations is given for each language (including by sentence). For the strict version, each language has 204 annotated mentions, while for the relaxed version, each language has 674 annotated mentions. In the relaxed version, 6.2%, 10.8%, 20.3% and 62.7% of the entries correspond to *persons*, *organizations*, *places* and *others* respectively, while in the strict version the entities that fall in the first three classes constitute 16.9%, 28.7% and 54.4% (*others* are excluded by definition under the strict guidelines). Again, this homogeneity of text and annotations across languages was non-trivial to achieve, but facilitates comparison of evaluation results not only across systems, but across languages.

5.3 Multilingual EL performance

We now use our proposed VOXEL dataset to conduct experiments in order to explore the behavior of state-of-the-art EL systems for multilingual settings.

In order to address the research questions **RQ2b** and **RQ2c**, we ran the multilingual EL systems Babelify, DBpedia Spotlight, FRED, TagME and THD over both versions of VOXEL in all five languages. These experiments were conducted with the GERBIL [164] EL evaluation framework, which provides unified access to the public APIs of multiple EL tools, abstracting different input and output formats using the NIF vocabulary, translating identifiers across KBs, and allowing to apply standard metrics to measure the performance of results with respect to a labelled dataset. GERBIL calls these systems via their REST APIs maintaining default (non-language) parameters, except for the case of Babelify, for which we analyse two configurations: one that applies a more liberal interpretation of entities to include conceptual entities (Babelify_R), and another configuration that applies a stricter definition of entities (Babelify_S), where the two configurations correspond loosely with the relaxed/strict versions of our dataset.

The results of these experiments are shown in Table 5.6, where we present micro-measures for Precision (mP), Recall (mR) and F_1 (mF), with all systems, for all languages, in both versions of the dataset.¹⁴ From first impressions, we can observe that two systems – TagME

¹⁴The GERBIL results are available at https://users.dcc.uchile.cl/~hrosales/ISWC2018_experiment_GERBIL.html

Table 5.6: GERBIL Evaluation of EL systems with Micro Recall (mR), Precision (mP) and F_1 (mF). A value “-” indicates that the system does not support the corresponding language. The results in bold are the best for that metric, system and dataset variant comparing across the five languages (i.e., the best in each row, split by Relax/Strict).

		Relaxed					Strict				
		DE	EN	ES	FR	IT	DE	EN	ES	FR	IT
Babelfy _R	mP	0.840	0.649	0.835	0.824	0.810	0.932	0.785	0.929	0.889	0.907
	mR	0.461	0.522	0.549	0.488	0.451	0.676	0.735	0.710	0.632	0.578
	mF	0.595	0.578	0.662	0.613	0.579	0.784	0.759	0.805	0.739	0.706
Babelfy _S	mP	0.903	0.722	0.916	0.912	0.884	0.942	0.816	0.923	0.912	0.894
	mR	0.181	0.219	0.210	0.200	0.192	0.558	0.524	0.593	0.563	0.583
	mF	0.301	0.336	0.342	0.328	0.316	0.701	0.638	0.722	0.697	0.706
DBspot	mP	0.731	0.745	0.691	0.658	0.682	0.781	0.854	0.690	0.691	0.800
	mR	0.508	0.577	0.399	0.360	0.488	0.544	0.602	0.382	0.406	0.549
	mF	0.600	0.650	0.506	0.466	0.569	0.641	0.706	0.492	0.512	0.651
FREME	mP	0.762	0.803	0.655	0.737	0.857	0.750	0.871	0.660	0.739	0.858
	mR	0.161	0.267	0.175	0.129	0.213	0.426	0.764	0.553	0.416	0.652
	mF	0.266	0.400	0.276	0.219	0.342	0.543	0.814	0.602	0.532	0.740
TagME	mP	0.635	0.754	-	-	0.494	0.875	0.946	-	-	0.742
	mR	0.232	0.488	-	-	0.182	0.652	0.784	-	-	0.509
	mF	0.340	0.592	-	-	0.266	0.747	0.857	-	-	0.604
THD	mP	0.831	0.806	-	-	-	0.857	0.809	-	-	-
	mR	0.109	0.253	-	-	-	0.352	0.647	-	-	-
	mF	0.194	0.386	-	-	-	0.500	0.719	-	-	-

and THD – cannot be configured for all languages, where we leave the corresponding results blank.

With respect to **RQ2b**, for the Relaxed version, the highest F_1 scores are obtained by Babelfy_R (0.662: ES) and DBpedia Spotlight (0.650: EN). On the other hand, the highest F_1 scores for the Strict version are TagME (0.857: EN) and Babelfy_R (0.805: ES). In general, the F_1 scores for the Strict version were higher than those for the Relaxed version: investigating further, the GERBIL framework only considers annotations to be false positives when a different annotation is given in the labelled dataset at an overlapping position; hence fewer labels in the Strict dataset will imply fewer false positives overall, which seems to outweigh the effect of the extra true positives that the Relaxed version would generate. Comparing the best Strict/Relaxed results for each system, we can see that Babelfy_R, DBpedia Spotlight and FREME have less of a gap between both, meaning that they tend to annotate a broader range of entities; on the other hand, Babelfy_S and THD are more restrictive in the entities they link.

With respect to **RQ2c**, considering all systems, we can see a general trend that English

had the best results overall, with the best mF for DBpedia Spotlight, FReME and TagME. For THD, German had higher precision but much lower recall; a similar result can be seen for FReME in Italian in the Relaxed version. On the other hand, Babelify generally had best results in German and Spanish, where, in fact, it often had the *lowest* precision in English.

With respect to possible factors that explain such differences across languages, there are variations between languages that may make the EL task easier or harder depending on the features used; for example, systems that rely on capitalisation may perform differently for Spanish, which uses less capitalisation, (e.g., “*Jungla de cristal*”: a Spanish movie title in sentence case); and German, where all nouns are capitalised. Furthermore, the quality of EL resources available for different languages – in terms of linguistic components, training sets, contextual corpora, KB meta-data, etc. – may also vary across languages.

5.3.1 Why not translate to English?

Regarding **RQ2d**, we present another experiment to address the question of the efficacy of using machine translations. First we note that, although works in related areas – such as cross-lingual ontology matching [53] – have used machine translation to adapt to multilingual settings, to the best of our knowledge, no system listed in Table 5.1 uses machine translations over the input text (though systems such as Babelify do use machine translations to enrich the lexical knowledge available in the KB). Hence we check to see if translating a text to English using a state-of-the-art approach – Google Translate¹⁵ – and applying EL over the translated English text would fare better than applying EL directly over the target language; we choose one target language to avoid generating results for a quadratic pairing of languages, and we choose English since it was the only language working for/supported by all systems in Table 5.6.

A complication for these translation experiments is that while VOXEL contains annotations for the texts in their original five languages, including English, it does not contain annotations for the texts translated to English. While we considered manually annotating such documents produced by Google Translate, we opted against it partly due to the amount of labour it would again involve, but more importantly because it would be specific to one translation service at one point in time: as these translation services improve, these labelled documents would quickly become obsolete. Instead, we apply evaluation on a per-sentence basis, where for each sentence of a text in a non-English language, we translate it and then compare the set of annotations produced against the set of manually-annotated labels from the original English documents; in other words, we check the annotations produced by sentence, rather than by their exact position. This is only possible because in the original VOXEL dataset, we defined a one-to-one correspondence between sentences across the five different languages.

Note that since GERBIL requires labels to have a corresponding position, we thus needed

¹⁵<https://translate.google.com/>; April 1st, 2018

to run these experiments locally outside of the GERBIL framework. Hence, for a sentence s , let A denote the IRIs associated with manual labels for s in the original English text, and let B denote the IRIs annotated by the system for the corresponding sentence of the translated text; we denote true positives by $A \cap B$, false positives by $B - A$, and false negatives by $A - B$.¹⁶

In Table 5.7, we show the results of this second experiment, focusing this time on the Micro- F_1 (mF) score obtained for each system over the five languages of VOXEL, again for the relaxed and strict versions. For each system, we consider three experiments: (1) the system is configured for the given language and run over text for the given language, (2) the system is configured for English and run over the text translated from the given language, (3) the system is configured for English and run over the text in the given language without translation. We use the third experiment to establish how the translation to English – rather than the system configuration to English – affects the results. First we note that without using positional information to check false positives (as per GERBIL), the results change from those presented in Table 5.6; more generally, the gap between the Relaxed and Strict version is reduced.

With respect to **RQ2d**, in Table 5.7, for each system, language and dataset variant, we underline which of the three configurations performs best. For example, in DBpedia Spotlight, all values on the (EN,EN_t) line – which denotes applying DBpedia Spotlight configured for English over text translated to English – are underlined, meaning that for all languages, prior translation to English outperformed submitting the text in its original language to DBpedia Spotlight configured for that language.¹⁷ In fact, for almost all systems, translating the input text to English generally outperforms using the available language configurations of the respective EL systems, with the exception of Babelfy, where the available multilingual settings generally outperform a prior translation to English (we may recall that in Table 5.6, Babelfy performed best for texts other than English). We further note that the translation results are generally competitive with those for the original English text – shown below the name of the system for the Relaxed and Strict datasets – even slightly outperforming those results in some cases. We also observe from the generally poor (EN,_) results that translation is important; in other words, one cannot simply just apply an EL system configured for English over another language and expect good results.

To give a better impression of the results obtained from the second experiment, in Figure 5.1, for the selected systems, we show the following aggregations: (1) *Calibrated* (_, _): the mean Micro- F_1 score across the four non-English languages with the EL system configured for that language; (2) *Translation* (EN,EN_t): the mean Micro- F_1 score across the four non-English languages with the text translated to English and the EL system configured for English; (3) *English* (EN,EN): the (single) Micro- F_1 score for the original English text.

¹⁶To compute Precision, Recall and F_1 , we do not require true negatives.

¹⁷... it also implies that it outperforms running English EL on text in the original language, though this is hardly surprising and just presented for reference.

Table 5.7: Micro F_1 scores for systems performing EL with respect to the VOXEL dataset. For each system and each non-English language, we show the results of three experiments: first, for $(_,_)$ the system is configured for the same language as the input text; second, for (EN,EN_t) , the system is configured for English and applied to text translated to English from the original language (EN,EN) ; third, for $(\text{EN},_)$, the system is configured for English and run for the text in the current (original) language. Below the name of each system, we provide the relaxed and strict results for the English text. Underlined results indicate the best of the three configurations for the given system, language and dataset variant (e.g., the best for the columns of three values). The best result for each system across all variations (excluding English input) is bolded.

		Relaxed				Strict			
		DE	ES	FR	IT	DE	ES	FR	IT
Babelfy _R (0.545,0.319)	$(_,_) $	<u>0.523</u>	0.541	0.493	<u>0.504</u>	<u>0.344</u>	<u>0.362</u>	0.309	<u>0.365</u>
	(EN,EN_t)	<u>0.507</u>	<u>0.515</u>	<u>0.505</u>	<u>0.501</u>	<u>0.298</u>	<u>0.298</u>	<u>0.314</u>	<u>0.301</u>
	$(\text{EN},_) $	0.215	0.170	<u>0.195</u>	0.140	0.253	0.239	<u>0.220</u>	0.179
Babelfy _S (0.308,0.567)	$(_,_) $	0.279	<u>0.325</u>	0.290	<u>0.311</u>	<u>0.572</u>	<u>0.611</u>	<u>0.583</u>	0.616
	(EN,EN_t)	<u>0.311</u>	<u>0.309</u>	<u>0.322</u>	<u>0.303</u>	0.518	<u>0.523</u>	<u>0.559</u>	<u>0.532</u>
	$(\text{EN},_) $	<u>0.201</u>	0.179	<u>0.189</u>	0.137	<u>0.376</u>	<u>0.372</u>	<u>0.395</u>	0.258
DBpedia Spotlight (0.466,0.707)	$(_,_) $	0.400	0.331	0.240	0.342	0.510	0.477	0.481	0.653
	(EN,EN_t)	<u>0.441</u>	<u>0.454</u>	<u>0.464</u>	<u>0.449</u>	<u>0.696</u>	<u>0.694</u>	<u>0.721</u>	0.729
	$(\text{EN},_) $	0.209	0.161	<u>0.180</u>	<u>0.188</u>	<u>0.374</u>	<u>0.259</u>	<u>0.326</u>	<u>0.323</u>
FREME (0.407,0.708)	$(_,_) $	0.282	0.302	0.268	0.373	0.483	0.583	0.479	0.726
	(EN,EN_t)	<u>0.404</u>	<u>0.403</u>	<u>0.401</u>	<u>0.408</u>	<u>0.701</u>	<u>0.713</u>	<u>0.692</u>	<u>0.711</u>
	$(\text{EN},_) $	0.166	<u>0.183</u>	<u>0.196</u>	<u>0.222</u>	<u>0.190</u>	<u>0.338</u>	<u>0.342</u>	<u>0.374</u>
TagME (0.462,0.327)	$(_,_) $	0.414	–	–	–	0.272	–	–	–
	(EN,EN_t)	<u>0.431</u>	0.450	<u>0.441</u>	<u>0.439</u>	<u>0.330</u>	<u>0.333</u>	<u>0.321</u>	<u>0.336</u>
	$(\text{EN},_) $	<u>0.188</u>	<u>0.181</u>	<u>0.200</u>	<u>0.148</u>	<u>0.212</u>	<u>0.202</u>	<u>0.197</u>	<u>0.164</u>
THD (0.392,0.625)	$(_,_) $	0.241	–	–	–	0.546	–	–	–
	(EN,EN_t)	<u>0.394</u>	<u>0.392</u>	<u>0.386</u>	<u>0.387</u>	<u>0.597</u>	<u>0.620</u>	<u>0.595</u>	0.623
	$(\text{EN},_) $	<u>0.207</u>	<u>0.175</u>	<u>0.217</u>	<u>0.174</u>	<u>0.251</u>	<u>0.332</u>	<u>0.403</u>	<u>0.352</u>

From this figure, we can see that translation is comparable to native English EL, and that translation often considerably outperforms EL in the original language.

We highlight that using translation to English, the result will be an annotated text in English rather than the original language. However, given that translation is done per-sentence, the EL annotations for the translated English text could potentially be “mapped” back per sentence to the text in the original language; at the very least, the translated English annotations would be a useful reference.

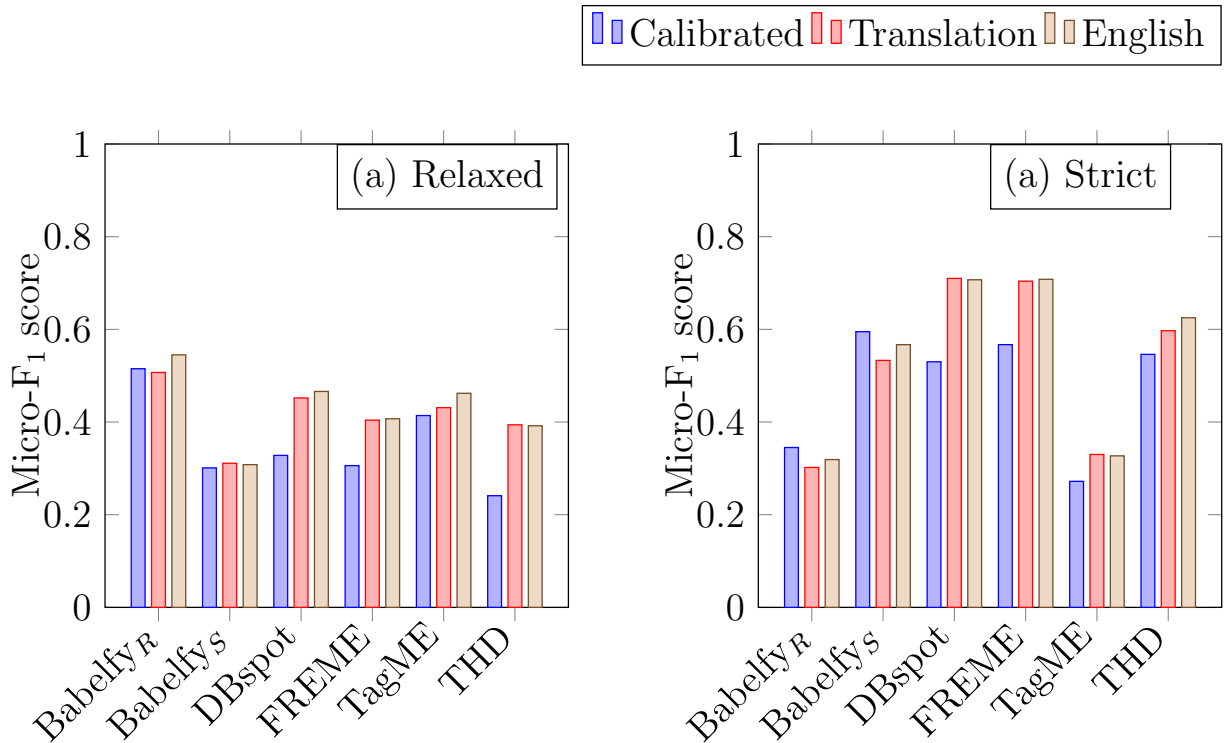


Figure 5.1: Summary of the Micro- F_1 results over VoxEL Relaxed/Strict for the translation experiments, comparing mean values for setting the EL system to the language of the text (*Calibrated*), translating the text to English first (*Translation*), and the corresponding F_1 score for EL over the original English text (*English*)

5.4 Translating across languages

We conduct experiments using VoxEL to compare the behavior of the four aforementioned multilingual EL systems for the five different languages offered by the dataset: German (DE), English (EN), Spanish (ES), French (FR) and Italian (IT). All systems were configured with their default parameters, except BabelFy, which allows to select a more strict or more relaxed notion of entity; we study the performance of both, denoted henceforth as BabelFy_S and BabelFy_R respectively. Aside from testing EL over the text in its native language, we also include results for EL applying machine translation – namely Google Translate¹⁸ – from each language of VoxEL to the other four languages; the purpose of this approach is to simulate an EL approach supporting one language and see if EL performs competitively when input text is translated from other languages.

The results are given in Table 5.8, where we present the F_1 -measure for various configurations. On the left we present the system and language configured. At the top of the table

¹⁸<https://translate.google.com>; June 1st, 2018.

Table 5.8: Comparison of EL systems for native and translated texts (F_1 measure)

		Relaxed					Strict				
		DE \rightarrow	EN \rightarrow	ES \rightarrow	FR \rightarrow	IT \rightarrow	DE \rightarrow	EN \rightarrow	ES \rightarrow	FR \rightarrow	IT \rightarrow
Babelfy _R	$\rightarrow DE$	0.523	0.498	0.495	0.492	0.490	0.344	0.342	0.365	0.369	0.367
	$\rightarrow EN$	0.508	0.545	0.515	0.506	0.502	0.299	0.319	0.299	0.315	0.301
	$\rightarrow ES$	0.525	0.558	0.541	0.552	0.548	0.344	0.356	0.362	0.357	0.348
	$\rightarrow FR$	0.493	0.485	0.502	0.493	0.493	0.332	0.331	0.342	0.309	0.341
	$\rightarrow IT$	0.513	0.527	0.512	0.533	0.504	0.366	0.379	0.377	0.378	0.365
Babelfy _S	$\rightarrow DE$	0.279	0.271	0.275	0.285	0.273	0.572	0.584	0.589	0.606	0.588
	$\rightarrow EN$	0.312	0.308	0.309	0.323	0.304	0.518	0.567	0.523	0.559	0.533
	$\rightarrow ES$	0.318	0.327	0.325	0.334	0.336	0.577	0.607	0.611	0.610	0.590
	$\rightarrow FR$	0.301	0.299	0.312	0.290	0.310	0.574	0.601	0.608	0.583	0.606
	$\rightarrow IT$	0.306	0.319	0.318	0.321	0.311	0.604	0.634	0.640	0.638	0.616
DBpedia Spotlight	$\rightarrow DE$	0.400	0.139	0.177	0.155	0.166	0.510	0.220	0.292	0.248	0.280
	$\rightarrow EN$	0.442	0.466	0.454	0.465	0.450	0.697	0.707	0.695	0.722	0.730
	$\rightarrow ES$	0.159	0.121	0.373	0.130	0.199	0.292	0.209	0.513	0.234	0.350
	$\rightarrow FR$	0.176	0.177	0.181	0.314	0.180	0.245	0.252	0.252	0.464	0.255
	$\rightarrow IT$	0.184	0.163	0.221	0.158	0.382	0.272	0.219	0.335	0.223	0.601
FREME	$\rightarrow DE$	0.282	0.072	0.132	0.114	0.160	0.483	0.154	0.240	0.179	0.261
	$\rightarrow EN$	0.401	0.407	0.402	0.397	0.406	0.700	0.708	0.715	0.694	0.713
	$\rightarrow ES$	0.174	0.117	0.302	0.147	0.232	0.319	0.231	0.583	0.269	0.417
	$\rightarrow FR$	0.167	0.143	0.169	0.268	0.214	0.287	0.278	0.314	0.483	0.322
	$\rightarrow IT$	0.164	0.127	0.205	0.136	0.373	0.321	0.253	0.413	0.256	0.726
TagME	$\rightarrow DE$	0.414	0.100	0.127	0.119	0.124	0.272	0.122	0.153	0.137	0.152
	$\rightarrow EN$	0.432	0.462	0.450	0.442	0.440	0.331	0.327	0.334	0.321	0.336

we present the Relaxed and Strict versions of the dataset, where for each version, we present the language of the input text, which is machine translated to the configured language; for example, row $\rightarrow ES$, column $DE \rightarrow$, gives the result for a German input text translated to Spanish ($DE \rightarrow ES$) and processed by the given EL systems configured for Spanish. Where input and translated languages coincide, we use the input text directly (such results are indicated with boxes). The best result per column for each dataset version and system is presented in bold. TagME supports English and German only.

