

Fine-Grained Entity Linking

Henry Rosales-Méndez, Aidan Hogan, Barbara Poblete
Department of Computer Science, University of Chile & IMFD Chile
{hrosales, ahogan, bpoblete}@dcc.uchile.cl

Abstract

The Entity Linking (EL) task involves linking mentions of entities in a text with their 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? In this work, we first design a questionnaire to understand what kinds of mentions and links the EL research community believes should be targeted by the task. Based on these results we propose a fine-grained categorization scheme for EL that distinguishes different types of mentions and links. We propose a vocabulary extension that allows to express 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.

Keywords: Entity Linking, Fine-Grained Entity Linking, Information Extraction, Benchmark

1. Introduction

Numerous Knowledge Bases (KB) are now available online, including semi-structured KBs such as Wikipedia, and structured KBs such as BabelNet [1], DBpedia [2], Freebase [3], Wikidata [4], YAGO [5], 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. Such KBs have been used in diverse applications, including, but not limited to, Information Extraction from text [6], the goal of which is to enhance the machine readable structure of natural language.

A foundational task for Information Extraction is the goal of Entity Linking (EL), which involves identifying entity mentions in a text (or potentially a semi-structured source [6]) 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.¹ 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. As such, EL provides an important bridge between unstructured sources of data (text) and (semi-)structured sources of data (KBs). 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 manager?”), among others [7, 6].

Given the central importance of the EL task, a broad number of EL techniques and systems have been proposed in recent years [7]. The EL task can generally be sub-divided into two high-level sub-tasks [7, 6]. The first sub-task is *recognition*, where entity mentions in the text – e.g., “Michael Jackson” and “Joseph Jackson” – are identified. The second is *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:

¹We use prefixes as denoted in <http://prefix.cc/>.

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: Annotations of Babelfy (b), DBpedia Spotlight (d), FRED (f) and TagME (t) on the same sentence [11]

- 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”).

Many techniques have then been proposed down through the years to address these sub-tasks [7, 6]; we can distinguish two high-level strategies employed by different systems, which we term: *Named Entity Recognition & Linking (NERL)* systems [8] and *End-to-End Entity Linking (E2E)* systems [9].²

NERL systems decouple the recognition and disambiguation steps of the Entity Linking task [12, 13, 14, 15, 16]. 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 targetted by NER systems was defined in the Message Understanding Conference 6 (MUC-6) [17], 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 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 [18, 19, 1, 20], 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 [10].

²We remark, however, that the precise definitions vary from author to author, where we introduce the convention used here; e.g., Luo et al. [10] refer to E2E systems as *Joint Entity Recognition and Linking (JERL)*.

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 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 targetted by NER tools, the album “Got to Be There” may not be detected although it is a named entity.³ 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 [22]. Recognizing these relative strengths and weaknesses, hybrid [23] and ensemble [24] approaches propose to combine NERL and E2E results.

In summary, 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 [22, 25, 26, 27, 28]. 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 presents the entity mentions recognized by a selection of popular online EL systems – Babelfy (strict configuration) [1], DBpedia Spotlight [19], FRED [16] and TagME [18] – 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”, behaviour differs across EL systems for *common entities* based on common noun phrases like “interview” [22]; 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

³It 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 [21] of a fine-grained classification of named entities; however, NER tools still predominantly follow MUC-6 definitions.

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 [22, 25, 26, 27, 28]. A typical approach to address this issue 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 designing an EL system, labeling an EL dataset, or performing evaluation. In this paper, 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.

This paper extends upon a previous conference paper [11], where we initially presented a fine-grained categorization of EL mentions and links, and performed experiments with respect to popular EL systems with interfaces available online. In comparison with our previous work, the novel contributions include:

- a detailed discussion of related works on EL benchmark datasets, formats, and design issues;
- a vocabulary that extends the NLP Interchange Format [29], supporting fine-grained labeling of EL datasets;
- the extension of existing EL systems with techniques for Coreference Resolution and Word Sense Disambiguation and their subsequent evaluation using our datasets;
- detailed guidelines used for labeling fine-grained EL datasets with our categorization scheme;
- extensions of a system used for annotating and validating datasets using our proposed categorization scheme;
- proofs of desirable properties of the metrics presented (previously presented as propositions without proofs).

We also provide extended discussion throughout.

The structure of this paper is then as follows:

Section 2 We discuss previous works that address the lack of consensus on the EL task, as well as surveying the existing EL datasets that have been proposed in the literature.

Section 3 We prepare a short questionnaire intending to understand what consensus on the goals of the EL task exist among authors of EL papers, further presenting the results.

Section 4 We propose a fine-grained categorization of EL mentions and links that allows for understanding the performance of different systems for different entity types.

Section 5 We propose an extension of existing vocabularies in order to express EL datasets annotated with fine-grained categories, as well as alternative links.

Section 6 Selecting three existing EL datasets in English – namely KORE50, VoxEL and a subset of ACE2004 – we relabel them with fine-grained categories and further links.

Section 7 Selecting five popular EL systems with online APIs – AIDA, Babelify, DBpedia Spotlight, FREME and TagME – we evaluate their performance for different categories of mentions and links using our relabeled datasets.

Section 8 We extend the previous five systems with off-the-shelf techniques for Coreference Resolution and Word Sense Disambiguation to extend their coverage, evaluating the impact on performance for different categories.