In Table 5.8, we see that DBpedia Spotlight, FREME and TagMe often perform markedly better when the input text is either in English, or translated to English; the one exception to this trend is that FREME performs slightly better over the untranslated Italian text in the Strict version of the dataset than over the translated English text. On the other hand, Babelfy generally performs best for (translated) Spanish texts in the Relaxed version, and (translated) Italian texts in the Strict version, though performance across languages is more balanced in general than for the former systems. These results suggest that prior machine translation makes little difference in the case of Babelfy, but markedly improves the performance of other systems when dealing with non-English texts; the reasons for this may include the quality of language-specific components, the richness of KB information available for a particular language, etc.

It is important to highlight in such cases that the output of the EL process after translation is still in the translated language; e.g., if we process text in French by translating it to English and performing EL configured for English, we may get better results, but the output text is in English, not French. But we put forward that given (1) a high(er) quality annotation in the translated English text, (2) a sentence-to-sentence correspondence between the French and translated English text, and (3) cross-language links provided by KBs; it would not be difficult to “transfer” the annotations back to the original French text.

In any case, these results raise the question of what role machine translation should play for EL, and indeed, in what circumstances it makes sense to develop multilingual EL systems, and in what circumstances it makes sense to develop monolingual EL systems with *a priori* translation. Addressing such questions is important for EL to reach its full potential and be applied for languages other than English.

Chapter 6

Fine-Grained Entity Linking

An issue that arose during the labeling of the VoxEL dataset described by the previous chapter was how to deal with the lack of consensus on the goals of the EL task, as was also observed by the questionnaire of Chapter 4. We initially addressed this issue by defining both a strict and relaxed notion of entity annotation. However, we believe that in general there are more categories of annotations than these two. In this chapter we propose a finer-grained classification to address this lack of consensus in more detail.

6.1 Fine-Grained Categories

Following the discussion of EL design issues by numerous authors [88, 169, 43, 78, 141] and the results of the questionnaire described in Chapter 4, we propose a fine-grained categorization of EL annotations to make explicit the different types of entity mentions and links that the EL task may consider, which can subsequently be used for the development of EL systems, their evaluation, or indeed, to configure them for application in a given setting. The categories are shown in Figure 6.1. The overall scheme has four distinct dimensions (described in more detail presently): `BASE FORM`, `PART OF SPEECH`, `OVERLAP` and `REFERENCE`. In order to label an EL annotation, we propose that precisely one leaf category (a category without children, shaded in Figure 6.1) should be selected from each dimension, giving four labels per annotation.

The categorization scheme was designed in parallel with the labeling of three EL datasets (described in Section 6.3), with the scheme being extended until it was sufficient to capture all of the cases that we encountered in these datasets. However, the categorization scheme should not be considered complete; for example, in the case of applying EL to Twitter, further categories to cover user mentions, hashtags, misspelled names, etc., might be of interest; the scheme we propose could be extended in future along such lines. Conversely, we do not claim that the EL task *should* always consider all of the annotations covered by this scheme; rather the goal is to capture the types of annotations that *could* be considered by the EL task. We now discuss each of the four dimensions of the scheme in turn.

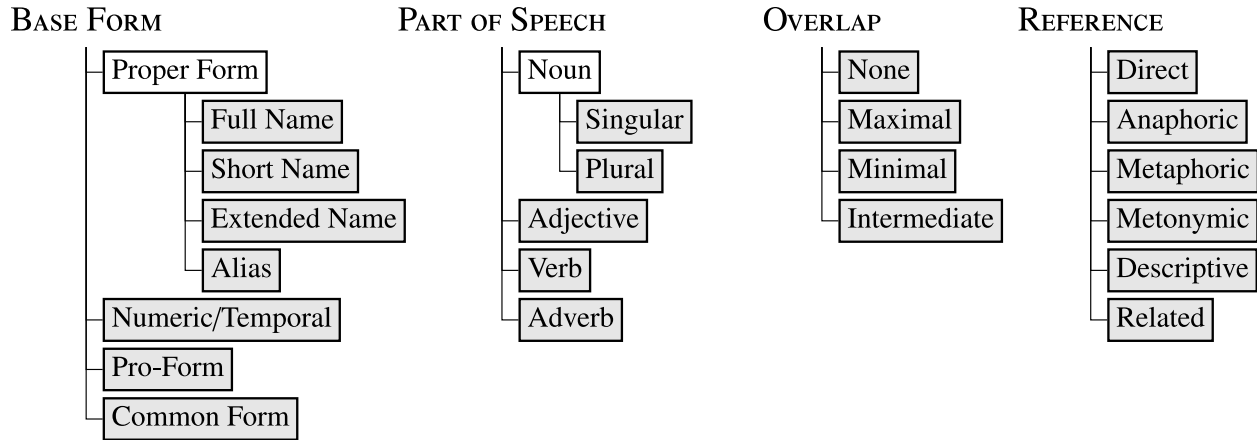


Figure 6.1: EL categorization scheme with concrete alternatives (leaf-nodes) shaded for each dimension

6.1.1 Base Form

The BASE FORM dimension of the scheme refers to the general form of the mention; more specifically, it indicates if the mention refers to one of the following categories:

- **Proper Form**: Denotes a mention based on a name, i.e., based on a proper noun; note however that not all such mentions are nouns, as in the case of “Russian” which, though it may be an adjective, is based on the name of the country, and is thus categorized as a proper *form*. For an annotation in this category, one of the following more specific categories must be selected, based on the primary label of the linked entity in the KB.¹
 - **Full Name**: Denotes that the mention corresponds to the primary label of the entity in the KB, or is a minor variation thereof²; for example, “Russia”, “RUSSIA” or “Russian” referring to `wiki:Russia`.
 - **Short Name**: Denotes that the mention corresponds to an abbreviated form of the primary label of the entity in the KB or an abbreviation of a substring/superstring of this primary label; for example, “M. Jackson”, “Jackson”, “Micheal”, “M.J.”, “M.” or “MJJ” referring to `wiki:Michael_Jackson`.
 - **Extended Name**: Denotes that the mention corresponds to an extended form of the primary label of the entity in the KB; for example, “Michael Joseph Jackson”, “Michael J. Jackson”, “Micheal ‘the King of Pop’ Jackson”, etc., referring to

¹For the more specific sub-categories, we assume that the KB has a primary label in a particular language; this is true of Wikipedia, DBpedia, Freebase, Wikidata and YAGO. In the absence of a particular KB, we recommend to use Wikipedia’s primary labels by default as they are shared by DBpedia and YAGO; these are the local names of the URLs of the corresponding entity article without parenthetical expressions added for disambiguation; for example the primary label for `wiki:Joe_Jackson_(manager)` is “Joe Jackson”.

²More specifically, we consider that the (case-normalized) lemmas of each word in the mention and the primary label correspond in the same order.

wiki:Michael_Jackson.³

- **Alias**: Denotes that the mention – though a proper form – does not correspond to the primary label of the entity per one of the previous three categorizations; for example, “Jackson, Michael” or “King of Pop” referring to wiki:Michael_Jackson.
- **Numeric/Temporal**: Denotes that the mention names a specific temporal or numeric form; for example, “2014”, “fourteen”, “May”, etc., but not “next year”.
- **Pro-Form**: Denotes that the mention is a (simple) pronoun, pro-adjective, etc., that refers (through coreference/anaphor) to a named entity; for example, linking “he” or “his” to wiki:Michael_Jackson.
- **Common Form**: Denotes that the mention is not one of the above categories; such mentions may refer to common entities (e.g., “interview”, “gas”, etc.) or to named entities (e.g., “he and his four siblings”, “his father”, etc.).

6.1.2 Part of Speech

The *Part of Speech* dimension of the scheme denotes the grammatical function of the head word of the mention in the sentence; it includes six categories (five leaves), as follows:

- **Noun**: Denotes a mention whose head term is a (proper or common) noun; for example, “Russia”, “Jackson”, “siblings”, “the capital of Russia”, etc.
 - **Singular**: Denotes that the head noun of the mention is singular; for example, “Russia”.
 - **Plural**: Denotes that the head noun of the mention is plural; for example, “siblings”.
- **Adjective**: Denotes a mention whose head term is an adjective; for example, “Russian”, “covalent”.
- **Verb**: Denotes a mention whose head term is a verb; for example, “assassinated”, “genetically modifying”.
- **Adverb**: Denotes a mention whose head term is an adverb; for example, “exponentially”, “Socratically”.

³The mention should contain the (case-normalized) lemmas of the primary label in order, possibly interrupted by other lemmas.

6.1.3 Overlap

The *Overlap* dimension indicates whether or not the text of a mention overlaps with that of other mentions, and if so, in what way; we illustrate its four categories for the text “The New York City Police Museum is located in Manhattan.”:

- **None**: Denotes a mention whose text does not overlap with that of another mention; for example, “Manhattan”.
- **Maximal**: Denotes a mention whose text contains an inner mention but is not contained in another mention; for example, “New York City Police Museum”.
- **Minimal**: Denotes a mention contained in another mention but that does not itself contain another mention; for example, “New York”, “Museum”, “Police”.
- **Intermediate**: Denotes a mention that does not fall into one of the above categories; for example, “New York City Police” is contained by and contains other mentions.

6.1.4 Reference

The *Reference* dimension indicates the manner in which the mention makes reference to the linked KB entity [154]. This dimension is flat, containing six leaf categories:

- **Direct**: Denotes a mention that makes direct reference to an entity, be it by name, abbreviation, alias, etc. in the case of named entities (e.g., “Jackson”, “King of Pop”, “Russian”), or a recognized surface form for a common entity (e.g., “interview”, “genetically modifying”).
- **Anaphoric**: Denotes a mention that uses a pro-form to refer to a named entity; for example, “he”, “his”, etc., referring to `wiki:Michael_Jackson`.
- **Metaphoric**: Denotes a mention that figuratively references a KB entity for their characteristics; for example “[King] of Pop” referring to `wiki:King`, or “the British version of [Trump]” referring to `wiki:Donald_Trump`.
- **Metonymic**: Denotes a mention that references a given KB entity by common association; for example “Moscow” being used to refer to `wiki:Government_of_Russia` or “Portugal” being used to refer to `wiki:Portugal_national_football_team`.
- **Descriptive**: Denotes a mention that refers to a named entity by description; for example, “he and his four siblings” referring to `wiki:Jackson_5`, “his father” referring to `wiki:Joe_Jackson_(manager)`, “Russia’s capital” referring to `wiki:Moscow`, “Hendix’s band” referring to `wiki:The_Jimi_Hendrix_Experience`, etc.

- **Related**: Denotes a mention that does not fall into one of the above categories. This category includes mentions for which the precisely matching entity does not exist in the KB, but a closely-related one does; for example, “the Russian [daily]” being linked to `wiki:Newspaper`.⁴ This category complements metonymic references, where “[Moscow] will supply” will also be linked to `wiki:Moscow` in an annotation with the related category. We also use this category to include entities related by hierarchies such as spatial hierarchies (e.g., New York City vs. Manhattan) or organisational hierarchies (e.g., India hockey team vs. the Olympics-2000 India hockey team).


6.2 Fine-Grained EL Format

In Section 3.5, we described formats for serializing EL datasets; however, none of the existing formats support our newly defined categorization scheme. In order to allow these categories to be used in EL datasets, we construct a novel vocabulary that allows for them to be used in conjunction with RDF formats. We then use this vocabulary to extend the existing NIF format; we further describe a convention for how multiple links can be added to a single mention in the NIF format. We first describe the vocabulary and then the NIF extension; thereafter we introduce a tool we have developed to aid in the creation and validation of EL datasets in this format.

6.2.1 Vocabulary

We show the Fine-Grained Entity Linking (FEL) vocabulary in Figure 6.2, with newly defined terms using the `fel:` prefix. The categories of Figure 6.1 are defined as classes, forming a sub-class hierarchy. We follow a set of rules proposed by Baker et al. [4] with respect to the description, preservation and governance of the vocabulary. They propose two types of rules: local ones act in favor of the quality of the vocabulary while global ones are aimed at governing their accessibility to third parties.

Towards fulfilling the local rules, our vocabulary has the following properties:

- Each category is resolvable by a unique and machine-readable URI.
- We use the DOAP⁵ vocabulary to specify the maintainer.
- We provide labels and definitions for each category in natural language to improve human readability.
- We publish the vocabulary under a CC-BY 3.0  license⁶ encouraging its re-use.

⁴In the case of Wikipedia, for example, redirects are sometimes provided to related entities if the target entity does not exist; other times the target entity may point to a section of the article of the related entity with a fragment id.

⁵<http://usefulinc.com/ns/doap>

⁶<https://creativecommons.org/licenses/by/3.0/>

- Further changes will be managed with the GitHub⁷ platform. We separate changes according to their significance. Minor changes (e.g., spelling, punctuation, orthography of comments, etc.) and the incorporation of triples that do not change the semantics of the vocabulary will be addressed in the current namespace. On the other hand, any change with a negative impact to the current semantics will be separated into a new namespace.
- We re-use existing terms from well-known vocabularies; in particular we map our vocabulary classes with similar ones in existing vocabularies using SKOS links [103] (as shown in Figure 6.2).

To satisfy global rules, we submit the FEL vocabulary to the Linked Open Vocabularies system [165]⁸: a catalog of reusable vocabularies that serves as a monitoring tool; the goal is to allow our vocabulary to be discovered by interested third parties, as well as to track its usage over time. Along these lines, we also fulfill the following criteria:

- We use the VoID⁹ vocabulary to allow data providers to discover what terms the vocabulary uses.
- We guarantee the persistence of our URIs storing our vocabulary on a server¹⁰ of the DCC, University of Chile. However, to deal with any problem in the future about institutional persistence, we use a permanent identifier provided by W3C Permanent Identifier Community Group¹¹ which can be redirected to another destination.
- To embrace the “safety through redundancy” principle [4] which advocates for mirroring information online, we make a second copy available in a GitHub repository¹².

6.2.2 Extending NIF

One benefit of using RDF as a core data model is that NIF can be readily extended with further class and property terms, as needed. For example, for the purposes of the Wes2015 dataset [168], for Document Retrieval, novel properties and classes (e.g., `si:Query`, `si:result`, `yv:queryId`) were used alongside NIF. We now describe a minor extension to NIF to specify entity annotation categories, entity types, as well as specifying alternative links for a mention.

Per Table 3.1, some EL datasets type annotations according to a list of predefined classes; this practice was prevalent in earlier Named Entity Recognition (NER) works, whose goal was to identify entities of different types but without having to link them to a KB. The entity type

⁷<https://github.com/henryrosalesmendez/fel>

⁸<https://lov.linkeddata.es/dataset/lov/>

⁹<http://vocab.deri.ie/void>

¹⁰<https://cutt.ly/2yEvqp0>

¹¹<https://www.w3.org/community/perma-id/>

¹²<https://github.com/henryrosalesmendez/fel>

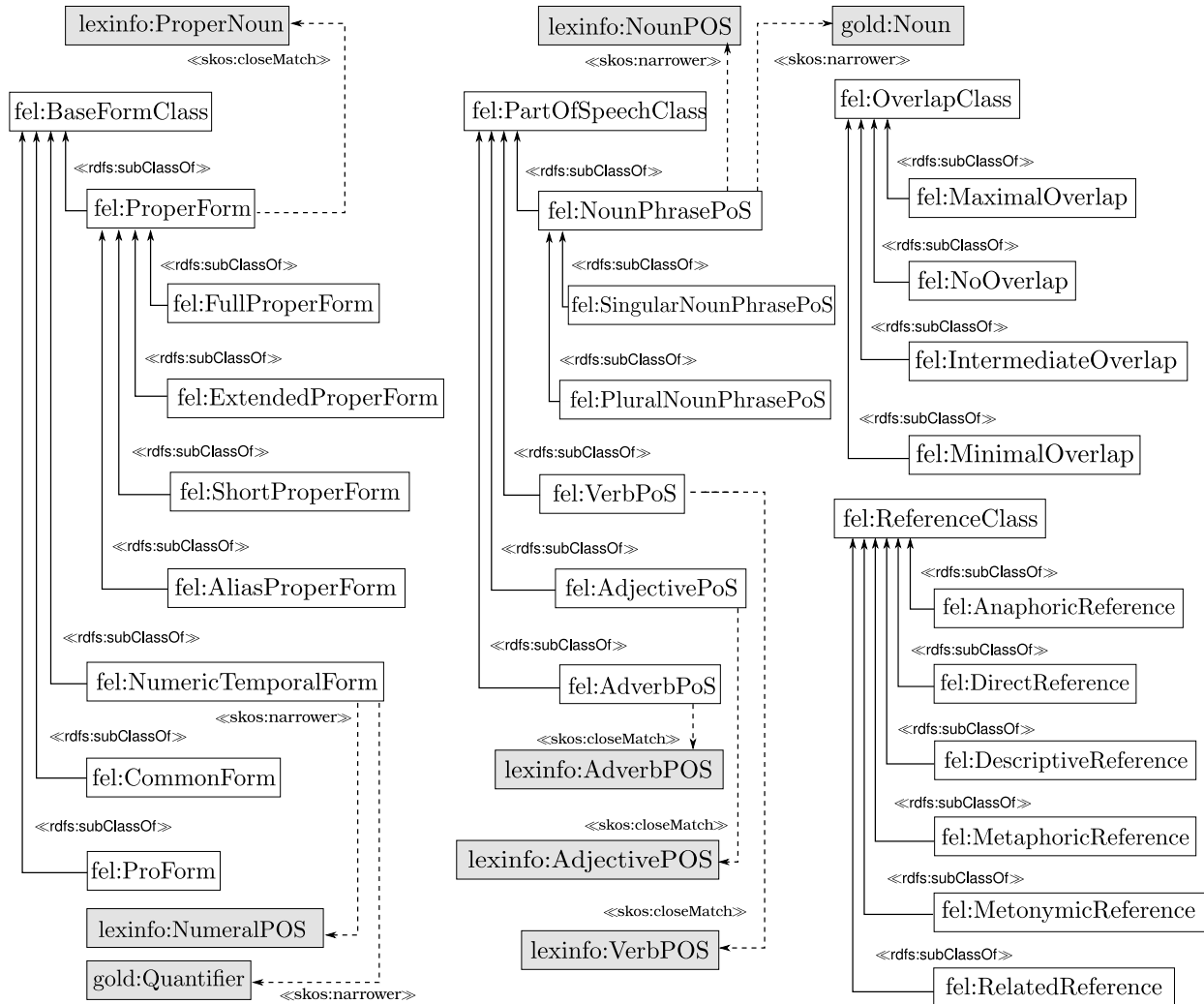


Figure 6.2: Hierarchy of classes belonging to the Fine-Grained Entity Linking vocabulary and its links to external vocabularies.

can be specified in NIF on an annotation with the property `itsrdf:taClassRef`.¹³ However, problematic situations emerge when the same mention may be considered as referring to more than one URI in the KB: although the general expectation is that EL systems will only yield one link per entity mention, multiple links may be acceptable in cases where the context is not enough to fully disambiguate the entity mention, the entity mention is intrinsically ambiguous, or multiple types of entities may be considered correct, per the following two examples:

S2 “Bush was president of the United States of America.”

¹³See example: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/example.ttl>; October 5th, 2019.

S3 “Iran is not capable of developing a nuclear program without Moscow’s help.”

In sentence **S2**, without further context, it remains unclear if the entity mention “**Bush**” refers to the 41st U.S. president George H. W. Bush, *OR* to his son, the 43rd U.S. president; when creating a gold standard for evaluating EL systems, we may thus wish to allow both possibilities. On the other hand, in sentence **S3**, the entity mention “**Moscow**” could be seen as referring to `wiki:Moscow`, the capital of Russia, *OR* perhaps rather as referring to help from the Government of Russia (`wiki:Government_of_Russia`). Hence we may wish to capture multiple links for a given mention.

Conversely, consider the following sentence:

S4 “Barack met Michelle in June 1989; they married three years later.”

If we support coreference in this case, then we may wish to capture that “**they**” refers to `wiki:Barack_Obama` *AND* `wiki:Michelle_Obama`, again requiring multiple links.

Although NIF can support the specification of multiple links, there are no indications on how such cases should be handled. We propose a simple convention, which is to put multiple links on the same annotation in the case of multiple *AND* links, and rather use multiple annotations with the same offset in the case of multiple *OR* links (both can also be combined). Further complications arise, however, when labeling types, where different types may apply to different links; while this would not be a problem for **S2** (both are *Persons*), in **S3**, one link is a *Place* while the other is an *Organization*. Along these lines, we propose to separate the entity type specification from the annotation scope with a triple $s \text{ fel:entityType } o$ for each link in the annotation, where s denotes the KB identifier, not the mention.

In Figure 6.3 we show the annotation of Moscow from sentence **S3** with NIF, displaying two alternative links (*OR*), with two triples specifying the entity type for each alternative; furthermore, we see that `wiki:Government_of_Russia` is indicated as a metonymic reference, while `wiki:Moscow` is indicated as a related reference. On the other hand, Figure 6.4 shows the annotation of the coreference “**them**” from sentence **S4**; in this case, both links are presented on the same annotation. Unlike in the case of *OR* links, we cannot assign different categories for different links in the *AND* case: in all such *AND* cases that we have observed in real datasets, the type of reference is either descriptive or anaphoric (per **S4**), where categories do not change for the different links; this assumption allows us to annotate *AND* cases in a lightweight manner (e.g., without having to introducing further vocabulary or nodes in the annotation).

In summary, our NIF extension includes the following additional features useful for annotating fine-grained EL datasets:

- *Categories*: we include terms to identify categories, such as `fel:FullProperForm`, `fel:NoOverlap`, etc.

Figure 6.3: NIF triples to specify the annotation of “Moscow” from sentence **S3**; we use multiple annotations to denote an *OR* over the links

```

<http://example.org#char=88,94;1>
  a nif:String, nif:Context, nif:Phrase,
  nif:RFC5147String , fel:SingularNounPhrasePoS,
  fel:MetonymicReference, fel:NoOverlap,
  fel:AliasProperForm;
  nif:anchorOf ""Moscow""^^xsd:string;
  nif:beginIndex "88"^^xsd:nonNegativeInteger;
  nif:endIndex "94"^^xsd:nonNegativeInteger;
  itsrdf:taIdentRef </wiki/Government_of_Russia>.

</wiki/Government_of_Russia> fel:entityType
  fel:Organisation .

<http://example.org#char=88,94;2>
  a nif:String, nif:Context, nif:Phrase,
  nif:RFC5147String , fel:FullProperForm,
  fel:SingularNounPhrasePoS, fel:RelatedReference,
  fel:NoOverlap;
  nif:anchorOf ""Moscow""^^xsd:string ;
  nif:beginIndex "88"^^xsd:nonNegativeInteger ;
  nif:endIndex "94"^^xsd:nonNegativeInteger ;
  itsrdf:taIdentRef </wiki/Moscow> .

</wiki/Moscow> fel:entityType fel:Place .

```

- *Typing entities*: the predicate `fel:entityType` can be used to type the entity independently of a mention.

We further propose conventions to represent multiple links on a single mention with OR and AND semantics (or potentially a mix of OR and AND using a disjunctive normal form).

As previously discussed, in the context of other future applications and (F)EL scenarios, it may be of interest to extend our categorization scheme, for example, to consider hash-tags, user mentions, misspellings, hyperlinks, etc.; our vocabulary could be further extended along these lines in a similar fashion to how we extend upon the NIF vocabulary.

Figure 6.4: NIF triples to specify the annotation of “them” from sentence **S4**; we use multiple `itsrdf:taIdentRef` values to denote an *AND* over the links

```

<https://example.org#char=33,37;1>
  a nif:String, nif:Context, nif:Phrase,
  nif:RFC5147String , fel:ProForm,
  fel:PluralNounPhrasePoS, fel:NoOverlap,
  fel:AnaphoricReference ;
  nif:anchorOf """"they""""^^xsd:string;
  nif:beginIndex "33"^^xsd:nonNegativeInteger;
  nif:endIndex "37"^^xsd:nonNegativeInteger;
  itsrdf:taIdentRef </wiki/Michelle_Obama>,
  itsrdf:taIdentRef </wiki/Barack_Obama> .

</wiki/Barack_Obama> fel:entityType fel:Person.
</wiki/Michelle_Obama> fel:entityType fel:Person.

```

6.3 Fine-Grained Entity Annotation

As discussed in Section 3.6, there are varying definitions on the EL task, and varying opinions regarding what should be included or excluded as part of the task. In terms of the datasets described in Section 3.4, while some make their annotation criteria explicit, others do not. When designing our criteria, our overall goal was to capture the types of mentions and links for which there was some support in the results of the questionnaire (see Figure 4.1) as captured by the categories previously outlined; this proven challenging in some cases. We now outline the annotation criteria we applied along these lines. These guidelines aim to be comprehensive in terms of annotating fine-grained EL datasets. The datasets we label – as will be described in Section 6.5 – are published online and further provide thousands of examples of annotations that can be referenced.

- With respect to the entities considered, we aim to adopt an inclusive definition, where we thus take as a base the definition provided by Guo et al. [60], who consider entities that are described by “*a nonambiguous, terminal page (e.g., The Town (the film)) in Wikipedia (i.e., a Wikipedia page that is not a category, disambiguation, list, or redirect page)*”. We refine this definition slightly, as follows:
 - We explicitly exclude Wikipedia articles that refer to *syntactic entities* – i.e., entities denoting their own syntactic form – which includes articles about names (e.g., `wiki:Jackson_(name)`), and symbols (e.g., `wiki:Exclamation_mark`). We do, however, include numbers, units, dates, etc.
 - We include named entities not appearing in Wikipedia as emerging entities (aka., *Not In Lexicon (NIL)* entities).

– We explicitly allow overlapping mentions.

- Each annotation is labeled with one leaf-node from each of the four category dimensions outlined in Figure 6.1.
- Entity boundaries are based on the primary label of the Wikipedia page. For example, in the case of the mention “[The Beatles]”, we include the article “The” as the link includes the article: `wiki:The_Beatles`. On the other hand, in the case of the mention “The [BBC]”, we do not include “The” as the link is to `wiki:BBC`. Furthermore, in the case of “President [Putin]”, we do not include “President” in the mention as the link is to `wiki:Vladimir_Putin` (without “President” in the label).
- Per the previous guidelines, different entity boundaries may be used for related entities, which are considered distinct annotations (rather than alternatives linked by OR); for instance, in the text “[The {Guardian}] is owned by [Scott Trust Limited]”, the mention “[The Guardian]” links to `wiki:The_Guardian` (i.e., the newspaper) whose primary label includes “The”, while “{Guardian}” links to `wiki:Guardian_Media_Group` (i.e., the company) whose primary label does not include “The”.
- The primary labels of KB entities may be abbreviations, in which case the corresponding mention falls into the Full Name category; for example, the mention “CNN” has the corresponding entity `wiki:CNN`, and thus will be labeled as a Full Name, rather than a Short Name.
- We only consider pro-forms when they clearly refer to a named entity or an enumeration of named entities in the KB. For example, in the text “The bill was passed in 2014; [it] was ...” we will not annotate “it” linked to `wiki:Bill_(law)`¹⁴, but rather only annotate the mention if it can be resolved from context to a specific bill, such as the `wiki:Ukraine_Support_Act`. In the sentence “Barack met Michelle in June 1989; [they] married three years later.”, we will link “they” to `wiki:Barack_Obama` AND `wiki:Michelle_Obama` as both are named entities.
- Descriptive mentions are likewise only annotated when pointing to named entities. Defining the boundaries of descriptive mentions proved challenging, where we settled on annotating noun phrases up to a participle clause. In the case of “he was managed by [his father]” linked to `wiki:Joe_Jackson`, we include “his” as part of the annotation; likewise in the case of “he was visited by [the president of Russia]” linked to `wiki:Russia` we include the definite article “the” and the clause “of Russia” in the mention.¹⁵ We argue that the inclusion of the definite article in such cases helps to distinguish general and specific links; for example, with the text “The World Cup was held in Russia”,

¹⁴We argue that “it” does not refer to `wiki:Bill_(law)` here, but rather refers to *something that is a wiki:Bill_(law)*.

¹⁵This decision was made after the questionnaire was conducted; for this reason, Table 4.1 uses an old convention for “.. the [Russian President].”; under the final convention, “.. [the Russian President].” would be considered the mention for `wiki:Vladimir_Putin`.

we link “The World Cup” to `wiki:2018_FIFA_World_Cup`, while “World Cup” is linked to `wiki:FIFA_World_Cup`. In the case of “[The bill] passed by Congress in 2014 in order to provide aid to Ukraine received bipartisan support.” linked to `wiki:Ukraine_Support_Act`, we cut the mention before the participle clause “passed by ...”; on the other hand, in the case of “[The passed bill] received bipartisan support.”, we maintain the simple participle “passed”.

- We do not annotate descriptive annotations that result in a reflexive (e.g., “is”) relation or an adjacent link. For example, in the text “His father was Joe Jackson, ...”, we do not annotate “His father” as it would correspond to the reflexive relation “*Joe Jackson was Joe Jackson, ...*”; furthermore, in the text “His father, Joe Jackson, was ...”, we do not annotate “His father” as it corresponds to the redundant phrase “*Joe Jackson, Joe Jackson, was ...*”.
- As aforementioned, in meronymic cases such as “[Moscow] will supply ...”, we add alternative (*OR*) links: a link to `wiki:Government_of_Russia` with the *Meronymic* category, and a link to `wiki:Moscow` with the *Related* category.
- If a mention in the text does not have a corresponding entity in the KB, we label it if and only if the mention is a proper form referring to a named entity; these are known as Not In Lexicon (NIL) or emerging entities. We link such entities to a reserved IRI used by Röder et al. [135], namely <http://en.wikipedia.org/wiki/NotInLexicon>. The PART OF SPEECH and OVERLAP categories follow the standard rules. The BASE FORM and REFERENCE categories should be selected with respect to how the NIL entity would most likely be described by the KB if added in future; for example, a mention “Smith” referring to a person not in the KB would be labeled as a Short Name (assuming the KB typically provides full names), and as a Direct reference.

Systematically covering all cases with support in the questionnaire – including more complex cases such as the descriptive mention “he and his four siblings” (0.50) – thus requires a complex set of guidelines. Though we argue that such guidelines are necessary to subsume the varying perspectives regarding the EL task, they do greatly complicate the annotation process when compared with (for example) only annotating named entities. We now describe the process of labeling our selected three datasets, providing statistics on the resulting annotations.

6.4 NIFify

Even without considering fine-grained classification, the creation of gold standards for EL is still a complex, error-prone and time-consuming work; hence a number of tools have been proposed to help experts in this task. Röder et al. [136] craft three NIF datasets from texts written in English and German that were tagged manually using their own tool, but to the best of our knowledge the tool is not openly available. Looking for mistakes

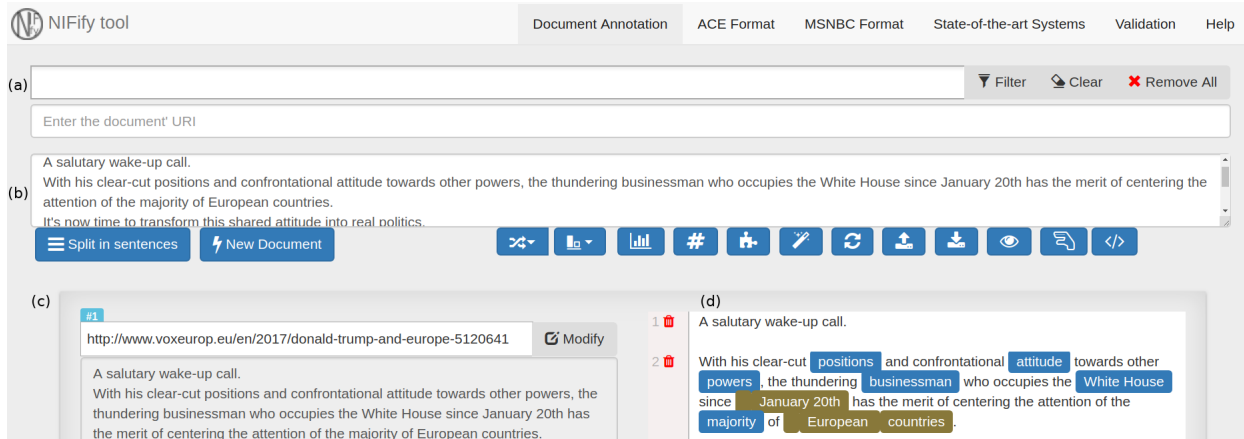


Figure 6.5: The main view of NIFify showing: (a) the class-reference input to filter annotations; (b) the document text input; (c) the mention identification field; and (d) the annotation visualization.

in datasets, Kunal et al. [78] propose guidelines to validate EL datasets, providing the EAGLET system that checks a variety of quality rules, helping experts to reduce errors; however, some important errors, such as verifying that the target of a link is not a redirect page, are not covered. On the other hand, other works have focused on standardizing the assessment process, providing benchmarking suites (e.g., GERBIL [164], Orbis [118]) that can quickly compare results for state-of-the-art EL systems against a variety of datasets. More generally, all of these NIF operations – creating, validating and performing experiments with EL datasets – have, to the best of our knowledge, been addressed as independent systems.

Here, we thus describe NIFify: a tool that simultaneously supports the creation, visualization, and validation of NIF datasets, as well as the comparison of EL systems. With our tool – shown in Figure 6.4 – we include some functionalities not covered by previous approaches for creating, modifying and validating NIF datasets. Additionally, we allow for visualizing the results of EL systems at both a sentence and document level.

6.4.1 NIF Construction

A number of EL datasets have either been computed from existing sources, or computed automatically. For example, DBpedia Abstracts is too large for human labeling to be feasible.¹⁶ On the other hand, the recently proposed BENGAL tool [117] adopts a creative strategy for automatically generating gold standard datasets: rather than start with text, the authors propose to start with facts about entities from structured datasets (in RDF) and use verbalization components to convert these facts to text, recording which entities

¹⁶Details of the annotation process are not provided, but we assume it uses links already present in the corresponding Wikipedia texts.

are used to generate which sentences; while this approach has the benefit of being able to generate very large and accurate gold standards, how representative the generated text is of real-world corpora depends on the quality of the verbalization component.

On the other hand, per Table 5.2, most datasets are constructed with manual intervention, and a number of systems have been proposed to help in this process. Addressing these limitations, we propose NIFify: an open source tool that provides end-to-end support for EL annotation, including the import of text corpora¹⁷; the import (including the conversion of MSNBC formats to NIF) of existing EL datasets; the addition and revision of annotations; custom tagging systems for annotations; visualizations of annotations; overlapping mentions; and finally, visualisations of the results of EL systems over the resulting dataset. The tool requires no installation and can be used either online or offline in a browser¹⁸. For space reasons, rather than describe all the features of NIF, we focus on two group of features of particular importance to NIFify: *validation* and *result visualization*.

6.4.2 Validation

Validation is a crucial step to help human experts ensure the production of a ground truth for gold standards, and EL datasets are no exception. Legacy EL datasets have been observed to contain errors or design choices that may affect the results of evaluation [43, 78, 141]; furthermore, target KBs may evolve, rendering some links obsolete.

Erp et al. [43], analyze characteristics of seven EL datasets and find biases introduced by the decisions taken in the annotation process; they highlight the need for a more standard creation of datasets. Jha et al. [78] propose a set of validation rules and propose the EAGLET system to check these rules when constructing EL datasets; however, these rules are sometimes dogmatic, considering, for example, overlapping mentions to be errors when they are considered valid by other definitions [141]; furthermore, EAGLET requires execution on a command-line to highlight errors in the visualization, rather than being supported by the interface.

NIFify allows for detecting possible errors present in terms of the mentions and the identifiers to which they are linked; specifically, the following rules are checked:

- **SPELLING ERROR (SE)**: Mentions should neither start nor end in the middle of a word.
- **LINK ERROR (LE)**: When linking to Wikipedia or DBpedia, identifiers should be the URLs/IRIs corresponding to an unambiguous, non-redirect page on Wikipedia.
- **FORMAT ERROR (FE)**: We check the consistency of the NIF representation with two sub-rules:

¹⁷https://users.dcc.uchile.cl/~hrosales/MSNBC_ACE2004_to_NIF.html; Jan. 1st, 2019

¹⁸https://github.com/henryrosalesmendez/NIFify_v2; January 1st, 2019

Table 6.1: Errors found in current NIF datasets; the last dataset was labeled by us

Dataset	SE	LE	FE	CE
DBpedia Spotlight	8	23	4	–
N3-RSS 500	1	34	–	–
Reuters 128	4	71	–	–
News-100	9	1515	–	–
Wes2015	–	609	–	–
VoxEL	–	8	–	–

- Annotations are typically assigned a subject IRI `http://example.org#char= x , y` , where x and y should correspond with the values given for `nif:beginIndex` and `nif:endIndex` respectively.
- The substring identified by these positions should correspond with that denoted by the `nif:anchorOf` property.
- **CATEGORY ERROR (CR):** For those datasets with classes specified by the predicate `itsrdf:taClassRef`, NIFify allows the specification of custom rules in order to detect inconsistencies in the annotation classes. For example, the classes `dbo:Person` and `dbo:Event` should not be present on the same annotation as they are disjoint: an entity is typically not a person and an event at the same time.

NIFify then encodes rules to detect these errors and thus validate EL datasets. In order to test the prevalence of these errors in existing datasets, we ran NIFify’s validation over EL datasets currently available in the NIF format (excluding those that we converted ourselves to NIF – MSNBC and ACE2004 – since we resolve such errors as part of the conversion). In Table 6.1, we show the results of this validation process, where we can observe that all datasets considered contain errors of at least one type.

In the majority of the cases, SE errors are introduced in the construction of the dataset with the addition of characters that do not belong to the mention, or on the contrary, leaving out part of a word that completes a mention; for example, in the DBpedia Spotlight dataset, the URI `wiki:Man` is associated with the three characters of the word *performance*. Other SE errors contained in the datasets involve missing spaces between words.

The most frequent type of error encountered in the NIF dataset was LE: this is mainly due to the fact that KBs are constantly evolving, which may affect link consistency. For example, in Wikipedia, pages about specific entities may become disambiguation pages, or redirects to other pages. Such changes explain why our own dataset (VoxEL, created using NIFify) contains such errors: the external KB has evolved since its creation. The News-100 and Wes2015 contain a large number of LE errors beyond what can be explained by the KB

changing: for example, in the Wes2015 dataset, 520 of its LE errors correspond to redirect pages, 48 to disambiguation pages, while the rest do not point to valid pages.

Finally, the only dataset we found with FE-type errors was DBpedia Spotlight, which had problems with its NIF representation. On the other hand, we did not find any errors of type CE.

We have published all errors found online for reference¹⁹. We conclude that most of the validation features of NIFify can help to improve the quality of EL datasets, including to find problems caused by the evolution of a KB over time.

As part of the final review of a dataset, we have further extended NIFify to provide a “validation tree” view, which allows to view the annotations grouped by mention, and thereafter by category, thus helping to ensure that mentions are labeled consistently (where appropriate) across a text. We provide an example in Figure 6.6 for the mention “Tehran”, showing all of its annotations in the ACE2004 dataset. We see that two mentions are labeled with two links, indicating a meronymic reference to the Government of Iran and a related reference to Tehran, while a third mention is linked directly to Tehran.