Section 9 We next propose novel metrics for recall and F_1 that address the lack of consensus by considering fuzzy sets, thus weighting annotations differently.

Section 10 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.

2. Related Works

Having provided an overview of EL strategies and tools in the introduction, we now focus on related works in three aspects pertinent to this work: EL benchmark datasets, EL formats, as well as previous works discussing design issues relating to EL.

2.1. 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. Evaluation benchmarks such as GERBIL [30] then allow for computing and visualizing such measures with respect to different EL datasets and systems.

In Table 1, we provide a brief overview of existing EL datasets [31, 6]. We see that a selection of datasets have been proposed, where most have been manually labeled; note that most marked \times were previously NER datasets to which KB links were added, with one exception being DBpedia Abstracts [32], 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 [33], DBpedia Abstracts [32], MEANTIME [34] and VoxEL [31] were designed specifically for evaluating multilingual EL systems,

Table 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 [35]	✗	✗	MSNBC
AQUAINT [36]	✗	✗	MSNBC
IITB [37]	✓	✗	IITB
ACE2004 [38]	✗	✗	MSNBC
AIDA/CoNLL [12]	✓	✗	AIDA
DBpedia Spotlight [19]	✓	✗	Lexvo
KORE50 [13]	✓	✗	AIDA
N ³ -RSS 500 [39]	✓	✗	NIF
Reuters 128 [39]	✓	✗	NIF
News-100 [39]	✓	✗	NIF
Wes2015 [40]	✓	✗	NIF
SemEval2015 Task 13 [33]	✓	✗	SemEval
Thibaudet [41]	✗	✓	REDEN
Bergson [41]	✗	✓	REDEN
DBpedia Abstracts [32]	✗	✗	NIF
MEANTIME [34]	✓	✓	CAT
VoxEL [31]	✓	✗	NIF

providing annotated texts in multiple languages. On the other hand, KORE50 [13] is intended as a succinct but challenging collection of highly-ambiguous entities in short sentences. Furthermore, DBpedia Abstracts [32] is intended for the purposes of training multilingual EL systems. Further details on these datasets can be found in the survey by Martinez-Rodriguez et al. [6] as well as in the discussions by Usbeck et al. [30], van Erp et al. [26] and Jha et al. [27] on EL evaluation.

2.2. EL Formats

As seen previously in Table 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 [42], which uses an XML-based format; we provide an example of the format in Figure 2, describing the mention “Jackson” 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 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”⁴ 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 4 that all words for sentence **S1** are tagged with O, except “Jackson”, which is the only annotation in this example.

⁴<https://www.clips.uantwerpen.be/conll2003/ner/>

Figure 2: MSNBC format for EL annotations

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

Figure 3: IITB format for EL annotations

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

In 2015, SemEval competitions began including a track dedicated to Entity Linking, further introducing a new format for EL benchmark datasets [33]. In Figure 5 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 [34] dataset, which consists of 120 news articles from WikiNews11 with manual annotations of entities, events, temporal information and semantic roles. MEANTIME was built with the CAT⁵ 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 6 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 [43]⁶ 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* [44]⁷ – and its extensions *lemon-LexInfo*⁸ and *ontolex-lemon* [45]⁹ – 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

⁵<https://dh.fbk.eu/resources/cat-content-annotation-tool>

⁶<http://linguistics-ontology.org/gold-2010.owl>

⁷<https://lemon-model.net/lemon>

⁸<https://www.lexinfo.net/ontology/3.0/lexinfo.ttl>

⁹<https://www.w3.org/2016/05/ontolex/>

Figure 4: 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
```

Figure 5: 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
```

Figure 6: 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>
```

Figure 7: 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>.
```

Due to the advantages and popularity of NIF, benchmark tools – such as GERBIL [30]¹⁰ and NIFify [48]¹¹ – are based on the NIF format, and support converting other EL formats to NIF.

2.3. 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) [17], the concept of a “named entity” was defined as those phrases in a text that refer to instances 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 [21] 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. [49] proposed

al. [46, 47] proposed Lexvo as a 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. [29] 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 7 for the running example.

Recalling Table 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 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, Wes2015 and VoxEL were created with NIF, where others have further been transcribed from their own formats to NIF (e.g., ACE04, DBpedia Spotlight and KORE50).

¹⁰<http://aksw.org/Projects/GERBIL.html>

¹¹https://github.com/henryrosalesmendez/NIFify_v3

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 [50], 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.

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. [22] 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. [51] 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, 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 [52]. 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. [51], Ling et al. [22], van Erp et al. [26]¹², Jha et al. [27], and more besides, with differing opinions; for example, Ling et al. [22] consider overlapping mentions to be useful to include, while Jha et al. [27] consider overlapping mentions to be an error.

Ling et al. [22] further raise two other (more subtle) 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 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 we saw in Figure 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 [33] and DBpedia Spotlight [19] datasets allow overlapping entities, while datasets such as ACE2004 [38] and AIDA/CoNLL [12] do not; in fact, Jha et al. [27] consider the overlapping mentions in DBpedia Spotlight to be errors and remove them. On the other hand, VoxEL [31] provides a strict and a relaxed version of the dataset, with the former containing non-overlapping named entities, and the latter further containing overlapping common entities. We also note that MEANTIME [34] 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.

Our goal in this paper 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.

3. Questionnaire on the Goals of EL

Based on the previous discussion, we see that there are often diverging perspectives with respect 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 8 (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:

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)?

¹²This 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.

In an [interview]^{0.19} with [Martin Bashir]^{1.00} for the [2003]^{0.28} [documentary]^{0.28} [Living with {Michael Jackson}^{0.75}]^{0.97}, the [{King}^{0.08} of {Pop}^{0.33}]^{0.94} [recalled]^{0.06} that [Joe]^{1.00} often [sat]^{0.08} with a [white]^{0.11} [belt]^{0.14} at [hand]^{0.14} as [{he}^{0.56} and {his}^{0.39} {four}^{0.08} {siblings}^{0.14}]^{0.50} [rehearsed]^{0.08}.

[Russian]^{0.61} [daily]^{0.14} [Kommersant]^{0.97} [reports]^{0.06} that [Moscow]^{0.94} will [supply]^{0.06} the [Greeks]^{0.94} with [gas]^{0.36} at [(rock)^{0.00} bottom {prices}^{0.19}]^{0.28} as [Tsipras]^{0.92} [prepares]^{0.03} to [meet]^{0.06} the [(Russian)^{0.53} {President}^{0.12}]^{0.97}.

Figure 8: 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 2 [11].

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 8, 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.¹³

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. [7], 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 online¹⁴, where in Figure 8 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 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

¹³The questionnaire design can be reviewed online: <https://users.dcc.uchile.cl/~hrosales/questionnaire>

¹⁴<https://users.dcc.uchile.cl/~hrosales/questionnaire>

Table 2: The ratio of respondents choosing particular links for mentions with multiple choices (underlined> in Figure 8; the questions were multiple choice, so respondents could choose multiple possibilities [11]

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

be considered, with “he” (0.56) and “he and his four siblings” (0.5)¹⁵ 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 8 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 2, 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 8, 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 common entities as part of the EL task. Most respondents agree that some non-noun phrases can

¹⁵One 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.

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 [50] – 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 [52] – 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. Fine-Grained Categories

Following the discussion of EL design issues by numerous authors [22, 25, 26, 27, 28] and the results of the questionnaire, 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 9. 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 9) 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), with the scheme being extended until it was sufficient to capture all of the cases that we encountered in these datasets. However, the

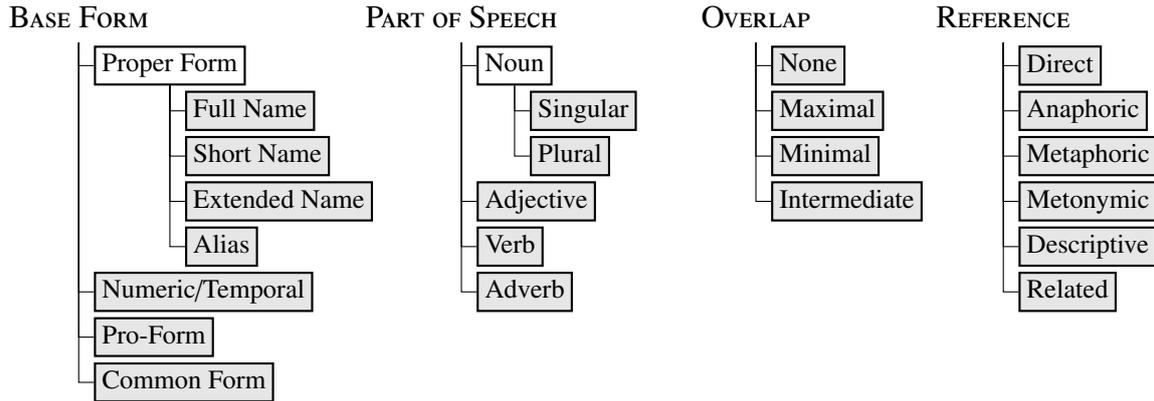


Figure 9: EL categorization scheme with concrete alternatives (leaf-nodes) shaded for each dimension

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.

4.1. 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.¹⁶
 - **Full Name**: Denotes that the mention corresponds to the primary label of the entity in the KB, or is a minor variation thereof¹⁷; 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,

¹⁶For 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”.

¹⁷More specifically, we consider that the (case-normalized) lemmas of each word in the mention and the primary label correspond in the same order.

“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 `wiki:Michael_Jackson`.¹⁸
- **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.).

4.2. 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”.

¹⁸The mention should contain the (case-normalized) lemmas of the primary label in order, possibly interrupted by other lemmas.

- **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”.

4.3. 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.

4.4. Reference

The *Reference* dimension indicates the manner in which the mention makes reference to the linked KB entity [53]. 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`.¹⁹ We also use this category to complement metonymic references, where “[Moscow] will supply” will also be linked to `wiki:Moscow` in an annotation with the related category.

5. Fine-Grained EL Format

In Section 2.2, 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.

5.1. Vocabulary

We show the Fine-Grained Entity Linking (FEL) vocabulary in Figure 10, with newly defined terms using the `fel:` prefix. The categories of Figure 9 are defined as classes, forming a subclass hierarchy. We follow a set of rules proposed by Baker et al. [54] 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 DOAP²⁰ 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  license²¹ encouraging its re-use.

¹⁹In 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.

²⁰<http://usefulinc.com/ns/dojo>

²¹<https://creativecommons.org/licenses/by/3.0/>

- Further changes will be managed with the GitHub²² 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 [55] (as shown in Figure 10).