6.4.3 Result Visualization

Once an EL dataset has been generated, the next step is to evaluate and compare EL systems using the dataset. A number of systems have been proposed to help evaluate and compare EL systems. Cornolti et al. [32] proposed the BAT framework, which they used to compare five EL systems over five datasets. Along similar lines, Usbeck et al. proposed GERBIL [164], which extends the systems and (NIF) datasets supported. However, both frameworks produce comparative metrics, rather than visualizing the actual output of the EL tool(s). Another EL benchmark framework called Orbis [118] was recently proposed that includes visualization of systems’ responses; however, Orbis is not available in the provided URL.²⁰

Given that there is no clear definition for what EL systems should link [141], we argue that metrics like precision and recall may not tell the full story, and that results may be due not only to the quality of the output produced by an EL system, but also whether or not it targets the same types of entities as labeled in the dataset. Comparing EL results with the ground truth labeled in a dataset under construction/revision may even lead to changes in the dataset.²¹ Hence with NIFify we propose a benchmark framework to visualize the results of EL systems over the NIF dataset, highlighting both *true positives* or *false positives*, which allows a more qualitative assessment of both a given EL tool and an EL dataset, possibly in

¹⁹https://users.dcc.uchile.cl/~hrosales/dataset_errors.html; January 1st, 2019.

²⁰<https://github.com/htwchur>; January 1st, 2019.

²¹Of course, we urge caution to ensure that bias is not introduced by adapting a dataset to suit a subset of tools evaluated.

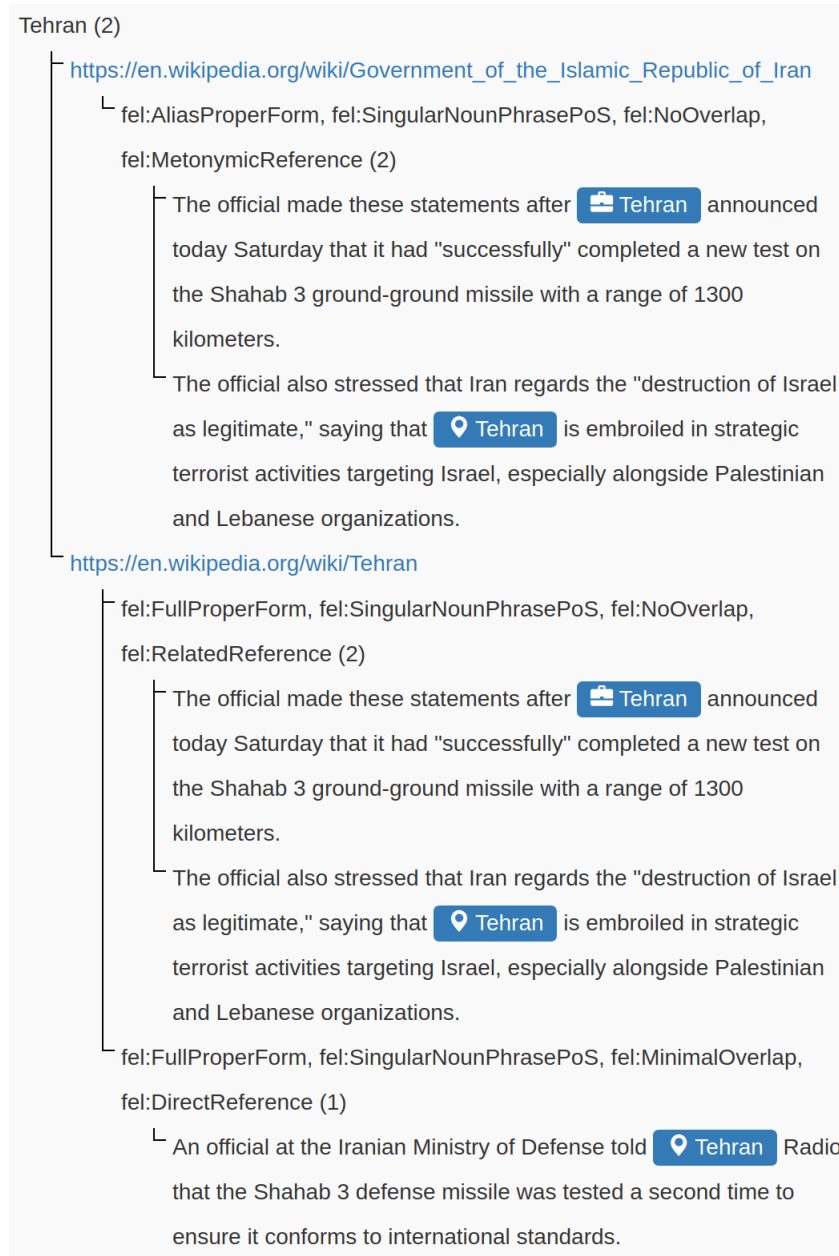


Figure 6.6: NIFify’s validation tree view for the mention “Tehran” in the ACE2004 dataset

the context of a given application. Additionally, NIFify can be used to demo EL systems, offering a visual, friendly user interface.

6.5 Relabeling KORE50, ACE2004 and VoxEL

We relabeled our three selected EL datasets – KORE50, ACE2004, and VoxEL – according to the aforementioned criteria and categorizations. In the case of KORE50 and ACE2004 – which focused on named entities – this required adding (many) novel annotations not considered in the original datasets. It is important to note that when we started the labeling process, our initial criterion was to label the entities of the three datasets per Guo et al.’s definition [60], also including emerging named entities; in other words, we did not have the previously discussed categories and guidelines prepared before we began the process, but rather these were also generated and refined as part of the process. More generally, given that the requirements for relabeling the datasets were not clear at the start of the process, we followed an agile methodology [6] of iterative refinement, involving not only the datasets themselves, but also the categories, the guidelines, and the tool used for annotation.

Specifically, the author began with an initial extension and relabeling of the KORE50 and VoxEL datasets, generating a list of difficult cases – such as descriptive mentions, meronymic references, etc. – that were discussed with his supervisors, leading to a refinement of the categories and guidelines. The two supervisors then iteratively reviewed the annotations produced for these datasets, which were also validated in semi-automated fashion using the extended NIFify tool (see Section 6.4). With consensus reached on these two datasets, the author then began an initial labeling of the larger ACE2004 dataset, highlighting further difficult cases that were discussed with his supervisors and, in some cases, leading to modifications of the categories, guidelines, and all three datasets. Given the time consuming nature of the annotation process, it was decided to limit the relabeling of ACE2004 to the first twenty of fifty-seven documents; we remark, for example, that the number of annotations in these twenty documents increases from 108 in the original data to 3,351 in our fine-grained version. Finally, the datasets were iteratively verified one last time by the author and checked with the tool.²² The resulting datasets – as well as the previously discussed categories and guidelines – reflect the consensus of the three participants. Furthermore, the categories and guidelines were sufficient to cover all cases encountered in the datasets.

Overall, the labeling process was very time consuming (spanning six months), due in part to the iterative refinement of the categories and guidelines, as well as the sheer number of annotations needed to satisfy the modified version of Guo et al.’s definition [60]. In Table 6.2, we provide statistics for the three relabeled datasets, further counting annotations in different categories. Of note is the large quantity of common entities labeled in the ACE2004 and VoxEL datasets; furthermore, we see that most entities do not correspond to the original MUC-6 definitions of entity types. The datasets are available online.²³

²²During the final validation, we also found and fixed a number of issues with the original datasets. Of particular note were some spelling errors in ACE2004 of entity names, e.g., *Stewart Talbot* as a misspelling of *Strobe Talbott*, *Coral Islands* as a spelling variant of *Kuril Islands*, etc.; we decided to keep the original spelling but link to the intended entities in such cases.

²³https://github.com/henryrosalesmendez/categorized_EMNLP_datasets

Table 6.2: Statistics on the three relabeled datasets [138]

	KORE50	ACE2004	VoxEL
Documents	1	20	15
Sentences	50	214	94
Annotations	372	3,351	1,107
Full Name	41	588	227
Short Name	114	307	97
Extended Name	1	8	–
Alias	5	94	15
Numeric/Temporal	17	276	111
Common Form	157	1,974	615
Pro-form	37	107	42
Singular Noun	248	1,943	683
Plural Noun	39	670	182
Adjective	45	501	149
Verb	40	232	85
Adverb	–	5	8
No Overlap	307	2,161	792
Maximal Overlap	23	392	95
Intermediate Overlap	4	62	14
Minimal Overlap	38	736	206
Direct	262	2,280	750
Anaphoric	37	107	42
Metaphoric	8	27	38
Metonymic	3	60	21
Related	54	698	224
Descriptive	8	179	32
Person	117	278	66
Organisation	40	199	120
Place	19	519	168
Miscellany	196	2,352	753

6.6 Fine-Grained Evaluation

Table 6.3: Results per category for Babelfy (strict/relaxed), TagME, DBpedia Spotlight, AIDA and FREDME on the unified dataset [138]

	A	B_s			B_r			T			D			A			F		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Full Mention	766	0.93	0.46	0.61	0.75	0.53	0.62	0.82	0.59	0.69	0.84	0.57	0.68	0.78	0.57	0.66	0.82	0.55	0.65
Short Mention	497	0.44	0.16	0.23	0.37	0.24	0.29	0.54	0.44	0.48	0.5	0.30	0.37	0.50	0.36	0.42	0.39	0.28	0.33
Extended Mention	9	1.00	0.56	0.71	0.83	0.56	0.67	1.00	0.44	0.62	1.00	0.44	0.62	1.00	0.44	0.62	0.80	0.44	0.57
Alias	112	0.56	0.16	0.25	0.33	0.21	0.25	0.52	0.32	0.40	0.67	0.38	0.48	0.60	0.29	0.40	0.55	0.29	0.38
Numeric/Temporal	404	0.45	0.01	0.02	0.82	0.24	0.37	0.14	0.03	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Common Form	2,452	0.21	0.00	0.01	0.66	0.33	0.44	0.49	0.28	0.35	0.88	0.04	0.08	0.43	0.00	0.00	0.56	0.00	0.01
Pro-form	153	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Singular Noun	2,623	0.79	0.17	0.28	0.73	0.45	0.56	0.62	0.38	0.47	0.87	0.24	0.38	0.79	0.20	0.32	0.74	0.19	0.31
Plural Noun	746	0.33	0.01	0.02	0.61	0.33	0.43	0.56	0.28	0.37	0.83	0.03	0.06	0.70	0.03	0.07	0.66	0.04	0.07
Adjective	516	0.77	0.02	0.04	0.26	0.07	0.11	0.56	0.24	0.34	0.65	0.14	0.23	0.72	0.21	0.32	0.60	0.14	0.22
Verb	334	0.00	0.00	0.00	0.86	0.02	0.04	0.37	0.17	0.23	1.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Adverb	12	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.42	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Non-Overlapping	2,871	0.75	0.12	0.20	0.67	0.33	0.45	0.58	0.38	0.46	0.84	0.19	0.32	0.78	0.19	0.30	0.71	0.17	0.27
Maximal Overlap	464	0.87	0.17	0.29	0.85	0.36	0.50	0.73	0.34	0.46	0.89	0.19	0.32	0.84	0.08	0.15	0.84	0.12	0.22
Intermediate Overlap	71	0.76	0.18	0.30	0.71	0.52	0.60	0.57	0.30	0.39	0.56	0.13	0.21	0.54	0.10	0.17	0.78	0.10	0.17
Minimal Overlap	825	0.82	0.04	0.09	0.61	0.37	0.46	0.50	0.15	0.23	0.80	0.09	0.17	0.72	0.09	0.16	0.66	0.06	0.12
Direct	3,106	0.79	0.13	0.23	0.71	0.43	0.53	0.63	0.38	0.47	0.83	0.21	0.33	0.76	0.19	0.30	0.70	0.17	0.27
Anaphoric	153	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Metaphoric	69	0.00	0.00	0.00	0.57	0.29	0.38	0.43	0.35	0.38	0.91	0.14	0.25	0.00	0.00	0.00	0.00	0.00	0.00
Metonymic	73	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Related	829	0.64	0.06	0.11	0.36	0.14	0.20	0.39	0.24	0.30	0.76	0.10	0.18	0.81	0.08	0.15	0.83	0.09	0.16
Descriptive	189	0.33	0.01	0.01	0.44	0.02	0.04	0.16	0.02	0.03	0.60	0.02	0.03	0.00	0.00	0.00	0.6	0.02	0.03
All	4231	0.77	0.11	0.19	0.67	0.35	0.46	0.59	0.33	0.42	0.84	0.17	0.29	0.77	0.16	0.26	0.72	0.14	0.24

We now apply our three fine-grained datasets to evaluate the performance of five EL systems with APIs available online, namely: Babelfy (**B**), TagME (**T**), DBpedia Spotlight (**D**), AIDA (**A**) and FREDME (**F**). All of these systems are applied to the texts with their default online configurations (and set for English). In the case of Babelfy, it provides two high-level options: *strict*, which focuses on named entities (B_s); and *relaxed*, which also includes common entities (B_r); we decide as an exception in this case to evaluate both versions of Babelfy.

We then compute the micro Precision (**P**), micro Recall (**R**) and micro F_1 score (**F**₁) for these systems; in other words, we compute precision, recall and F_1 over a dataset composed of the concatenation of our three datasets. Following the precedent of GERBIL [164], we consider false positives to be annotations that overlap with a dataset annotation but with a different link. True positives must have the same link and mention boundaries as labeled in the dataset; although systems sometimes propose annotations with the same target KB entity but a different overlapping boundary, such cases represented 0.013% of the total annotations identified, where on manual review, most of these cases were mentions based on partial names, such as linking “Merkel” instead of her full name “Angela Merkel”.

We recall that mentions may be associated with multiple link options while current EL systems suggest one link per mention. In the case of *OR* links, we consider a system annotation to be a true positive if it matches any of the alternatives, removing the other alternatives from consideration (i.e., they are not considered as false negatives); in the case

of *AND* links, we compute a local precision and recall measure for that mention, averaging the scores for all mentions in the combined datasets.²⁴

Table 6.3 then presents the results, broken down by annotations of each individual category, further indicating the number of mentions labeled with that category ($|A|$); the last row provides the overall results considering all mentions.²⁵ Given the large number of results, we shade better results (closer to one) with a darker color to aid visual comparison. From these results, we observe the following high-level trends:

- In terms of categories well-supported by the evaluated systems, in the BASE FORM dimension, we see that the best results are given for *Proper Forms* (named entities), with *Full* and *Extended Mentions*, in particular, having good results; results were poorer in the case of *Aliases* and *Short Mentions*. In the PART OF SPEECH dimension, results were best for *Nouns* and *Adjectives* (note that many adjectives, like “Russian”, are based on proper forms). In the OVERLAP category, we do not see any notable trends across the different categories, which was perhaps unexpected; we remark, however, that a system not allowing inner overlapping mentions may still find annotations labeled as *Minimal Overlap* assuming it does not recognize the outer mention, and hence the results do not necessarily reflect system policies regarding such mentions. Finally, in the REFERENCE dimension, we see that *Direct* and *Related* links have the broadest support, though recall is often low.
- Conversely, looking at categories of annotations with negligible support, in the BASE FORM dimension we found that *Pro-form* mentions have negligible support in all systems, while in the REFERENCE category, we found that *Anaphoric* and *Metonymic* links also have negligible support. Other categories, such as *Descriptive* links in the REFERENCE category, have uniformly poor support across the systems.
- On the other hand, some categories received mixed support across the evaluated systems. In particular, in the BASE FORM category, we see mixed results for *Common Form* annotations, where Babelfy_r and TagME find a considerable number of such mentions, whereas other systems find few or none. Likewise, in the PART OF SPEECH dimension, we see a further distinction, where TagME captures more verbs and adverbs than even Babelfy_r, indicating that the latter system, while permitting common entities, perhaps limits the detection of entity mentions to noun phrases. We see these particular variations across systems as revealing the different design choices made for those EL systems.

²⁴The *AND* case only came into play for extended versions of EL systems since all such cases came from *Pro-form* or *Descriptive* annotations not considered by off-the-shelf systems. We do not have combinations of *OR* and *AND*; in such a case, we suggest that the maximum score for all *OR* alternatives be taken as the score for that mention.

²⁵Counts are given by mention; for this reason, the sum of $|A|$ for categories in the dimension REFERENCE is greater than the total amount as one mention may have, for example, a separate *Related* and *Metonymic* link.

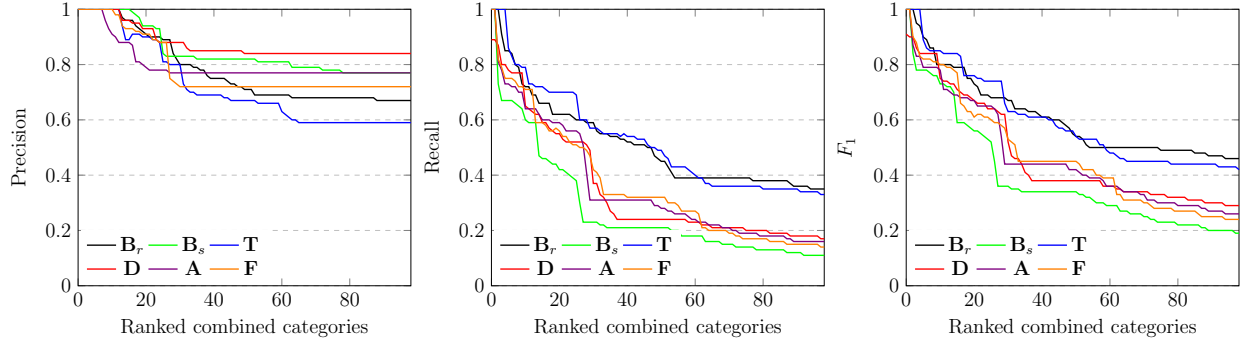


Figure 6.7: Cumulative best-first progression of precision, recall and F_1 scores for Babelfy (relaxed/strict), TagME, DBpedia Spotlight, AIDA and FREME for the unified dataset considering combinations of categories [138]

It is also interesting to contrast some of these results with those of the questionnaire. For example, while systems do not support *Metonymic* references, the results of Table 4.1 indicate that such references were preferred by respondents in the community when compared with the entity directly named (e.g., linking “Moscow” in the given sentence to `wiki:Government_of_Russia` rather than `wiki:Moscow`).

While Table 6.3 provides detailed results per individual categories, each annotation is labeled with four categories – one from each dimension – resulting in $7 \times 5 \times 4 \times 6 = 840$ combinations of categories applicable to an annotation across the four dimensions. However, not all 840 combinations do (or can) occur, where, for example, a *Pro-form* mention is always labeled as an *Anaphoric* reference. We found 123 combinations of these categories to have at least one annotation in the unified dataset. Rather than present the results for all such combinations across the systems, in Figure 6.7, we rather present a best-first cumulative progression of performance across the combinations, presenting Precision, Recall and F_1 as separate charts. At $x = 1$, we select the combination with the best score for the current metric and system, presenting the score for that metric; at $x = 2$, we add the annotations of the second-best combination to the current set of annotations and present the resulting score; and so forth. Although precision remains *relatively* high as combinations are added – i.e., the majority of annotations given by systems tend to remain correct – recall drops drastically as combinations not well-covered by the systems are added; this is likewise reflected in the F_1 scores. In these results, we can distinguish two groups of systems: Babelfy_r and TagME have lower precision towards the end of the progression, but maintain a much higher recall; on the other hand, Babelfy_s, DBpedia Spotlight, AIDA and FREME maintain higher precision throughout the progression, but lose recall much more rapidly than the first group. Again, we see this division as revealing different design issues in the two groups of systems, particularly relating to the inclusion/exclusion of common entities.

6.7 Fuzzy Recall and F_1 Measures

Thus far we have presented the results of the EL systems on a category-by-category basis, providing insights into the performance of EL systems for fine-grained categories of annotations. However, these results may perhaps be considered *too* fine-grained, making it somewhat difficult to compare systems at a glance. On the other hand, we mentioned that some categories of annotations appear to belong to the “core” definition of EL, while other categories are only considered by some authors; furthermore, we mentioned that some EL annotations might be more important in certain application scenarios than others. These observations lead us to propose a framework in the following that assigns different weights to different annotations, which may denote the level of consensus that annotation should be the target of the EL task, or the importance of that annotation to a particular application scenario, and so forth. Thereafter we instantiate this framework with a concrete measure and use it to evaluate the EL systems.

6.7.1 Fuzzy Framework

We propose a configurable evaluation framework based on *Fuzzy Set Theory* [175] for weighting annotations during the evaluation of EL systems. More specifically, given a universe of elements U , a *fuzzy set* A^* is associated with a *membership function* $\mu_{A^*} : U \rightarrow [0, 1]$ which denotes the *degree* to which a member of the universe $x \in U$ is a member of A^* ; we denote this degree by $\mu_{A^*}(x)$. Noting that a traditional *crisp set* B can be defined with a membership function $\mu_{A^*} : U \rightarrow \{0, 1\}$ – mapping elements of the universe to a value 0 or 1 instead of a value between 0 and 1 – fuzzy sets thus generalize crisp sets. We can consider the gold standard as providing a fuzzy set of annotations, where the degree of the annotation may intuitively denote the importance, consensus, etc., for that annotation in the given setting; more concretely, we propose metrics that penalize systems more for missing annotations with higher degree. This framework helps to answer **RQ1c**.