To satisfy global rules, we submit the FEL vocabulary to the Linked Open Vocabularies system [56]²³: 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 VoID²⁴ vocabulary to allow data providers to discover what terms the vocabulary uses.
- We guarantee the persistence of our URIs storing our vocabulary on a server²⁵ 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 Group²⁶ which can be redirected to another destination.
- To embrace the “safety through redundancy” principle [54] which advocates for mirroring information online, we make a second copy available in a GitHub repository²⁷.

5.2. 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 [40], 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 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 can be specified in NIF on an annotation with the property `itsrdf:taClassRef`.²⁸ 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.”

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 fel:entityType o` for each link in the annotation, where `s` denotes the KB identifier, not the mention.

In Figure 11 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 12 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

²²<https://github.com/henryrosalesmendez/fel>

²³<https://lov.linkeddata.es/dataset/lov/>

²⁴<http://vocab.deri.ie/void>

²⁵<https://cutt.ly/2yEvqp0>

²⁶<https://www.w3.org/community/perma-id/>

²⁷<https://github.com/henryrosalesmendez/fel>

²⁸See example: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/example.ttl>; October 5th, 2019.

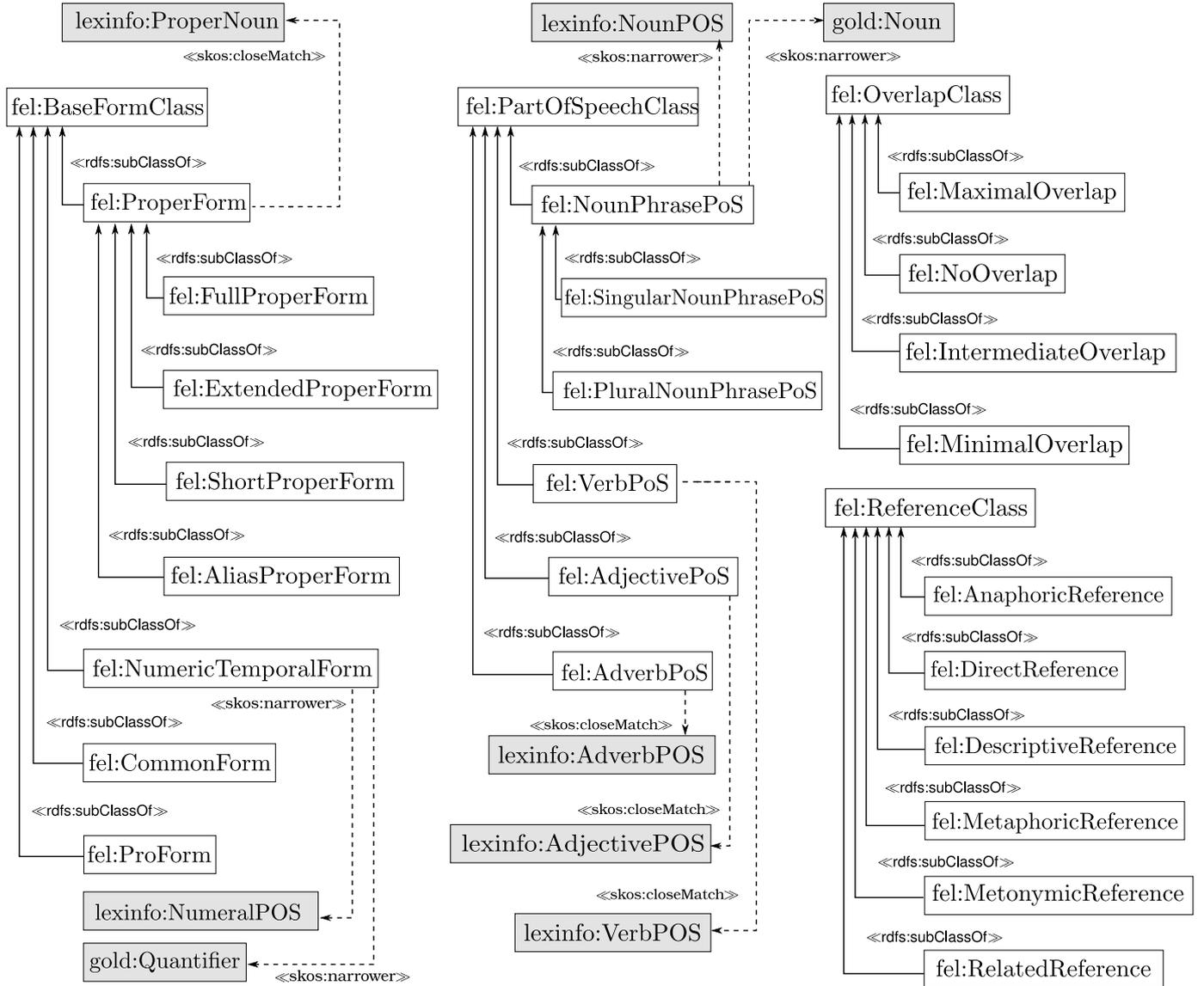


Figure 10: Hierarchy of classes belonging to the Fine-Grained Entity Linking vocabulary and its links to external vocabularies.

change for the different links; this assumption allows us to annotate *AND* cases in a lightweight manner (e.g., without having to introduce 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.
- *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.

5.3. Extending NIFify

While our vocabulary allows for fine-grained annotation of EL datasets, in the case of benchmark datasets, such annotation is typically performed by humans; by providing a fine-grained categorization and format for EL, we will be able to distinguish the performance of systems for different types of mentions and links. However, this adds significant additional cost when labeling the dataset by hand. To mitigate these costs, we extend our NIFify tool, described in previous work [48], which provides a user interface for manually and/or semi-automatically annotating, visualizing and validating NIF datasets, as well as for benchmarking EL systems. We initially created NIFify to help with the annotation of the VoxEL [31] dataset. Since then,

Figure 11: 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 .
```

we have further extended the NIFify tool²⁹ to support our fine-grained EL format. We briefly describe the extensions here.

In terms of the fine-grained categories, some can be labeled automatically with relatively high precision while others cannot. In particular, for the purposes of the PART OF SPEECH and OVERLAP dimensions, NIFify allows for generating suggestions that can be modified by the annotator. NIFify further implements a number of validation services (similar in principle to other tools such as Eaglet [27], though the rules vary) that help to detect and review common types of errors; the base version of NIFify already included the following services [31]:

Boundary Error: detects when a label includes characters that it should not have on its borders (e.g., whitespace, periods, etc.), or when characters are missing.

Link Error: detects entity links that are not valid targets; for example, when considering Wikipedia as a target KB, detects links to redirect pages, disambiguation pages, etc.

Format Error: detects contradictions in the format, for example, when the label of an annotation does not match with the substring generated from the initial and final offset.

The extended version of NIFify further supports defining and executing rules capturing inter-dependencies between categories; for example, a *Pro-Form* annotation will always be labeled with *Anaphoric* reference. Such rules can be used to help the annotator complete or validate the current annotations:

Figure 12: 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.
```

Category Error: detects annotations belonging to incompatible categories, e.g., *Proper Form* and *Anaphoric*.

Some other validation services are provided in the case of specific categories, such as the following:

Overlapping error: detects when an annotation does not have the correct class in the OVERLAP dimension.

Pro-form error: detects when a *Pro-form* annotation links to an entity not mentioned elsewhere in the text.

The annotator may generate a list of violations for such cases; they are offered the choice to either leave the annotation as it is, or to make changes to the annotation, as they deem appropriate.

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 13 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. Fine-Grained Datasets

To put our fine-grained categorization scheme into practice, we now relabel a selection of existing EL datasets according to our categories. We currently focus on English texts, where we select three datasets for relabeling: KORE50 [13] (a concise but challenging dataset with highly ambiguous entities); ACE2004 [38] (a large and widely-used dataset for evaluating EL systems); and VoxEL [31] (a multilingual dataset with strict and relaxed annotations; we currently relabel the English text only). We first describe some of the guidelines used for annotation and the annotation process itself. We subsequently provide key statistics for the relabeled datasets.

²⁹https://github.com/henryrosalesmendez/NIFify_v4

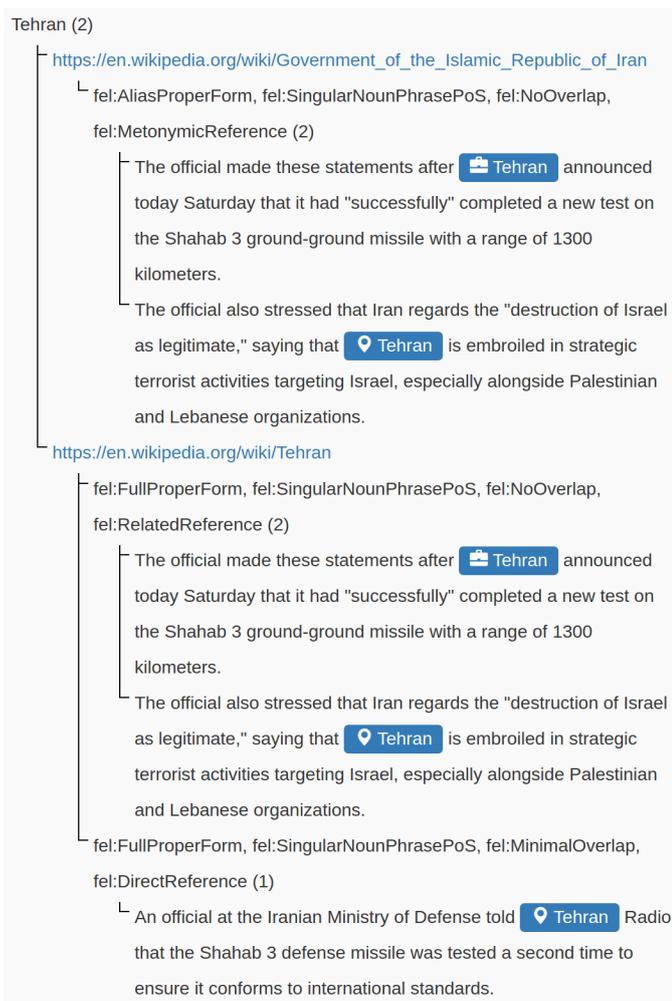


Figure 13: NIFify’s validation tree view for the mention “Tehran” in the ACE2004 dataset

6.1. Annotation Guidelines

As discussed in Section 2.3, 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 2.1, 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 8) 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 6.2 – 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. [51], 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 9.
 - 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)`³⁰, 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.