To define such measures, we first define an annotation as a triple $a = (o, o', l)$, where o and o' denotes the start and end offset of the mention in the input text ($o < o'$), and l denotes a link represented by a KB identifier or a special not-in-lexicon (NIL) value. We then consider a (crisp) gold standard G to be a set of annotations, and the results of an EL system to be a set of annotations. For a given gold standard G and system result S , the set of *true positives* is defined as $TP = G \cap S$, *false positives* as $FP = S - G$, and *false negatives* as $FN = G - S$. In the fuzzy setting, we still consider S to be a crisp set; however, we allow the gold standard G^* to be a fuzzy set, with $\mu_{G^*} : G \rightarrow [0, 1]$; slightly abusing notation, for annotations $a \notin G$, we assume $\mu_{G^*}(a) = 0$. We will later discuss how this membership function can be defined in practice for a given gold standard, but first we will discuss how Precision, Recall and F_1 measures are defined with respect to the fuzzy gold standard G^* .

For a given system result S , gold standard G and its fuzzy version G^* , we define the *fuzzy recall measure* R^* with respect to G^* as $R^* = \frac{\sum_{a \in S} \mu_{G^*}(a)}{\sum_{a \in G} \mu_{G^*}(a)}$, thus applying different costs for

missing annotations (type II errors) depending on the annotation in question. On the other hand, we propose that precision be computed in the traditional way for the crisp version of the gold standard – $P = \frac{|TP|}{|S|}$ – with the intuition that false positives proposed by the system (type I error) be weighted equally: if the system proposes an annotation, it should be correct, independently of the type of annotation.²⁶ We then define the fuzzy F_1 measure as simply the harmonic mean of the fuzzy recall measure and the traditional precision measure: $F_1^* = \frac{2 \cdot P \cdot R^*}{P + R^*}$.

The following properties are now verified for R^* and F_1^* :

- **PROP1:** the values for R^* and F_1^* both range between 0 and 1, inclusive.

Proof: The lower bound is given when no annotation of the system is in the gold standard, and thus, $\mu_{G^*}(a) = 0$ for all $a \in S$. On the other hand, the upper bound is given when $S = G$, and thus $R^* = \frac{\sum_{a \in S} \mu_{G^*}(a)}{\sum_{a \in G} \mu_{G^*}(a)} = \frac{\sum_{a \in G} \mu_{G^*}(a)}{\sum_{a \in G} \mu_{G^*}(a)} = 1$. Otherwise, observe that the numerator and denominator of R^* remain positive because they are the sum of membership degrees that are positive by definition. Furthermore, the numerator’s sum only includes non-zero summands for annotations of the system that are contained in G , and therefore the numerator is always lower than the denominator, and thus we conclude that R^* ranges between 0 and 1, inclusive. Given that both R^* and P (the traditional precision measure) range between 0 and 1 inclusive, so too does F_1^* : the harmonic mean of both measures. \square

- **PROP2:** when $\mu_{G^*} : G \rightarrow \{1\}$ (i.e., when memberships are binary), the fuzzy measures R^* and F_1^* correspond to the traditional measures R and F_1 .

Proof: When memberships are binary, $\mu_{G^*}(a) = 1$ for all $a \in S \cap G$ and $\mu_{G^*}(a) = 0$ for all $a \in S - G$, respectively. In this context, $R^* = \frac{\sum_{a \in S} \mu_{G^*}(a)}{\sum_{a \in G} \mu_{G^*}(a)} = \frac{|S \cap G|}{|G|} = \frac{|TP|}{|G|}$ per the traditional recall measure R , and as a consequence, F_1^* behaves the same as the traditional F_1 measure. \square

- **PROP3:** for a given system result, missing annotations with higher membership degree are penalized more in R^* and F_1^* than those with lower membership degree.

Proof: Given a fuzzy gold standard G^* , let a_1 and a_2 be two annotations such that $\mu_{G^*}(a_1) < \mu_{G^*}(a_2)$. Further let S be a set of system annotations that includes both a_1 and a_2 . In order to prove the result for R^* , we must prove the following inequality,

²⁶Furthermore observe that if we were to hypothetically define a fuzzy precision measure in the natural way, for the weighted denominator, we would end up having to assign weights to false positive annotations in $S - G$, which will not be available; an option would be to assign weights of 1 to such false positives, but this is not so natural since correct annotations may be assigned lower weights. In summary, defining a fuzzy precision measure would require a “fudge” where we thus prefer the traditional precision measure as discussed.

where the left-hand side represents the R^* measure for S removing a_2 , while the right-hand side represents the R^* measure for S removing a_1 :

$$\frac{\sum_{a \in S} \mu_{G^*}(a) - \mu_{G^*}(a_2)}{\sum_{a \in G} \mu_{G^*}(a)} < \frac{\sum_{a \in S} \mu_{G^*}(a) - \mu_{G^*}(a_1)}{\sum_{a \in G} \mu_{G^*}(a)} \quad (6.1)$$

We can simplify this inequality as follows:

$$\sum_{a \in S} \mu_{G^*}(a) - \mu_{G^*}(a_2) < \sum_{a \in S} \mu_{G^*}(a) - \mu_{G^*}(a_1) \quad (6.2)$$

$$-\mu_{G^*}(a_2) < -\mu_{G^*}(a_1) \quad (6.3)$$

$$\mu_{G^*}(a_1) < \mu_{G^*}(a_2) \quad (6.4)$$

Hence we see that inequality (6.1) holds if and only if the assumed inequality (6.4) holds, proving the result for R^* .

For precision, there are two possibilities such that $\mu_{G^*}(a_1) < \mu_{G^*}(a_2)$: either $a_1 \in G$ or $a_1 \notin G$ (in both cases $a_2 \in G$). In the case that $a_1 \in G$, then P is affected equally by the omission of either a_1 or a_2 . In the case that $a_1 \notin G$, then P is less in the case that a_2 is omitted than in the case that a_1 is omitted. Since P missing a_2 is less than or equals P missing a_1 , and R^* missing a_2 is strictly less than R^* missing a_1 , we conclude that F_1^* missing a_2 is strictly less than F_1^* missing a_1 , proving the result for F^* . \square

This same behavior cannot be achieved by extending Precision in the same way as Recall. Precision operates over FP , but annotations that belong to FP are not covered in the benchmark dataset, and thus, are not included in the domain of μ_{G^*} . On the other hand, just setting all membership degrees to zero for FP annotations makes Precision yield a constant value equal to one. Thus we keep the traditional behavior of Precision rather than defining a fuzzy variant of the measure.

Having defined the fuzzy framework in an abstract way and proven some natural properties that it satisfies, we are left to discuss how the values for the membership function μ_{G^*} can be defined in practice. In fact, we argue that the definition of μ_{G^*} is dependent on the setting, and may be manually configured based on categories, automatically learned from labeled examples in a given setting, and so forth. In the following, we propose a straightforward instantiation of this membership function and use it to evaluate the selected EL systems.

6.7.2 Fuzzy Evaluation

We propose to generate a membership function for the annotations of our datasets based on the questionnaire results seen in Figure 4.1 and our categorization scheme. Specifically, we select combined categories that consistently score greater than 0.9 in Figure 4.1 and assign them a degree of 1, considering them to be *strict annotations*; as a result, the strict annotations are those labeled as *Proper Form*, *Noun*, *No Overlap* with *Direct* reference. We

call all other annotations *relaxed* and assign them a membership degree of α . By varying the value of α , we can then place more importance in the evaluation results on achieving a greater ratio of relaxed annotations; more specifically, when $\alpha = 0$, missing a relaxed annotation does not affect R^* , but when $\alpha = 1$, missing a relaxed annotation affects R^* the same as missing a strict annotation. Given that the gold standard may offer multiple alternative links for a mention, we apply the same procedure discussed previously for the traditional measures. In the case of *OR* annotations, we check for each mention that the predicted link matches one of the alternatives in the gold standard where in the case of R^* , the membership degree for a mention in G^* is given as the maximum membership score over all annotations/links for that mention in G^* ; e.g., if a system predicts a link for a mention with weight α in G^* but there exists another link for that mention with weight 1 in G^* , the system will score $\frac{\alpha}{\max\{1,\alpha\}} = \alpha$ for that mention in R^* . On the other hand, in the case of *AND* annotations, we compute a local R^* value for that mention, thereafter averaging the R^* values for all mentions (i.e., we apply macro- R^* on different mentions).

The F_1^* results are shown in Figure 6.8 for the off-the-shelf EL systems and for varying degrees of α .²⁷ Here we see that all systems perform worse as more emphasis is given to relaxed annotations. We can further see two different behaviors in the systems: when less emphasis is placed on relaxed annotations, the four system configurations not linking common entities perform better, but as more emphasis is placed on relaxed annotations, the two system configurations that do link common entities perform better relative to the other system configurations.

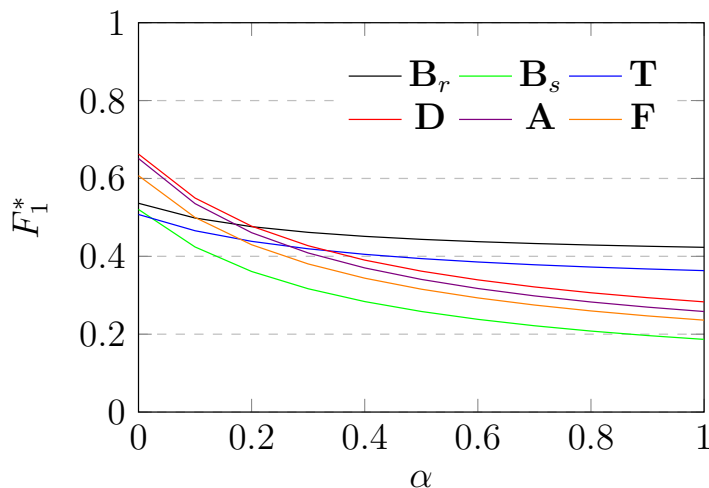


Figure 6.8: α -based fuzzy F_1 scores for off-the-shelf systems

In summary, the fine-grained classification proposed in this chapter helps to understand

²⁷We do not show results for P as they do not change for varying α , and with P being constant, R^* follows the same trend as F_1^* . Also the results for the extended FEL systems look largely identical, being slightly flatter.

the performance of different EL systems for different types of entities, which different applications may prioritize in different ways. The fuzzy measure we have defined here allows for compressing the information across different categories, based on application-specific weighting for the different categories. The result is a higher-level measure of the performance of different EL systems in the context of specific applications.

Chapter 7

Fine-Grained Entity Linking Systems

Our categorization scheme considers a number of types of mentions and links that – although indicated as annotations that EL systems should ideally give by some respondents in the questionnaire – are not supported by the evaluated EL systems. As previously discussed, this may be due to design choices made for particular systems; for example, in the case of *Pro-form* mentions – not supported by any evaluated system – one may argue that this part of a separate Coreference Resolution (CR) task [157]; on the other hand, though Babelfy_r and TagME support *Common Form* annotations, one may likewise argue that this is part of a separate Word Sense Disambiguation (WSD) task [116]. Conversely, some systems choose to incorporate CR [64, 41] and WSD [110] methods for the EL task.

In this Chapter, we extend five EL systems with CR and WSD methods to create prototypes of what we call Fine-Grained Entity Linking (FEL) systems and evaluate them on our datasets to understand how far state-of-the-art methods can reach considering our more inclusive, fine-grained view of the potential goals of EL. We expect these extended systems to exhibit increased recall on our datasets, particularly for *Pro-form* annotations (all cases) and *Common Form* annotations (particular for AIDA, Babelfy_s, DBpedia Spotlight and FREME).

7.1 Adding Coreference Resolution

We first extend the existing EL systems with techniques for CR. In particular, we employ two off-the-shelf tools provided by Stanford CoreNLP [91] for these purposes. Both of these models provide scores indicating the likelihood of a particular mention having a particular antecedent in the text.

SCR: Refers to the statistical coreference resolution model [31] trained on the CoNLL 2012 data, which uses logistic classification and ranking, with features based on the distance between coreferent mentions, syntax (e.g., POS tags, mention length), semantics (e.g., the type of entity), rules (matching known patterns), and lexical elements (e.g., the head term of a mention).

Table 7.1: Changing results per category for Babelfy, TagME, DBpedia Spotlight, AIDA and FRED extended with *SCR* on the unified dataset.

	A	B _s			B _r			T			D			A			F		
		P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
Pro-form	153	0.44	0.18	0.26	0.55	0.24	0.33	0.52	0.27	0.36	0.60	0.24	0.34	0.55	0.27	0.36	0.44	0.22	0.30
Singular Noun	2623	0.77	0.18	0.29	0.72	0.45	0.56	0.62	0.39	0.48	0.86	0.25	0.39	0.78	0.21	0.33	0.73	0.20	0.31
Adjective	518	0.49	0.05	0.09	0.33	0.12	0.17	0.55	0.29	0.38	0.64	0.18	0.28	0.67	0.25	0.37	0.54	0.18	0.27
Non-Overlapping	2871	0.71	0.13	0.21	0.67	0.34	0.45	0.58	0.40	0.47	0.82	0.20	0.33	0.76	0.20	0.32	0.68	0.18	0.28
Minimal Overlap	826	0.78	0.05	0.09	0.62	0.38	0.47	0.51	0.15	0.24	0.81	0.10	0.18	0.72	0.10	0.17	0.66	0.07	0.13
Anaphoric	153	0.44	0.18	0.26	0.55	0.24	0.33	0.52	0.27	0.36	0.60	0.24	0.34	0.55	0.27	0.36	0.44	0.22	0.30
All	4231	0.74	0.12	0.20	0.67	0.36	0.46	0.59	0.34	0.43	0.82	0.18	0.30	0.76	0.17	0.27	0.69	0.15	0.25

NCR: Refers to the neural coreference resolution model [30], which uses reinforcement learning on word embeddings and features, with hidden layers based on rectified linear units (ReLU) and a fully-connected scoring layer.

Using SCR and NCR, we can then extract antecedents for a mention. Subsequently taking the results of a given EL tool, if a particular mention is not annotated with a link, but the CR tool identifies an antecedent for that mention and the EL tool annotates the antecedent with a link, we can propose that link for the original mention. For example, in the text “Michael Jackson is a pop singer. He was managed by Joe Jackson.”, assuming that the EL system links “Michael Jackson” to `wiki:Michael_Jackson` but does not annotate “He”, and assuming that the CR tool states that “Michael Jackson” is the antecedent for “He”, then we will extend the results of the EL system by linking “He” to `wiki:Michael_Jackson`.

We provide the results extended with SCR in Table 7.1 and the results extended with NCR in Table 7.2 where we display only those categories (rows) where results changed versus the off-the-shelf results from Table 6.3; this time we shade cells blue in case of improvement or red in case of deterioration of results, with more intense shading indicating greater change. As expected, we see an improvement in the results for *Pro-form* and *Anaphoric* categories using both CR techniques. In both cases, we also see some deterioration in the precision of adjectives, which we attribute to the CR extensions having lower precision for pro-form adjectives such as “her” than the baseline ER systems have for proper-form adjectives such as “Russian”; the recall and F_1 indeed improves slightly for this category. Comparing SCR with NCR, we see the neural variant affecting more categories, including changes in the *Descriptive* category; we attribute this to the Deep Learning architecture of NCR being able to detect coreference for more complex forms of mentions than the logistic framework employed for SCR.

7.2 Adding Word Sense Disambiguation

Word Sense Disambiguation (WSD) refers to the task of disambiguating the sense of a word used in a particular context [116]. A typical target for WSD is to link words with

Table 7.2: Changing results per category for Babelfy, TagME, DBpedia Spotlight, AIDA and FREDME extended with *NCR* on the unified dataset.

	A	B _s			B _r			T			D			A			F		
		P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
Short Mention	497	0.46	0.17	0.25	0.37	0.25	0.30	0.54	0.44	0.49	0.50	0.31	0.38	0.50	0.37	0.42	0.39	0.29	0.34
Common Form	2452	0.21	0.00	0.01	0.66	0.34	0.44	0.49	0.28	0.36	0.88	0.04	0.08	0.69	0.00	0.01	0.67	0.01	0.01
Pro-form	153	0.39	0.15	0.22	0.39	0.16	0.23	0.47	0.22	0.30	0.55	0.20	0.30	0.51	0.23	0.32	0.42	0.18	0.25
Singular Noun	2623	0.77	0.18	0.29	0.72	0.46	0.56	0.62	0.39	0.48	0.86	0.25	0.39	0.78	0.21	0.33	0.73	0.20	0.32
Adjective	518	0.48	0.04	0.07	0.29	0.10	0.15	0.55	0.28	0.37	0.63	0.17	0.27	0.66	0.24	0.36	0.55	0.17	0.26
Non-Overlapping	2871	0.72	0.13	0.22	0.66	0.34	0.45	0.58	0.40	0.47	0.83	0.21	0.33	0.76	0.20	0.32	0.69	0.18	0.29
Maximal Overlap	464	0.83	0.17	0.29	0.83	0.36	0.50	0.73	0.35	0.47	0.90	0.20	0.33	0.86	0.09	0.17	0.85	0.14	0.23
Intermediate Overlap	71	0.78	0.20	0.31	0.72	0.54	0.61	0.57	0.30	0.39	0.59	0.14	0.23	0.57	0.11	0.19	0.80	0.11	0.20
Minimal Overlap	826	0.72	0.05	0.09	0.60	0.38	0.46	0.50	0.15	0.24	0.77	0.10	0.18	0.68	0.09	0.16	0.63	0.07	0.12
Anaphoric	153	0.39	0.15	0.22	0.39	0.16	0.23	0.47	0.22	0.30	0.55	0.20	0.30	0.51	0.23	0.32	0.42	0.18	0.25
Descriptive	189	0.25	0.01	0.02	0.40	0.03	0.06	0.32	0.04	0.07	0.82	0.05	0.09	1.00	0.03	0.06	0.82	0.05	0.09
All	4231	0.73	0.12	0.20	0.67	0.35	0.46	0.59	0.34	0.43	0.82	0.18	0.30	0.76	0.17	0.27	0.70	0.15	0.25

WordNet [104], which provides groups of words representing synonyms (aka. *synsets*) in English, relations between synsets, as well as definitions of words. Words with multiple senses (meanings) can be found in different synsets: one for each sense of the word. Tools performing WSD can then link a word with a particular synset in a database like WordNet, thus disambiguating the sense of the word used in the text. Of the EL systems evaluated in Chapter 6, only Babelfy includes a WSD component in its relaxed configuration [110]. To extend all of the evaluated EL systems, we use the following WSD tools:

WSD-NLTK: We use the WSD system packaged with the Natural Language Toolkit (NLTK) based on the Lesk algorithm [87], which ranks the senses of a word in a text based on how many neighboring words in the text also appear in the dictionary definition of the word sense. The WSD-NLTK tool then links words to WordNet synsets.

WSD-DIS: Refers to the “*disambiguate*” system proposed by Vial et al. [166], which aggregates word senses in Wordnet into higher-level clusters of sense based on the semantic relations it contains. These are then used in the context of a neural WSD system combined with a pre-trained BERT model, achieving state-of-the-art results.

Given that our goal is to link to Wikipedia and not WordNet, and that neither WordNet nor Wikipedia link to each other, we use the third-party alignment provided by Miller and Gurevych [106] to map from the WordNet-based WSD results to Wikipedia articles. Thereafter, given the results of an EL system, any word that can be linked to Wikipedia through the WSD tools and that is not already a mention returned by the EL system is added (with the corresponding link) to the results.

The results of the EL systems extended with WSD-NLTK are shown in Table 7.3, while the results with WSD-DIS are shown in Table 7.4; as before, we only include categories whose results change. Across both systems, we see that a broader range of categories are affected versus the extensions with CR; however, annotations with the category *Adverb*

Table 7.3: Changing results per category for Babelfy, TagME, DBpedia Spotlight, AIDA and FREDME extended with *WSD-NLTK* on the unified dataset.