³⁰We argue that “it” does not refer to `wiki:Bill_(law)` here, but rather refers to *something that is a* `wiki:Bill_(law)`.

- 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.³¹ 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”, 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. [39], 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

³¹This decision was made after the questionnaire was conducted; for this reason, Table 2 uses an old convention for “.. the [Russian President].”; under the final convention, “.. [the Russian President].” would be considered the mention for `wiki:Vladimir_Putin`.

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.

6.2. 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 [51], 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 [57] of iterative refinement, involving not only the datasets themselves, but also the categories, the guidelines, and the tool used for annotation.

Specifically, the first 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 among all authors, leading to a refinement of the categories and guidelines. The other two authors then iteratively reviewed the annotations produced for these datasets, which were also validated in semi-automated fashion using the extended NIFify tool. With consensus reached on these two datasets, the first author then began an initial labeling of the larger ACE2004 dataset, highlighting further difficult cases that were discussed among all authors 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 authors and checked with the tool.³² The resulting datasets – as well as the previously discussed categories and guidelines – reflect the consensus of the three authors. 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

³²During 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.

Table 3: Statistics on the three relabeled datasets [11]

	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

definition [51]. In Table 3, 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.³³

7. Fine-Grained Evaluation

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 FREME (**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.

³³https://github.com/henryrosalesmendez/categorized_EMNLP_datasets

We then compute the micro Precision (**P**), micro Recall (**R**) and micro F_1 score (**F₁**) for these systems; in other words, we compute precision, recall and F_1 over a dataset comprised of the concatenation of our three datasets. Following the precedent of GERBIL [30], 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.³⁴

Table 4 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.³⁵ 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

³⁴The *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.

³⁵Counts 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.

Table 4: Results per category for Babelfly (strict/relaxed), TagME, DBpedia Spotlight, AIDA and FREDME on the unified dataset [11]

	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

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 Babelfly_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 Babelfly_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.

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 2 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 4 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 combination 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 14, 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: Babelfly_r and TagME have lower precision towards the end of the progression, but maintain a much higher recall; on the other hand, Babelfly_s, DBpedia Spotlight, AIDA and FREDME 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.

8. 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

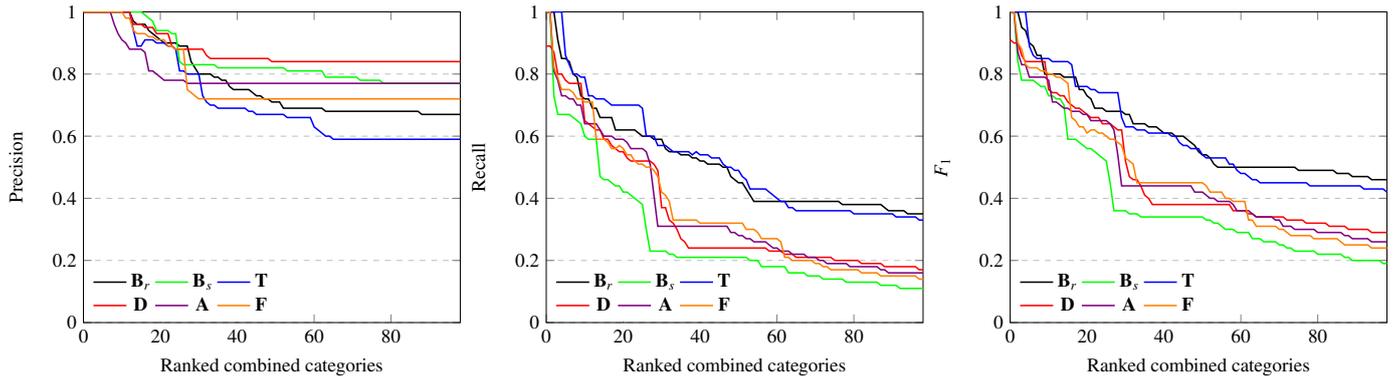


Figure 14: Cumulative best-first progression of precision, recall and F_1 scores for Babelfy (relaxed/strict), TagME, DBpedia Spotlight, AIDA and FREME for the unified dataset considering combinations of categories [11]

– one may argue that this part of a separate Coreference Resolution (CR) task [52]; 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 [50]. Conversely, some systems choose to incorporate CR [58, 59] and WSD [1] methods for the EL task.

In this section, we extend the five EL systems with CR and WSD methods to create initial versions 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).

8.1. 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 [60] 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 [61] 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).

NCR: Refers to the neural coreference resolution model [62], 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 5 and the results extended with NCR in Table 6 where we display only those categories (rows) where results changed versus the off-the-shelf results from Table 4; 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.

8.2. Adding Word Sense Disambiguation

Word Sense Disambiguation (WSD) refers to the task of disambiguating the sense of a word used in a particular context [50]. A typical target for WSD is to link words with WordNet [63], 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. To extend 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 algo-

Table 5: Changing results per category for Babelfy, TagME, DBpedia Spotlight, AIDA and FREDME 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

Table 6: 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

rithm [64], 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. [65], 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 [66] 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, while the results with *WSD-DIS* are shown in Table 8; 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* 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 (B_r and T), where B_r already incorporates WSD techniques [1]. Between both systems, WSD-NLTK tends to improve recall more than WSD-DIS, but WSD-DIS tends to maintain a higher precision.