	A	\mathbf{B}_s			\mathbf{B}_r			\mathbf{T}			\mathbf{D}			\mathbf{A}			\mathbf{F}		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Full Mention	766	0.73	0.51	0.60	0.69	0.57	0.62	0.75	0.62	0.68	0.77	0.60	0.67	0.72	0.58	0.65	0.74	0.57	0.64
Short Mention	497	0.31	0.21	0.25	0.35	0.29	0.31	0.53	0.46	0.49	0.43	0.31	0.36	0.46	0.38	0.42	0.36	0.30	0.33
Alias	112	0.39	0.18	0.25	0.32	0.22	0.26	0.49	0.32	0.39	0.63	0.39	0.48	0.57	0.29	0.39	0.54	0.29	0.37
Numeric/Temporal	404	0.56	0.18	0.28	0.68	0.25	0.37	0.34	0.14	0.19	0.59	0.18	0.28	0.59	0.18	0.28	0.59	0.18	0.28
Common Form	2452	0.25	0.13	0.17	0.56	0.36	0.44	0.43	0.32	0.37	0.30	0.16	0.21	0.25	0.12	0.16	0.25	0.13	0.17
Singular Noun	2623	0.46	0.29	0.36	0.64	0.48	0.55	0.56	0.44	0.49	0.53	0.35	0.42	0.48	0.31	0.38	0.47	0.30	0.37
Plural Noun	746	0.25	0.14	0.17	0.51	0.34	0.41	0.45	0.31	0.37	0.29	0.16	0.20	0.28	0.16	0.20	0.28	0.16	0.20
Adjective	518	0.21	0.06	0.09	0.24	0.11	0.15	0.43	0.26	0.33	0.41	0.16	0.23	0.51	0.24	0.32	0.41	0.16	0.23
Verb	334	0.00	0.00	0.00	0.46	0.02	0.03	0.35	0.17	0.23	0.10	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Non-Overlapping	2871	0.43	0.24	0.31	0.56	0.36	0.44	0.54	0.42	0.47	0.51	0.30	0.38	0.49	0.30	0.37	0.46	0.28	0.35
Minimal Overlap	826	0.23	0.15	0.18	0.54	0.41	0.46	0.33	0.23	0.28	0.28	0.19	0.23	0.28	0.19	0.23	0.25	0.17	0.20
Direct	3106	0.45	0.27	0.33	0.64	0.46	0.54	0.55	0.43	0.49	0.52	0.32	0.40	0.49	0.31	0.38	0.46	0.29	0.36
Metaphoric	69	0.14	0.07	0.10	0.42	0.30	0.35	0.40	0.36	0.38	0.42	0.22	0.29	0.16	0.07	0.10	0.16	0.07	0.10
Related	829	0.18	0.08	0.11	0.27	0.16	0.20	0.34	0.25	0.29	0.23	0.11	0.15	0.21	0.10	0.13	0.22	0.10	0.14
Descriptive	189	0.25	0.01	0.01	0.44	0.02	0.04	0.15	0.02	0.03	0.50	0.02	0.03	0.00	0.00	0.00	0.50	0.02	0.03
All	4231	0.40	0.21	0.28	0.58	0.37	0.45	0.52	0.37	0.43	0.48	0.26	0.34	0.45	0.25	0.32	0.43	0.24	0.30

are not affected, probably because there are only 12 such annotations; further annotations with *Maximal Overlap* and *Intermediate Overlap* are not affected as they require more than one word, whereas WSD targets individual words; finally annotations in *Anaphoric* and *Metonymic* categories are not affected as WSD does not provide any mechanism for resolving complex references of this form. Both WSD systems improve F_1 measures overall by boosting recall at the cost of precision; less improvement is seen for EL systems that already support common entities (\mathbf{B}_r and \mathbf{T})s, where \mathbf{B}_r already incorporates WSD techniques [110]. Between both systems, WSD-NLTK tends to improve recall more than WSD-DIS, but WSD-DIS tends to maintain a higher precision.

Overall we see that precision in general worsens but recall improves across the different systems and categories, suggesting that WSD does allow for finding additional annotations but with lower precision than what the baseline EL systems find; overall, F_1 measures with WSD tend to improve slightly.

7.3 Combined CR and WSD Results

Finally we present the results of the EL systems combined with both CR techniques and both WSD techniques. The results are shown in Table 7.5, where this time we present all categories to also emphasize those that were not affected. In particular, we see that although more annotations are found in many categories, the extended systems still fail to support *Metonymic* references in particular. Given that the extensions are monotonic – annotations are added to the baseline systems – the recall increases for some categories; conversely, with some exceptions, precision tends to decrease, with CR and WSD targeting more difficult cases not addressed by the baseline EL systems.

Table 7.4: Changing results per category for Babelfy, TagME, DBpedia Spotlight, AIDA and FRED extended with *WSD-DIS* on the unified dataset.

	A	\mathbf{B}_s			\mathbf{B}_r			\mathbf{T}			\mathbf{D}			\mathbf{A}			\mathbf{F}		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Full Mention	766	0.80	0.48	0.60	0.71	0.54	0.62	0.77	0.60	0.68	0.80	0.58	0.68	0.74	0.57	0.65	0.76	0.55	0.64
Short Mention	497	0.34	0.17	0.23	0.36	0.26	0.30	0.52	0.45	0.48	0.47	0.30	0.37	0.49	0.38	0.43	0.38	0.30	0.34
Extended Mention	9	0.83	0.56	0.67	0.83	0.56	0.67	0.80	0.44	0.57	0.80	0.44	0.57	1.00	0.44	0.62	0.80	0.44	0.57
Alias	112	0.42	0.16	0.23	0.33	0.21	0.25	0.50	0.32	0.39	0.65	0.38	0.47	0.58	0.29	0.39	0.53	0.29	0.37
Numeric/Temporal	404	0.58	0.06	0.12	0.81	0.24	0.37	0.25	0.07	0.11	0.64	0.06	0.10	0.64	0.06	0.10	0.64	0.06	0.10
Common Form	2452	0.21	0.04	0.07	0.62	0.34	0.44	0.46	0.29	0.35	0.34	0.08	0.13	0.21	0.04	0.07	0.22	0.04	0.07
Singular Noun	2623	0.55	0.21	0.30	0.69	0.46	0.55	0.58	0.40	0.47	0.66	0.28	0.39	0.59	0.23	0.33	0.56	0.23	0.32
Plural Noun	746	0.22	0.05	0.08	0.59	0.34	0.43	0.52	0.29	0.37	0.33	0.07	0.11	0.32	0.07	0.11	0.32	0.07	0.12
Adjective	518	0.31	0.04	0.07	0.29	0.09	0.14	0.53	0.26	0.35	0.56	0.15	0.24	0.65	0.22	0.33	0.52	0.15	0.23
Verb	334	0.25	0.00	0.01	0.75	0.02	0.04	0.37	0.17	0.23	0.40	0.01	0.01	0.25	0.00	0.01	0.25	0.00	0.01
Non-Overlapping	2871	0.51	0.15	0.24	0.63	0.34	0.44	0.57	0.40	0.47	0.64	0.23	0.34	0.60	0.22	0.32	0.55	0.20	0.30
Minimal Overlap	826	0.28	0.08	0.12	0.59	0.39	0.47	0.38	0.18	0.24	0.39	0.13	0.19	0.36	0.12	0.18	0.32	0.10	0.15
Direct	3106	0.53	0.18	0.26	0.68	0.44	0.53	0.59	0.40	0.47	0.63	0.24	0.35	0.58	0.23	0.33	0.54	0.21	0.30
Metaphoric	69	0.19	0.04	0.07	0.53	0.29	0.37	0.43	0.35	0.38	0.65	0.19	0.29	0.25	0.04	0.07	0.25	0.04	0.07
Related	829	0.27	0.06	0.10	0.30	0.14	0.20	0.37	0.24	0.29	0.38	0.10	0.16	0.35	0.09	0.14	0.37	0.09	0.15
Descriptive	189	0.25	0.01	0.01	0.44	0.02	0.04	0.15	0.02	0.03	0.50	0.02	0.03	0.00	0.00	0.00	0.50	0.02	0.03
All	4231	0.50	0.14	0.22	0.64	0.36	0.46	0.56	0.34	0.43	0.61	0.20	0.30	0.56	0.19	0.28	0.53	0.17	0.26

Table 7.6 provides a summary of overall results for the extensions. In terms of the F_1 measure, we see some improvements, except in the case of \mathbf{B}_r , whose F_1 measure remains the same. Overall we can conclude that extending EL systems with CR and WSD broadens the types of annotations that can be supported and increases recall, but at the cost of lower precision; however, *Metonymic* references remain unsupported.

Table 7.5: Results per category for Babelify, TagME, DBpedia Spotlight, AIDA and FRENDE extended with *SCR*, *NCR*, *WSD-NLTK* and *WSD-DIS* on the unified dataset.

	A	B _s			B _r			T			D			A			F		
		P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
Full Mention	766	0.70	0.51	0.59	0.67	0.57	0.62	0.73	0.62	0.67	0.75	0.60	0.67	0.70	0.58	0.64	0.72	0.57	0.63
Short Mention	497	0.30	0.21	0.24	0.35	0.29	0.31	0.52	0.46	0.49	0.43	0.32	0.36	0.45	0.39	0.42	0.35	0.31	0.33
Extended Mention	9	0.83	0.56	0.67	0.83	0.56	0.67	0.80	0.44	0.57	0.80	0.44	0.57	1.00	0.44	0.62	0.80	0.44	0.57
Alias	112	0.34	0.18	0.24	0.32	0.22	0.26	0.47	0.32	0.38	0.63	0.39	0.48	0.55	0.29	0.38	0.52	0.29	0.37
Numeric/Temporal	404	0.56	0.18	0.28	0.68	0.25	0.37	0.34	0.14	0.19	0.58	0.18	0.28	0.58	0.18	0.28	0.58	0.18	0.28
Common Form	2452	0.24	0.13	0.17	0.56	0.36	0.43	0.43	0.32	0.37	0.29	0.16	0.21	0.24	0.13	0.17	0.25	0.13	0.17
Pro-form	153	0.32	0.20	0.24	0.39	0.25	0.30	0.42	0.29	0.34	0.41	0.25	0.31	0.44	0.29	0.35	0.35	0.24	0.28
Singular Noun	2623	0.45	0.30	0.36	0.63	0.49	0.55	0.55	0.45	0.49	0.52	0.36	0.43	0.47	0.32	0.38	0.46	0.31	0.37
Plural Noun	746	0.24	0.14	0.17	0.51	0.34	0.41	0.45	0.31	0.37	0.28	0.16	0.20	0.27	0.16	0.20	0.27	0.16	0.20
Adjective	518	0.25	0.09	0.13	0.30	0.15	0.20	0.47	0.31	0.37	0.45	0.21	0.28	0.55	0.28	0.37	0.44	0.20	0.28
Verb	334	0.08	0.00	0.01	0.46	0.02	0.03	0.35	0.17	0.23	0.15	0.01	0.01	0.08	0.00	0.01	0.08	0.00	0.01
Adverb	12	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.42	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Non-Overlapping	2871	0.42	0.25	0.31	0.56	0.37	0.45	0.54	0.44	0.48	0.51	0.31	0.39	0.49	0.31	0.38	0.46	0.29	0.36
Maximal Overlap	464	0.83	0.17	0.29	0.83	0.36	0.50	0.73	0.35	0.47	0.90	0.20	0.33	0.86	0.09	0.17	0.85	0.14	0.23
Intermediate Overlap	71	0.78	0.20	0.31	0.72	0.54	0.61	0.57	0.30	0.39	0.59	0.14	0.23	0.57	0.11	0.19	0.80	0.11	0.20
Minimal Overlap	826	0.22	0.15	0.18	0.54	0.41	0.47	0.33	0.24	0.28	0.28	0.20	0.23	0.28	0.20	0.23	0.25	0.17	0.21
Direct	3106	0.44	0.27	0.33	0.64	0.46	0.53	0.55	0.43	0.48	0.51	0.33	0.40	0.48	0.31	0.38	0.45	0.29	0.36
Anaphoric	153	0.32	0.20	0.24	0.39	0.25	0.30	0.42	0.29	0.34	0.41	0.25	0.31	0.44	0.29	0.35	0.35	0.24	0.28
Metaphoric	69	0.13	0.07	0.09	0.42	0.30	0.35	0.40	0.36	0.38	0.39	0.22	0.28	0.14	0.07	0.10	0.14	0.07	0.10
Metonymic	73	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Related	829	0.17	0.08	0.11	0.26	0.16	0.20	0.34	0.25	0.29	0.22	0.11	0.15	0.20	0.10	0.13	0.22	0.10	0.14
Descriptive	189	0.22	0.01	0.02	0.40	0.03	0.06	0.31	0.04	0.07	0.75	0.05	0.09	0.86	0.03	0.06	0.75	0.05	0.09
All	4231	0.39	0.22	0.28	0.58	0.38	0.46	0.52	0.39	0.44	0.47	0.28	0.35	0.44	0.26	0.33	0.42	0.25	0.31

Table 7.6: High-level results comparing different EL systems and WSD/CR extensions.

	A	B _s			B _r			T			D			A			F		
		P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
EL	4231	0.77	0.11	0.19	0.67	0.35	0.46	0.59	0.33	0.42	0.84	0.17	0.29	0.77	0.16	0.26	0.72	0.14	0.24
EL + SCR	4231	0.74	0.12	0.20	0.67	0.36	0.46	0.59	0.34	0.43	0.82	0.18	0.30	0.76	0.17	0.27	0.69	0.15	0.25
EL + NCR	4231	0.73	0.12	0.20	0.67	0.35	0.46	0.59	0.34	0.43	0.82	0.18	0.30	0.76	0.17	0.27	0.70	0.15	0.25
EL + WSD-NLTK	4231	0.40	0.21	0.28	0.58	0.37	0.45	0.52	0.37	0.43	0.48	0.26	0.34	0.45	0.25	0.32	0.43	0.24	0.30
EL + WSD-DIS	4231	0.50	0.14	0.22	0.64	0.36	0.46	0.56	0.34	0.43	0.61	0.20	0.30	0.56	0.19	0.28	0.53	0.17	0.26
EL + All	4231	0.39	0.22	0.28	0.58	0.38	0.46	0.52	0.39	0.44	0.47	0.28	0.35	0.44	0.26	0.33	0.42	0.25	0.31

Chapter 8

Conclusions and Future Work

In this PhD thesis we highlight the evident disagreement about the definition of entities in the context of the Entity Linking task. We believe that the solution to this lack of consensus lies in the different applications of Entity Linking. While applications such as Relation Extraction require the annotation of the largest set of entities possible, other applications only require identifying entities under a MUC-like definition that involves only the recognition of persons, locations and organizations. However, current benchmark datasets are only focused on one definition, ignoring the diverse requirements for the task. In addition, many datasets and resources focus on English language texts, which again prevents EL systems from being used in applications involving other languages. These issues imply the need for further work to better characterize the goals of the EL tasks and to generalize the techniques found in the literature in order to be suitable for new applications.

With respect to the lack of consensus, we propose a fine-grained categorization that gathers current definitions in one formalization. Generally speaking, we created a fine-grained EL ecosystem composed of (1) a fine-grained categorization scheme, (2) a vocabulary that formalizes these categories, (3) a standalone web system to create benchmark datasets with these categories, (4) three benchmark datasets labelled according these categories, (5) a quality measure that computes fine-grained scores from categorized benchmark datasets, and (6) a system that combines EL, WSD and CR in order to detect and link a wider range of fine-grained entities.

On the other hand, we also focused on issues surrounding multilingual Entity Linking. Motivated by the lack of multilingual datasets, and in particular the lack of parallel datasets with the same entities in different languages, we propose the VoxEL dataset, containing news articles in English, Italian, Spanish, German, and French. Using this novel dataset, we conducted experiments in order to measure and compare the behavior of popular EL approaches over texts in different languages. We also explored the use of machine translation: due to the recent improvements achieved in machine translation, we ascertain the performance possible by translating the given text from non-supported languages to supported-languages.

In the next subsections we summary of our main contributions and results, a discussion

of limitations that could be addressed in future works, and our outlook on the EL task.

8.1 Contributions and results

- We designed a questionnaire to understand the varying perspectives on the goals of the EL task that exist within the EL research community. While there was a strong consensus that named entities should be linked and that overlapping mentions should be allowed, responses were mixed on the issue of including common entities, pro-form mentions, and descriptive mentions as part of the EL task. Respondents in general preferred linking to the KB entity to which the mention intends to refer rather than linking to the KB entity that the mention explicitly names; in particular, respondents preferred to resolve metonyms.
- While Entity Linking has traditionally focused on processing texts in English, in recent years there has been a growing trend towards developing techniques and systems that can support multiple languages. To support such research, in this paper we have described a new labelled dataset for multilingual EL, which we call VOXEL. The dataset contains 15 news articles in 5 different languages with 2 different criteria for labelling, resulting in a corpus of 150 manually-annotated news articles. In a Strict version of the dataset considering a core set of entities, we derive 204 annotated mentions in each language, while in a Relaxed version of the dataset considering a broader range of entities described by Wikipedia, we derive 674 annotated mentions in each language. The VOXEL dataset is distinguished by having a one-to-one correspondence of sentences – and annotated entities per sentence – between languages. The dataset (in NIF) is available online under a CC-BY 4.0 licence: <https://dx.doi.org/10.6084/m9.figshare.6539675>.
- We used the VOXEL dataset to conduct experiments comparing the performance of selected EL systems in a multilingual setting. We found that in general, Babelfy and DBpedia Spotlight performed the most consistently across languages. We also found that with the exception of Babelfy, EL systems performed best over English versions of the text. Next, we compared configuring the multilingual EL system for each non-English language versus applying a machine translation of the text to English and running the system in English; with the exception of Babelfy, we found that the machine translation approach outperformed configuring the system for a non-English language; even in the case of Babelfy, the translation sometimes performed better, while in others it remained competitive. This raises a key issue for research on multilingual EL: state-of-the-art machine translation is now reaching a point where we must ask if it is worth building dedicated multilingual EL systems, or if we should focus on EL for one language to which other languages can be machine translated.
- We proposed a fine-grained categorization of EL annotations, comprising of twenty-four categories along four dimensions. We propose a vocabulary for annotating EL

datasets with these categories, describe a tool to assist with the annotation process, and provide associated annotation guidelines. Relabeling three existing EL datasets accordingly, we find that the number of annotations increases greatly, particularly in the case of the ACE2004 dataset, with many common entities being added.

- Evaluating five off-the-shelf EL systems with respect to the relabeled datasets, we find good support for named entities being referred to through nouns or adjectives. On the other hand, we find little support for mentions using metonymic reference, or pro-forms. We also find a split between the systems in terms of common entities, with some systems considering such entities and others not.
- We describe fuzzy-recall and fuzzy- F_1 measures that allow for assigning different weights to different annotations, thus allowing to configure the evaluation results according to the priorities of a given setting, or according to a particular consensus. Dividing the annotations of our datasets into strict and relaxed annotations based on the results of our questionnaire, by varying the weight assigned to relaxed annotations, we observe how systems perform as more priority is assigned to such annotations; we find that systems targeting common entities start with lower F_1 scores as relaxed annotations are assigned low weights, but perform better than systems targeting only named entities as relaxed annotations are given higher priority.
- With the goal of achieving state-of-the-art results for our datasets in terms of Fine-Grained Entity Linking (FEL), we extend the EL systems with off-the-shelf Coreference Resolution tools and Word Sense Disambiguation tools in order to capture more annotations. As expected, these extensions improve the recall of the systems, particularly for pro-form and common-form mentions, but often at the cost of lower precision. The extended EL systems still do not capture metonymic references.

8.2 Limitations and Future Work

As an initial work on exploring and expanding the boundaries of the goals of the EL task towards more fine-grained annotations and evaluation, there are a number of limitations that could be addressed in future work.

- We mentioned in Chapter 1 that the notion of an entity may vary across different languages and cultures, which we have not directly addressed. Many concepts change their meaning depending on the particular culture. One example is the family of team sports that involve kicking a ball to score a goal, known as “Football” in Great Britain, but as “Soccer” in the United States. This is not an isolated case; in Ireland, for instance, the following words have particular meanings different to the rest of the world: coppers, wagon, locked, shift, massive, press, yoke, ride, notion, dose, gas, messages, and others. All of these cases need processes that consider cultural information in order to identify the appropriate corresponding KB-entity. Another challenging factor

stems from multilingual scenarios. Languages such as German or Dutch contain compound words that enrich and augment their lexicon [68]. For example, the 63-letter German word “Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz” refers to the “law for the delegation of monitoring beef labelling” [114] which may not be considered a coherent entity in English.

- Our questionnaire was targeted at researchers from the EL community, with the goal of understanding what consensus exists within that community on the goals of the EL task, asking which annotations an EL system would ideally return. We saw varying responses and perspectives, which may lean towards what EL systems have conventionally targeted, rather than what the goals of the EL task should be going forward. Regarding the latter question, it might be of interest to consider the perspectives of other sub-communities of computational linguistics, and also experts in areas that use EL tools in their work.
- While the VoxEL dataset that we propose covers a variety of languages, we only provide texts for a small selection of European languages, in particular because the source was from a European website. Also texts required curation to ensure a one-to-one correspondence of sentences and entities across the languages. It would be of interest to develop parallel datasets (with one-to-one correspondences across languages) for a much broader range of languages. One possibility would be to leverage crowdsourcing platforms. Another issue is that VoxEL only provides two categories of entities – Strict and Relaxed – where it was deemed too challenging to provide fine-grained labels without the expertise of native speakers. Maintaining parallel texts with one-to-one correspondences of fine-grained annotations would also be complicated; for example, the use of certain pronouns in Spanish is optional.
- Labeling EL datasets with fine-grained categories, as we propose, is far more challenging and costly than labeling datasets focused primarily on named entities: the number of annotations required increases roughly thirty-fold under the broader definition, mentions may link to multiple alternatives (e.g., under metonymy), each annotation must be labeled with specific categories, the guidelines to follow grow more complex, etc. Unlike named entities that are commonly capitalized (in many languages), another challenge relates to identifying the common-form words and phrases in the text that have corresponding KB entries. In order to assist in the annotation process, we developed the NIFify tool, which helps not only to generate, but also to semi-automatically validate, annotations. This tool could be extended to include further features, such as automatically suggesting annotations, perhaps based on similar mentions annotated previously. Another option to explore might be to use crowdsourcing, though given the challenging nature of the annotation process, designing human-intelligence tasks appropriate for non-experts is non-trivial; a viable approach might be to divide the annotation process into smaller tasks, for example, with one task for annotating named entities, another for common entities, another for resolving coreference, another for labeling categories, etc.

- At the outset of labeling our datasets, we did not have the categories and guidelines defined; rather we adopted a more agile methodology where the categories and guidelines were developed in parallel with – and adapted for – the labeling process itself, with decisions made based on a consensus between the authors. As such, we currently do not have an estimate for inter-rater agreement in terms of annotating datasets per our categories and guidelines. Based on our experience labeling our datasets, and relating to the previous point, we believe that such agreement would be a function of how well the annotators understand the guidelines and categories, and how much experience the annotators have with respect to what the KB includes/excludes. There is also some subjective judgment required for certain cases, such as in the case of “daily”, which may point to `wiki:Day` or `wiki:Newspaper`, or in the case of “nation”, where the options include `wiki:Nation`, `wiki:Nation_state`, `wiki:Country`, `wiki:State_(polity)`, etc., where the appropriate choice may be subjective and dependent on the context of the mention. With the categories and guidelines now defined, it would be interesting to design experiments to measure inter-rater agreement in order to better understand where differences occur between annotators.
- Our categorization scheme was designed to cover the cases we found in the three existing EL datasets that we relabeled. These EL datasets mainly pertain to news articles or extracts thereof, which tend to have a high density and diversity of named entities, making them suitable for traditional EL settings. Our categorization scheme may thus not cover the types of mentions that may occur in other settings, such as user-mentions or hashtags on Twitter. However, our categorization scheme is extensible, and could be expanded to cover other application scenarios in future.
- In order to ensure that our categorization scheme covered all the of the cases found in the three datasets, we extended the scheme with values such as *Extended Name*, *Adverb*, *Intermediate*, *Metaphoric*, etc., that occur in the texts, but do so infrequently (see Table 6.2). Rather than being a particular characteristic of our datasets, we believe that these types of annotations would occur relatively infrequently in general. For example, we find 9 instances of *Extended Name* (e.g., “Michael Joseph Jackson”) across our three datasets; such mentions are rare as even where they are used, they will typically appear at most once in a document to introduce an entity, with *Short Name* being used for subsequent references to that entity (“Jackson”, “Michael”, etc.). Likewise, we found 13 instances of *Adverb* in the datasets associated with Wikipedia articles; these were a small fraction of the *adverbs of form* (those that typically end with “-ly”), specifically those related to philosophical qualities or concepts (“simply” → `wiki:Simplicity`, “naturally” → `wiki:Nature`); or a handful of numeric values (“once” → `wiki:1`, “twice” → `wiki:2`). Still, the low number of examples for certain categories may be a limitation for training or evaluating systems focusing on particular (rare) types of entity mentions. Given that such types of entity mentions are rare, a lot of (general) text would need to be labeled to increase the number of their instances; for example, to reach 100 instances of *Extended Name* would require labeling around

10 times more text similar to what we labeled, potentially requiring years of manual annotation work. If required in future work, a more feasible approach would be to identify and label text with a higher density of particular categories of entity mentions.

- In our fine-grained EL evaluation, we include the results of two CR systems and two WSD systems, comparing a statistical and a neural model for both tasks. Both CR and WSD are active areas of research, with new techniques continuously under development. In future work, it would be interesting to include further CR (e.g. [85, 124, 79]) and WSD systems (e.g., [75, 89]) in our experiments.

8.3 Outlook

EL is an important technique that can help to bridge unstructured text and knowledge bases. In order for EL to reach its full potential, and to serve a wider range of applications, we have argued that it is important to offer better support for a wider range of languages, and to better understand the goals of EL in terms of what it should or should not link.

Our results generally reveal varying opinions on how broad/narrow the goals of EL should be set. Having a broader definition of the goals of the EL task allows for EL systems to capture a wider range of annotations that may be useful, in turn, for a wider range of applications; in particular, having an EL system produce more (correct) annotations is unlikely to be a negative for any application. However, a broader definition of EL’s goals makes the tasks of labeling datasets and developing high-performing EL systems considerably more demanding, posing new challenges for the research community. While we do not take a strong stance on this particular question, we believe that the categorization scheme, datasets¹, guidelines, metrics and results developed in this thesis may help to inform future conventions regarding the EL task, perhaps seeing it split into two separate tasks, with Entity Linking (EL) focusing primarily on named entities (essentially extending the NER task with disambiguation), and Fine-Grained Entity Linking (FEL) focusing on a broader range of entities appearing in a KB.

On the other hand, we also find that whether the goals of EL are set more broadly or more narrowly, there is a strong preference within the EL community for metonymic references to be resolved by EL systems, whereas we find that no evaluated system resolves such references and are not aware of any work that proposes methods to resolve such references (though Ling et al. [88] do discuss the issue). We thus identify this as an open challenge for EL research (and one that does not appear trivial).

Regarding support for multiple languages, our results suggest that Multilingual Entity Linking systems can be built upon three steps: translation, recognition and disambiguation. In other words, in the short-to-medium term, a promising and simple way to bridge EL with other languages is to leverage machine translation on the input text, though of course the

¹The three datasets are available from https://github.com/henryrosalesmendez/categorized_EMNLP_datasets

quality of such translation can still vary depending on the popularity of the language and the amount of parallel corpora available for it. In the longer term, it would be better to build EL systems that are adapted to a particular language or culture. Another open challenge then is to develop EL systems that can provide competitive results not only for popular languages, but also less widely-spoken languages.

Bibliography

- [1] B. Adida, M. Birbeck, S. McCarron, and I. Herman. *RDFa Core 1.1 - Third Edition*. <https://www.w3.org/TR/rdfa-core/>. 2015.
- [2] J. B. et al. *OWL 2 Web Ontology Language Document Overview (Second Edition)*. <https://www.w3.org/TR/owl2-overview/>. 2012.
- [3] A. Bagga and B. Baldwin. “Entity-Based Cross-Document Coreferencing Using the Vector Space Model”. In: *COLING-ACL*. 1998, pp. 79–85.
- [4] T. Baker, P. Vandenbussche, and B. Vatant. “Requirements for vocabulary preservation and governance”. In: *Library Hi Tech* 31.4 (2013), pp. 657–668.
- [5] D. Banerjee, D. Chaudhuri, M. Dubey, and J. Lehmann. *PNEL: Pointer Network based End-To-End Entity Linking over Knowledge Graphs*. <https://www.inf.uni-hamburg.de/en/inst/ab/lt/publications/2020-banerjee-iswc-pnel.pdf>. 2020.
- [6] K. Beck et al. “Manifesto for Agile Software Development”. In: (2001).
- [7] D. Beckett. *Introduction To RDF Query With SPARQL*. <https://www.dajobe.org/talks/200603-sparql-stanford/>. 2006.
- [8] D. Beckett. *N-Triples W3C RDF Core WG Internal Working Draft*. <https://www.w3.org/2001/sw/RDFCore/ntriples/>. 2001.
- [9] D. Beckett, T. Berners-Lee, E. Prud’hommeaux, G. Carothers, and L. Machina. *RDF 1.1 Turtle Terse RDF Triple Language*. <https://www.w3.org/TR/turtle/>. 2014.
- [10] T. Berners-Lee. *An RDF language for the Semantic Web*. <https://www.w3.org/DesignIssues/Notation3>. 2005.
- [11] T. Berners-Lee. *Primer: Getting into RDF & Semantic Web using N3*. <https://www.w3.org/2000/10/swap/Primer.html>. 2000.
- [12] C. Bizer, T. Heath, and T. Berners-Lee. “Linked Data - The Story So Far”. In: *Int. J. Semantic Web Inf. Syst.* 5.3 (2009), pp. 1–22.
- [13] S. Boag, D. Chamberlin, M. F. Fernández, D. Florescu, J. Robie, and J. Siméon. *XQuery 1.0: An XML Query Language (Second Edition)*. <https://www.w3.org/TR/2010/REC-xquery-20101214/>. 2010.
- [14] K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. “Freebase: a collaboratively created graph database for structuring human knowledge”. In: *SIGMOD*. 2008, pp. 1247–1250.

- [15] O. Borrega, M. Taulé, and M. A. Martí. “What do we mean when we speak about Named Entities”. In: *Proceedings of Corpus Linguistics* (2007).
- [16] C. Brando, F. Frontini, and J. Ganascia. “REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets”. In: *CSIMQ 7* (2016), pp. 60–80.
- [17] D. Brickley and R. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema*. <https://www.w3.org/TR/2002/WD-rdf-schema-20020430/>. 2002.
- [18] D. Brickley and R. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema*. <https://www.w3.org/TR/2004/REC-rdf-schema-20040210/>. 2004.
- [19] D. Brickley, R. Guha, and A. Layman. *Resource Description Framework (RDF) Schemas*. <https://www.w3.org/TR/1998/WD-rdf-schema-19980409/>. 1998.
- [20] M. Brümmer, M. Dojchinovski, and S. Hellmann. “DBpedia Abstracts: A Large-Scale, Open, Multilingual NLP Training Corpus”. In: *International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), 2016.
- [21] F. Bry, P. Patranjan, and S. Schaffert. “Xcerpt and XChange - Logic Programming Languages for Querying and Evolution on the Web”. In: *Logic Programming, 20th International Conference, ICLP*. Vol. 3132. Lecture Notes in Computer Science. Springer, 2004, pp. 450–451.
- [22] R. C. Bunescu and M. Pasca. “Using Encyclopedic Knowledge for Named entity Disambiguation”. In: *11st Conference of the European Chapter of the Association for Computational Linguistics, EACL*. The Association for Computer Linguistics, 2006.
- [23] J. J. Carroll and P. Stickler. “RDF triples in XML”. In: *Proceedings of the 13th international conference on World Wide Web - Alternate Track Papers & Posters, WWW*. ACM, 2004, pp. 412–413.
- [24] T. Cassidy, H. Ji, H. Deng, J. Zheng, and J. Han. “Analysis and Refinement of Cross-Lingual Entity Linking”. In: *CLEF*. Vol. 7488. Lecture Notes in Computer Science. Springer, 2012, pp. 1–12.
- [25] M. Chang, B. P. Hsu, H. Ma, R. Loynd, and K. Wang. “E2E: An End-to-End Entity Linking System for Short and Noisy Text”. In: *Workshop on Making Sense of Microposts*. Vol. 1141. CEUR-WS.org, 2014, pp. 62–63.
- [26] B. Chardin, E. Coquery, M. Pailloux, and J. Petit. “RQL: A SQL-Like Query Language for Discovering Meaningful Rules”. In: *IEEE International Conference on Data Mining Workshops, ICDM*. IEEE Computer Society, 2014, pp. 1203–1206.
- [27] E. Charton, M. Gagnon, and B. Ozell. “Automatic semantic web annotation of named entities”. In: *Canadian Conference on Artificial Intelligence*. Springer. 2011, pp. 74–85.
- [28] A. Chisholm and B. Hachey. “Entity disambiguation with web links”. In: *Transactions of the Association for Computational Linguistics 3* (2015), pp. 145–156.

- [29] J. Clark and S. DeRose. *XML Path Language (XPath) Version 1.0*. <https://www.w3.org/TR/1999/REC-xpath-19991116/>. 1999.
- [30] K. Clark and C. D. Manning. “Deep Reinforcement Learning for Mention-Ranking Coreference Models”. In: *EMNLP*. 2016, pp. 2256–2262.
- [31] K. Clark and C. D. Manning. “Entity-Centric Coreference Resolution with Model Stacking”. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2015, pp. 1405–1415.
- [32] M. Cornolti, P. Ferragina, and M. Ciaramita. “A framework for benchmarking entity-annotation systems”. In: *WWW*. 2013, pp. 249–260.
- [33] S. Cucerzan. “Large-Scale Named Entity Disambiguation Based on Wikipedia Data”. In: *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*. ACL, 2007, pp. 708–716.
- [34] S. Cucerzan. “Large-Scale Named Entity Disambiguation Based on Wikipedia Data”. In: *EMNLP-CoNLL (2007)*, p. 708.
- [35] R. Cyganiak, D. Wood, and M. Lanthaler. *RDF 1.1 Concepts and Abstract Syntax*. <https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>. 2014.
- [36] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. “Improving efficiency and accuracy in multilingual entity extraction”. In: *I-SEMANTICS*. ACM. 2013, pp. 121–124.
- [37] A. Delpuch. *OpenTapioca: Lightweight Entity Linking for Wikidata*. arXiv e-prints. 2019, pp. 640–650.
- [38] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva. “Analysis of named entity recognition and linking for tweets”. In: *Inf. Process. Manag.* 51.2 (2015), pp. 32–49.
- [39] M. Dojchinovski and T. Kliegr. “Recognizing, Classifying and Linking Entities with Wikipedia and DBpedia”. In: *WIKT (2012)*, pp. 41–44.
- [40] M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin. “Entity Disambiguation for Knowledge Base Population”. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010, pp. 277–285.
- [41] G. Durrett and D. Klein. “A Joint Model for Entity Analysis: Coreference, Typing, and Linking”. In: *TACL 2 (2014)*, pp. 477–490.
- [42] A. Eckhardt, J. Hresko, J. Procházka, and O. Smrs. “Entity linking based on the co-occurrence graph and entity probability”. In: *ERD’14, Proceedings of the First ACM International Workshop on Entity Recognition & Disambiguation*. ACM, 2014, pp. 37–44.

- [43] M. van Erp, P. N. Mendes, H. Paulheim, F. Ilievski, J. Plu, G. Rizzo, and J. Waitelonis. “Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job”. In: *International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), 2016.
- [44] O. Etzioni, M. J. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. “Unsupervised named-entity extraction from the Web: An experimental study”. In: *Artif. Intell.* 165.1 (2005), pp. 91–134.
- [45] A. Fahrni and M. Strube. “A Latent Variable Model for Discourse-aware Concept and Entity Disambiguation”. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL*. The Association for Computer Linguistics, 2014, pp. 491–500.
- [46] Z. Fang, Y. Cao, Q. Li, D. Zhang, Z. Zhang, and Y. Liu. “Joint Entity Linking with Deep Reinforcement Learning”. In: *The World Wide Web Conference, WWW*. 2019, pp. 438–447.
- [47] Z. Fang, Y. Cao, R. Li, Z. Zhang, Y. Liu, and S. Wang. “High Quality Candidate Generation and Sequential Graph Attention Network for Entity Linking”. In: *The Web Conference, WWW*. 2020, pp. 640–650.
- [48] S. Farrar and D. T. Langendoen. “A linguistic ontology for the semantic web”. In: *GLOT international* 7.3 (2003), pp. 97–100.
- [49] J. D. Fernández, M. A. Martínez-Prieto, and C. Gutiérrez. “Compact Representation of Large RDF Data Sets for Publishing and Exchange”. In: *9th International Semantic Web Conference, ISWC*. Vol. 6496. Lecture Notes in Computer Science. Springer, 2010, pp. 193–208.
- [50] P. Ferragina and U. Scaiella. “TAGME: on-the-fly annotation of short text fragments (by Wikipedia entities)”. In: *CIKM*. 2010, pp. 1625–1628.
- [51] M. Fleischman. “Automated Subcategorization of Named Entities”. In: *ACL (Companion Volume)*. 2001, pp. 25–30.
- [52] M. Fleischman and E. H. Hovy. “Fine Grained Classification of Named Entities”. In: *COLING*. 2002.
- [53] B. Fu, R. Brennan, and D. O’Sullivan. “Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web”. In: *Multilingual Semantic Web*. 2010, pp. 13–20.
- [54] T. Furche, B. Linse, F. Bry, D. Plexousakis, and G. Gottlob. “RDF Querying: Language Constructs and Evaluation Methods Compared”. In: *Reasoning Web, Second International Summer School*. Vol. 4126. Lecture Notes in Computer Science. Springer, 2006, pp. 1–52.
- [55] A. Gangemi, V. Presutti, D. R. Recupero, A. G. Nuzzolese, F. Draicchio, and M. Mongiovì. “Semantic Web Machine Reading with FRED”. In: *Semantic Web* 8.6 (2017), pp. 873–893.

- [56] S. Ghosh, P. Maitra, and D. Das. “Feature Based Approach to Named Entity Recognition and Linking for Tweets”. In: *Proceedings of the 6th Workshop on ‘Making Sense of Microposts’ co-located with the 25th International World Wide Web Conference WWW*. Vol. 1691. 2016, pp. 74–76.
- [57] K. Gimpel et al. “Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments”. In: *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011, pp. 42–47.
- [58] R. Grishman and B. Sundheim. “Message Understanding Conference- 6: A Brief History”. In: *COLING*. 1996, pp. 466–471.
- [59] T. Grütze, G. Kasneci, Z. Zuo, and F. Naumann. “CohEEL: Coherent and efficient named entity linking through random walks”. In: *J. Web Semant.* 37-38 (2016), pp. 75–89.
- [60] S. Guo, M. Chang, and E. Kiciman. “To Link or Not to Link? A Study on End-to-End Tweet Entity Linking”. In: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*. The Association for Computational Linguistics, 2013, pp. 1020–1030.
- [61] Z. Guo, Y. Xu, F. de Sá Mesquita, D. Barbosa, and G. Kondrak. “ualberta at TAC-KBP 2012: English and Cross-Lingual Entity Linking.” In: *TAC*. 2012.
- [62] C. Gutiérrez, C. A. Hurtado, and A. O. Mendelzon. “Formal aspects of querying RDF databases”. In: *Proceedings of SWDB’03, The first International Workshop on Semantic Web and Databases, Co-located with VLDB*. 2003, pp. 293–307.
- [63] P. Haase, J. Broekstra, A. Eberhart, and R. Volz. “A Comparison of RDF Query Languages”. In: *The Semantic Web - ISWC 2004: Third International Semantic Web Conference*. Vol. 3298. Lecture Notes in Computer Science. Springer, 2004, pp. 502–517.
- [64] H. Hajishirzi, L. Zilles, D. S. Weld, and L. S. Zettlemoyer. “Joint Coreference Resolution and Named-Entity Linking with Multi-Pass Sieves”. In: *EMNLP*. 2013, pp. 289–299.
- [65] S. Hakimov, H. ter Horst, S. Jebbara, M. Hartung, and P. Cimiano. “Combining Textual and Graph-Based Features for Named Entity Disambiguation Using Undirected Probabilistic Graphical Models”. In: *Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW*. Vol. 10024. 2016, pp. 288–302.
- [66] X. Han and J. Zhao. “NLPR_KBP in TAC 2009 KBP Track: A Two-Stage Method to Entity Linking”. In: *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*. NIST, 2009.
- [67] S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. “Integrating NLP Using Linked Data”. In: *International Semantic Web Conference (ISWC)*. Springer, 2013, pp. 98–113.