8.3. 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 9, 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 10 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 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;

Table 7: Changing results per category for Babelify, TagME, DBpedia Spotlight, AIDA and FREME extended with WSD-NLTK 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	0.73	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	0.10	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

Table 8: Changing results per category for Babelify, TagME, DBpedia Spotlight, AIDA and FREME extended with 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	0.80	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

however, *Metonymic* references remain unsupported.

9. 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.

9.1. Fuzzy Framework

We propose a configurable evaluation framework based on *Fuzzy Set Theory* [67] 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

Table 9: Results per category for Babelfy, TagME, DBpedia Spotlight, AIDA and FREME 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 10: 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

setting; more concretely, we propose metrics that penalize systems more for missing annotations with higher degree.

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 miss-

ing 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.³⁶ 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.

³⁶Furthermore 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.

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, 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 :

$$\frac{\sum_{a \in S} \mu_{G^*}(a) - \mu_{G^*}(a_2)}{\sum_{a \in G} \mu_{G^*}(a)} < \frac{\sum_{a \in S} \mu_{G^*}(a) - \mu_{G^*}(a_1)}{\sum_{a \in G} \mu_{G^*}(a)} \quad (1)$$

We can simplify this inequality as follows:

$$\sum_{a \in S} \mu_{G^*}(a) - \mu_{G^*}(a_2) < \sum_{a \in S} \mu_{G^*}(a) - \mu_{G^*}(a_1) \quad (2)$$

$$-\mu_{G^*}(a_2) < -\mu_{G^*}(a_1) \quad (3)$$

$$\mu_{G^*}(a_1) < \mu_{G^*}(a_2) \quad (4)$$

Hence we see that inequality (1) holds if and only if the assumed inequality (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_1^* . \square

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.

9.2. Fuzzy Evaluation

We propose to generate a membership function for the annotations of our datasets based on the questionnaire results seen in Figure 8 and our categorization scheme. Specifically, we select combined categories that consistently score greater than 0.9 in Figure 8 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 15 for the off-the-shelf EL systems and for varying degrees of α .³⁷ Here we see that all systems performs 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.

10. Conclusions

We conclude the paper with a 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.

³⁷We 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.

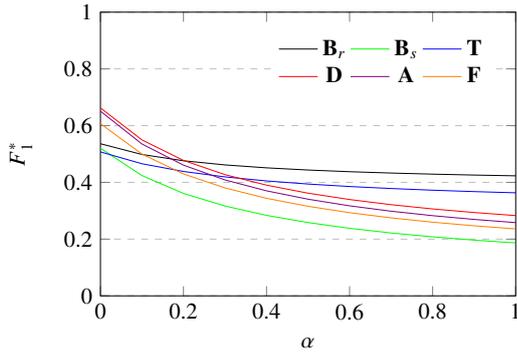


Figure 15: α -based fuzzy F_1 scores for off-the-shelf systems [11]

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.
- 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.
- With the goal of determining 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.
- We describe fuzzy-recall and fuzzy- F_1 measures that allow for assigning different weights to different annota-

tions, 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.

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.

- 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.
- 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, in parallel we further developed and extended 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 3). 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. [68, 69, 70]) and WSD systems (e.g., [71, 72]) in our experiments.

Outlook. 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, datasets³⁸, guidelines, metrics and results developed in this paper 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. [22] do discuss the issue). We thus identify this as an open challenge for EL research (and one that does not appear trivial).

Acknowledgments. The work of Henry Rosales-Méndez was supported by CONICYT-PCHA/Doctorado Nacional/2016-21160017. The work was also partially supported by the Millennium Institute for Foundational Research on Data (IMFD) and by Fondecyt Grant No. 1181896 and 1191604. We would also like to thank the respondents to our questionnaire, and the anonymous reviewers whose valuable feedback has helped to improve this paper.

Datasets and code. The NIFify system is available from https://github.com/henryrosalesmendez/NIFify_v3. The datasets and the code used for experiments are available from https://github.com/henryrosalesmendez/EL_exp.

³⁸The three datasets are available from https://github.com/henryrosalesmendez/categorized_EMNLP_datasets