- [68] M. Hlittmair-Delazer, B. Andree, C. Semenza, R. De Bleser, and T. Benke. “Naming by German compounds”. In: *Journal of Neurolinguistics* 8.1 (1994), pp. 27–41.
- [69] J. Hoffart, Y. Altun, and G. Weikum. “Discovering emerging entities with ambiguous names”. In: *WWW*. 2014, pp. 385–396.
- [70] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. “KORE: keyphrase overlap relatedness for entity disambiguation”. In: *CIKM*. 2012, pp. 545–554.
- [71] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. “Robust Disambiguation of Named Entities in Text”. In: *EMNLP*. 2011, pp. 782–792.
- [72] A. Hogan. “The Semantic Web: Two decades on”. In: *Semantic Web* 11.1 (2020), pp. 169–185.
- [73] A. Hogan et al. “Knowledge Graphs”. In: *CoRR* (2020).
- [74] M. Horridge. *Owl syntaxes*. <http://ontogenesis.knowledgeblog.org/88/>. 2010.
- [75] L. Huang, C. Sun, X. Qiu, and X. Huang. “GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2019, pp. 3507–3512.
- [76] I. Hulpuş, N. Prangnawarat, and C. Hayes. “Path-Based Semantic Relatedness on Linked Data and Its Use to Word and Entity Disambiguation”. In: *International Semantic Web Conference (ISWC)*. Springer, 2015, pp. 442–457.
- [77] F. Ilievski, G. Rizzo, M. van Erp, J. Plu, and R. Troncy. “Context-enhanced Adaptive Entity Linking”. In: *International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), 2016.
- [78] K. Jha, M. Röder, and A. N. Ngomo. “All that Glitters Is Not Gold – Rule-Based Curation of Reference Datasets for Named Entity Recognition and Entity Linking”. In: *ESWC*. 2017, pp. 305–320.
- [79] M. Joshi, O. Levy, L. Zettlemoyer, and D. S. Weld. “BERT for Coreference Resolution: Baselines and Analysis”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2019, pp. 5802–5807.
- [80] G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis, and M. Scholl. “RQL: a declarative query language for RDF”. In: *Proceedings of the Eleventh International World Wide Web Conference (WWW)*. ACM, 2002, pp. 592–603.
- [81] M. Kay. *XSL Transformations (XSLT) Version 2.0 (Second Edition)*. <https://www.w3.org/TR/2009/PER-xslt20-20090421/>. 2009.
- [82] N. Kolitsas, O. Ganea, and T. Hofmann. “End-to-End Neural Entity Linking”. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL*. 2018, pp. 519–529.
- [83] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. “Collective annotation of Wikipedia entities in web text”. In: *SIGKDD*. 2009, pp. 457–466.

- [84] O. Lassila and R. R. Swick. *Resource Description Framework (RDF) Model and Syntax Specification*. <https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>. 1999.
- [85] K. Lee, L. He, and L. Zettlemoyer. “Higher-Order Coreference Resolution with Coarse-to-Fine Inference”. In: *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2018, pp. 687–692.
- [86] J. Lehmann et al. “DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia”. In: *Semantic Web 6.2 (2015)*, pp. 167–195.
- [87] M. Lesk. “Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone”. In: *International Conference on Systems Documentation (SIGDOC)*. ACM, 1986, pp. 24–26.
- [88] X. Ling, S. Singh, and D. S. Weld. “Design challenges for entity linking”. In: *TACL 3 (2015)*, pp. 315–328.
- [89] D. Loureiro and A. Jorge. “Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation”. In: *Association for Computational Linguistics (ACL)*. 2019, pp. 5682–5691.
- [90] G. Luo, X. Huang, C. Lin, and Z. Nie. “Joint Entity Recognition and Disambiguation”. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. The Association for Computer Linguistics, 2015, pp. 879–888.
- [91] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *ACL*. 2014, pp. 55–60.
- [92] M. A. Martínez-Prieto, M. A. Gallego, and J. D. Fernández. “Exchange and Consumption of Huge RDF Data”. In: *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC*. Vol. 7295. Lecture Notes in Computer Science. Springer, 2012, pp. 437–452.
- [93] J. L. Martínez-Rodríguez, A. Hogan, and I. Lopez-Arevalo. “Information Extraction meets the Semantic Web: A Survey”. In: *Semantic Web Journal (2019)*.
- [94] P. H. Martins, Z. Marinho, and A. F. T. Martins. “Joint Learning of Named Entity Recognition and Entity Linking”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*. 2019, pp. 190–196.
- [95] J. P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, and P. Cimiano. “The OntoLemon Model: Development and Applications”. In: *eLex*. 2017, pp. 19–21.
- [96] J. P. McCrae, D. Spohr, and P. Cimiano. “Linking Lexical Resources and Ontologies on the Semantic Web with Lemon”. In: *ESWC*. 2011, pp. 245–259.
- [97] M. A. Medina, J. A. Sánchez, and R. O. Chávez. “RDF-based Model for Encoding Document Hierarchies”. In: *17th International Conference on Electronics, Communications and Computers, CONIELECOMP*. IEEE Computer Society, 2007, p. 22.

- [98] G. de Melo. “Lexvo.org: Language-related information for the Linguistic Linked Data cloud”. In: *Semantic Web 6.4* (2015), pp. 393–400.
- [99] G. de Melo and G. Weikum. “Language as a Foundation of the Semantic Web”. In: *Proceedings of the Poster and Demonstration Session at ISWC*. 2008.
- [100] J. Melton and J. Spiegel. *XQueryX 3.1*. <https://www.w3.org/TR/xqueryx-31/>. 2017.
- [101] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. “DBpedia Spotlight: shedding light on the web of documents”. In: *I-SEMANTICS*. 2011, pp. 1–8.
- [102] R. Mihalcea and A. Csomai. “Wikify!: linking documents to encyclopedic knowledge”. In: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM*. ACM, 2007, pp. 233–242.
- [103] A. Miles and S. Bechhofer. *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation. <https://www.w3.org/2004/02/skos/>. 2009.
- [104] G. A. Miller and C. Fellbaum. “WordNet then and now”. In: *Language Resources and Evaluation (LRE)* 41.2 (2007), pp. 209–214.
- [105] L. Miller, A. Seaborne, and A. Reggiori. “Three Implementations of SquishQL, a Simple RDF Query Language”. In: *The Semantic Web - ISWC 2002, First International Semantic Web Conference*. Vol. 2342. Lecture Notes in Computer Science. Springer, 2002, pp. 423–435.
- [106] T. Miller and I. Gurevych. “WordNet—Wikipedia—Wiktionary: Construction of a Three-way Alignment”. In: *International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), 2014, pp. 2094–2100.
- [107] D. N. Milne and I. H. Witten. “Learning to link with Wikipedia”. In: *ACM Conference on Information and Knowledge Management (CIKM)*. ACM, 2008, pp. 509–518.
- [108] A. Minard, M. Speranza, R. Urizar, B. Altuna, M. van Erp, A. Schoen, and C. van Son. “MEANTIME, the NewsReader Multilingual Event and Time Corpus”. In: *LREC*. 2016.
- [109] A. Moro and R. Navigli. “SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking”. In: *SemEval@NAACL-HLT*. ACL, 2015, pp. 288–297.
- [110] A. Moro, A. Raganato, and R. Navigli. “Entity Linking meets Word Sense Disambiguation: a Unified Approach”. In: *TACL 2* (2014), pp. 231–244.
- [111] D. Moussallem and et al. “MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach”. In: *K-CAP*. 2017, p. 9.
- [112] D. Nadeau. “Balie—baseline information extraction: Multilingual information extraction from text with machine learning and natural language techniques”. In: *Technical report*. University of Ottawa, 2005.

- [113] D. Nadeau and S. Sekine. “A survey of named entity recognition and classification”. In: *Linguisticae Investigationes* 30.1 (2007), pp. 3–26.
- [114] P. Nakov. “On the interpretation of noun compounds: Syntax, semantics, and entailment”. In: *Nat. Lang. Eng.* 19.3 (2013), pp. 291–330.
- [115] F. Narducci, M. Palmonari, and G. Semeraro. “Cross-language Semantic Matching for Discovering Links to e-gov Services in the LOD Cloud”. In: *Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data*. Vol. 992. 2013, pp. 21–32.
- [116] R. Navigli. “Word sense disambiguation: A survey”. In: *ACM Comput. Surv.* 41.2 (2009), 10:1–10:69.
- [117] A. N. Ngomo, M. Röder, D. Moussallem, R. Usbeck, and R. Speck. “BENGAL: An Automatic Benchmark Generator for Entity Recognition and Linking”. In: *Proceedings of the 11th International Conference on Natural Language Generation*. 2018, pp. 339–349.
- [118] F. Odoni, P. Kuntschik, A. M. P. Brasoveanu, and A. Weichselbraun. “On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance”. In: *SEMANTICS*. 2018, pp. 33–42.
- [119] C. Ogbuji. *Versa: Path-Based RDF Query Language*. <https://www.xml.com/pub/a/2005/07/20/versa.html>. 2005.
- [120] A. Olieman, H. Azaronyad, M. Dehghani, J. Kamps, and M. Marx. “Entity linking by focusing DBpedia candidate entities”. In: *Proceedings of the First International Workshop on Entity Recognition & Disambiguation, ERD*. ACM, 2014, pp. 13–24.
- [121] A. Pappu, R. Blanco, Y. Mehdad, A. Stent, and K. Thadani. “Lightweight Multilingual Entity Extraction and Linking”. In: *WSDM*. ACM. 2017, pp. 365–374.
- [122] S. Perera, P. N. Mendes, A. Alex, A. P. Sheth, and K. Thirunarayan. “Implicit Entity Linking in Tweets”. In: *Extended Semantic Web Conference (ESWC)*. Springer, 2016, pp. 118–132.
- [123] E. Pietriga. *Fresnel Selector Language for RDF (FSL)*. <https://www.w3.org/2005/04/fresnel-info/fsl/>. 2005.
- [124] J. Plu, R. Prokofyev, A. Tonon, P. Cudré-Mauroux, D. E. Difallah, R. Troncy, and G. Rizzo. “Sanaphor++: Combining Deep Neural Networks with Semantics for Coreference Resolution”. In: *International Conference on Language Resources and Evaluation (LREC)*. 2018.
- [125] J. Plu, G. Rizzo, and R. Troncy. “Enhancing Entity Linking by Combining NER Models”. In: *ESWC*. 2016, pp. 17–32.
- [126] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. “KIM - Semantic Annotation Platform”. In: *The Semantic Web - ISWC 2003, Second International Semantic Web Conference*. Vol. 2870. Lecture Notes in Computer Science. Springer, 2003, pp. 834–849.

- [127] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov. “KIM - a semantic platform for information extraction and retrieval”. In: *Nat. Lang. Eng.* 10.3-4 (2004), pp. 375–392.
- [128] S. Powers. *Practical RDF*. 1st. O’Reilly Media, 2003.
- [129] E. Prud’hommeaux and A. Seaborne. *SPARQL Query Language for RDF*. <https://www.w3.org/TR/rdf-sparql-query/>. 2008.
- [130] D. Rao, P. McNamee, and M. Dredze. “Entity Linking: Finding Extracted Entities in a Knowledge Base”. In: *Multi-source, Multilingual Information Extraction and Summarization*. Springer, 2013, pp. 93–115.
- [131] L. Ratinov and D. Roth. “Design Challenges and Misconceptions in Named Entity Recognition”. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL*. ACL, 2009, pp. 147–155.
- [132] L. Ratinov, D. Roth, D. Downey, and M. Anderson. “Local and Global Algorithms for Disambiguation to Wikipedia”. In: *ACL*. 2011, pp. 1375–1384.
- [133] T. Rebele, F. M. Suchanek, J. Hoffart, J. Biega, E. Kuzey, and G. Weikum. “YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames”. In: *International Semantic Web Conference (ISWC)*. 2016, pp. 177–185.
- [134] G. Rizzo and R. Troncy. “NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools”. In: *EACL*. The Association for Computer Linguistics, 2012, pp. 73–76.
- [135] M. Röder, R. Usbeck, S. Hellmann, D. Gerber, and A. Both. “N³ – A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format”. In: *International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), 2014, pp. 3529–3533.
- [136] M. Röder, R. Usbeck, S. Hellmann, D. Gerber, and A. Both. “N³ - A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format”. In: *LREC*. 2014, pp. 3529–3533.
- [137] H. Rosales-Méndez, A. Hogan, and B. Poblete. “Fine-Grained Entity Linking”. In: *J. Web Semant.* 65 (2020), p. 100600.
- [138] H. Rosales-Méndez, A. Hogan, and B. Poblete. “Fine-Grained Evaluation for Entity Linking”. In: *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP/IJCNLP)*. 2019.
- [139] H. Rosales-Méndez, A. Hogan, and B. Poblete. “NIFify: Towards Better Quality Entity Linking Datasets”. In: *WWW Companion Volume*. 2019, pp. 815–818.
- [140] H. Rosales-Méndez, A. Hogan, and B. Poblete. “VoxEL: A Benchmark Dataset for Multilingual Entity Linking”. In: *International Semantic Web Conference (ISWC)*. Springer, 2018, pp. 170–186.

- [141] H. Rosales-Méndez, B. Poblete, and A. Hogan. “What should Entity Linking link?” In: *Alberto Mendelzon Workshop (AMW)*. 2018.
- [142] M. Ruiz-Casado, E. Alfonseca, and P. Castells. “Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia”. In: *Natural Language Processing and Information Systems*. 2005, pp. 67–79.
- [143] F. Sasaki, M. Dojchinovski, and J. Nehring. “Chainable and Extendable Knowledge Integration Web Services”. In: *Knowledge Graphs and Language Technology - ISWC 2016 International Workshops: KEKI and NLP&DBpedia*. Vol. 10579. Lecture Notes in Computer Science. Springer, 2016, pp. 89–101.
- [144] A. Seaborne. *RDQL - A Query Language for RDF*. <https://www.w3.org/Submission/2004/SUBM-RDQL-20040109/>. 2004.
- [145] W. Shen, J. Wang, and J. Han. “Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions”. In: *IEEE Trans. Knowl. Data Eng.* 27.2 (2015), pp. 443–460.
- [146] W. Shen, J. Wang, P. Luo, and M. Wang. “LINDEN: linking named entities with knowledge base via semantic knowledge”. In: *Proceedings of the 21st World Wide Web Conference 2012, WWW*. ACM, 2012, pp. 449–458.
- [147] A. Sil, G. Kundu, R. Florian, and W. Hamza. “Neural Cross-Lingual Entity Linking”. In: *AAAI*. AAAI Press, 2018, pp. 5464–5472.
- [148] A. Sil and A. Yates. “Re-ranking for joint named-entity recognition and linking”. In: *22nd ACM International Conference on Information and Knowledge Management, CIKM*. 2013, pp. 2369–2374.
- [149] M. Sintek and S. Decker. “TRIPLE - An RDF Query, Inference, and Transformation Language”. In: *Proceedings of the 14th International Conference on Applications of Prolog, INAP*. The Prolog Association of Japan, 2001, pp. 47–56.
- [150] M. K. Smith, C. Welty, and D. L. McGuinness. *OWL Web Ontology Language Guide*. <https://www.w3.org/TR/2004/REC-owl-guide-20040210/>. 2004.
- [151] R. Speck and et al. “Ensemble Learning of Named Entity Recognition Algorithms using Multilayer Perceptron for the Multilingual Web of Data”. In: *K-CAP*. 2017, p. 26.
- [152] R. Speck and A. N. Ngomo. “Named Entity Recognition using FOX”. In: *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC*. Vol. 1272. 2014, pp. 85–88.
- [153] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, and N. Lindström. *JSON-LD 1.0*. <https://www.w3.org/TR/2014/REC-json-ld-20140116/>. 2014.
- [154] P. F. Strawson. “On referring”. In: *Mind* 59.235 (1950), pp. 320–344.

- [155] F. M. Suchanek, G. Ifrim, and G. Weikum. “LEILA: Learning to Extract Information by Linguistic Analysis”. In: *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge@COLING/ACL*. 2006, pp. 18–25.
- [156] F. M. Suchanek, G. Kasneci, and G. Weikum. “Yago: a core of semantic knowledge”. In: *Proceedings of the 16th International Conference on World Wide Web (WWW) , Banff, Alberta, Canada, May 8-12*. ACM, 2007, pp. 697–706.
- [157] R. Sukthanker, S. Poria, E. Cambria, and R. Thirunavukarasu. “Anaphora and Coreference Resolution: A Review”. In: *CoRR* abs/1805.11824 (2018).
- [158] P. Taufer. “Named Entity Recognition and Linking”. MA thesis. Univerzita Karlova, 2017.
- [159] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”. In: *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL*. 2003.
- [160] F. Tristram, S. Walter, P. Cimiano, and C. Unger. “Weasel: a Machine Learning Based Approach to Entity Linking combining different features”. In: *Proceedings of the Third NLP&DBpedia Workshop (NLP & DBpedia 2015) co-located with the 14th International Semantic Web Conference 2015 ISWC*. Vol. 1581. CEUR Workshop Proceedings. 2015, pp. 25–32.
- [161] C.-T. Tsai and D. Roth. “Cross-lingual wikification using multilingual embeddings”. In: *Proceedings of NAACL-HLT*. 2016, pp. 589–598.
- [162] V. S. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna. “Semantic annotation for knowledge management: Requirements and a survey of the state of the art”. In: *J. Web Semant.* 4.1 (2006), pp. 14–28.
- [163] R. Usbeck, A. N. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, and A. Both. “AGDISTIS - Agnostic Disambiguation of Named Entities Using Linked Open Data”. In: *European Conference on Artificial Intelligence (ECAI)*. Springer, 2014, pp. 1113–1114.
- [164] R. Usbeck et al. “GERBIL: General Entity Annotator Benchmarking Framework”. In: *International Conference on World Wide Web (WWW)*. ACM, 2015, pp. 1133–1143.
- [165] P. Vandenbussche, G. Atemezing, M. Poveda-Villalón, and B. Vatant. “Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web”. In: *Semantic Web* 8.3 (2017), pp. 437–452.
- [166] L. Vial, B. Lecouteux, and D. Schwab. “Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation”. In: *Proceedings of the 10th Global Wordnet Conference*. 2019.

- [167] D. Vrandečić and M. Krötzsch. “Wikidata: a free collaborative knowledgebase”. In: *Commun. ACM* 57.10 (2014), pp. 78–85.
- [168] J. Waitelonis, C. Exeler, and H. Sack. “Linked Data enabled generalized vector space model to improve document retrieval”. In: *NLP & DBpedia @ ISWC*. 2015.
- [169] J. Waitelonis, H. Jürges, and H. Sack. “Don’t compare Apples to Oranges: Extending GERBIL for a fine grained NEL evaluation”. In: *SEMANTICS*. 2016, pp. 65–72.
- [170] Z. Wang, J. Zhang, J. Feng, and Z. Chen. “Knowledge Graph and Text Jointly Embedding”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*. ACL, 2014, pp. 1591–1601.
- [171] Z. Wang, J. Li, and J. Tang. “Boosting Cross-Lingual Knowledge Linking via Concept Annotation.” In: *IJCAI*. 2013, pp. 2733–2739.
- [172] G. Wu, Y. He, and X. Hu. “Entity Linking: An Issue to Extract Corresponding Entity With Knowledge Base”. In: *IEEE Access* 6 (2018), pp. 6220–6231.
- [173] I. Yamada, H. Takeda, and Y. Takefuji. “Enhancing Named Entity Recognition in Twitter Messages Using Entity Linking”. In: *Proceedings of the Workshop on Noisy User-generated Text, NUT@IJCNLP*. 2015, pp. 136–140.
- [174] Y. Yang and M. Chang. “S-MART: Novel Tree-based Structured Learning Algorithms Applied to Tweet Entity Linking”. In: *ACL*. 2015, pp. 504–513.
- [175] L. A. Zadeh. “Fuzzy sets”. In: *Information and control* 8.3 (1965), pp. 338–353.
- [176] T. Zhang, K. Liu, and J. Zhao. “Cross Lingual Entity Linking with Bilingual Topic Model”. In: *Proceedings of the 23rd International Joint Conference on Artificial Intelligence, IJCAI*. IJCAI/AAAI, 2013, pp. 2218–2224.
- [177] W. Zhang, Y. C. Sim, J. Su, and C. L. Tan. “Entity Linking with Effective Acronym Expansion, Instance Selection, and Topic Modeling”. In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI*. IJCAI/AAAI, 2011, pp. 1909–1914.
- [178] S. Zhou, S. Rijhwani, J. Wieting, J. G. Carbonell, and G. Neubig. “Improving Candidate Generation for Low-resource Cross-lingual Entity Linking”. In: *Trans. Assoc. Comput. Linguistics* 8 (2020), pp. 109–124.
- [179] Z. Zuo, G. Kasneci, T. Gruetze, and F. Naumann. “BEL: Bagging for Entity Linking.” In: *COLING*. 2014, pp. 2075–2086.
- [180] S. Zwicklbauer, C. Seifert, and M. Granitzer. “DoSeR - A Knowledge-Base-Agnostic Framework for Entity Disambiguation Using Semantic Embeddings”. In: *The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC*. Vol. 9678. Springer, 2016, pp. 182–198.