References

- [1] A. Moro, A. Raganato, R. Navigli, Entity Linking meets Word Sense Disambiguation: a Unified Approach, *TACL 2* (2014) 231–244.
- [2] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia, *Semantic Web 6* (2015) 167–195.
- [3] K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: *SIGMOD*, 2008, pp. 1247–1250.
- [4] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledge-base, *Commun. ACM 57* (2014) 78–85.
- [5] T. Rebele, F. M. Suchanek, J. Hoffart, J. Biega, E. Kuzey, G. Weikum, YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames, in: *International Semantic Web Conference (ISWC)*, 2016, pp. 177–185.
- [6] J. L. Martinez-Rodriguez, A. Hogan, I. Lopez-Arevalo, Information Extraction meets the Semantic Web: A Survey, *Semantic Web Journal* (2019). (To appear: <http://www.semantic-web-journal.net/content/information-extraction-meets-semantic-web-survey>).
- [7] G. Wu, Y. He, X. Hu, Entity Linking: An Issue to Extract Corresponding Entity With Knowledge Base, *IEEE Access 6* (2018) 6220–6231.
- [8] P. Taufer, Named Entity Recognition and Linking, Master’s thesis, Univerzita Karlova, 2017.
- [9] M. Chang, B. P. Hsu, H. Ma, R. Loynd, K. Wang, E2E: An End-to-End Entity Linking System for Short and Noisy Text, in: *Workshop on Making Sense of Microposts*, volume 1141. CEUR-WS.org, 2014, pp. 62–63.
- [10] G. Luo, X. Huang, C. Lin, 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.
- [11] H. Rosales-Méndez, A. Hogan, 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.
- [12] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum, Robust disambiguation of named entities in text, in: *EMNLP*, 2011, pp. 782–792.
- [13] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, G. Weikum, KORE: keyphrase overlap relatedness for entity disambiguation, in: *CIKM*, 2012, pp. 545–554.
- [14] I. Hulpuş, N. Prangnawarat, 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.
- [15] T. Grütze, G. Kasneci, Z. Zuo, F. Naumann, CoHEEL: Coherent and efficient named entity linking through random walks, *J. Web Semant.* 37–38 (2016) 75–89.
- [16] A. Gangemi, V. Presutti, D. R. Recupero, A. G. Nuzzolese, F. Draicchio, M. Mongiovi, Semantic Web Machine Reading with FRED, *Semantic Web 8* (2017) 873–893.
- [17] R. Grishman, B. Sundheim, Message Understanding Conference- 6: A Brief History, in: *COLING*, 1996, pp. 466–471.
- [18] P. Ferragina, U. Scaiella, TAGME: on-the-fly annotation of short text fragments (by wikipedia entities), in: *CIKM*, 2010, pp. 1625–1628.
- [19] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, DBpedia Spotlight: shedding light on the web of documents, in: *I-SEMANTICS*, 2011, pp. 1–8.
- [20] R. Usbeck, A. N. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, A. Both, AGDISTIS - Agnostic Disambiguation of Named Entities Using Linked Open Data, in: *European Conference on Artificial Intelligence (ECAI)*, Springer, 2014, pp. 1113–1114.
- [21] M. Fleischman, E. H. Hovy, Fine Grained Classification of Named Entities, in: *COLING*, 2002.
- [22] X. Ling, S. Singh, D. S. Weld, Design challenges for entity linking, *TACL 3* (2015) 315–328.
- [23] F. Ilievski, G. Rizzo, M. van Erp, J. Plu, R. Troncy, Context-enhanced Adaptive Entity Linking, in: *International Conference on Language Resources and Evaluation (LREC)*, European Language Resources Association (ELRA), 2016.
- [24] G. Rizzo, 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.
- [25] J. Waitelonis, H. Jürges, H. Sack, Don’t compare apples to oranges: Extending GERBIL for a fine grained NEL evaluation, in: *SEMANTICS*, 2016, pp. 65–72.
- [26] M. van Erp, P. N. Mendes, H. Paulheim, F. Ilievski, J. Plu, G. Rizzo, 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.
- [27] K. Jha, M. Röder, 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.
- [28] H. Rosales-Méndez, B. Poblete, A. Hogan, What should Entity Linking link?, in: *Alberto Mendelzon Workshop (AMW)*, 2018.
- [29] S. Hellmann, J. Lehmann, S. Auer, M. Brümmer, Integrating NLP Using Linked Data, in: *International Semantic Web Conference (ISWC)*, Springer, 2013, pp. 98–113.
- [30] R. Usbeck, M. Röder, A. N. Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, L. Wesemann, GERBIL: General Entity Annotator Benchmarking Framework, in: *International Conference on World Wide Web (WWW)*, ACM, 2015, pp. 1133–1143.
- [31] H. Rosales-Méndez, A. Hogan, B. Poblete, VoxEL: A Benchmark Dataset for Multilingual Entity Linking, in: *International Semantic Web Conference (ISWC)*, Springer, 2018, pp. 170–186.
- [32] M. Brümmer, M. Dojchinovski, 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.
- [33] A. Moro, R. Navigli, SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking, in: *Semantic Evaluation Workshop (SemEval@NAACL-HLT)*, ACL, 2015, pp. 288–297.
- [34] A. Minard, M. Speranza, R. Urizar, B. Altuna, M. van Erp, A. Schoen, C. van Son, MEANTIME, the NewsReader Multilingual Event and Time Corpus, in: *LREC*, 2016.
- [35] 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.
- [36] D. N. Milne, I. H. Witten, Learning to link with Wikipedia, in: *ACM Conference on Information and Knowledge Management (CIKM)*, ACM, 2008, pp. 509–518.
- [37] S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti, Collective annotation of Wikipedia entities in web text, in: *SIGKDD*, 2009, pp. 457–466.
- [38] L. Ratinov, D. Roth, D. Downey, M. Anderson, Local and Global Algorithms for Disambiguation to Wikipedia, in: *ACL*, 2011, pp. 1375–1384.
- [39] M. Röder, R. Usbeck, S. Hellmann, D. Gerber, 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.
- [40] J. Waitelonis, C. Exeler, H. Sack, Linked Data enabled generalized vector space model to improve document retrieval, in: *NLP & DBpedia @ ISWC*, 2015.
- [41] C. Brando, F. Frontini, J. Ganascia, REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets, *CSIMQ 7* (2016) 60–80.
- [42] S. Cucerzan, Large-scale named entity disambiguation based on Wikipedia data, *EMNLP-CoNLL* (2007) 708.
- [43] S. Farrar, D. T. Langendoen, A linguistic ontology for the semantic web, *GLOT international 7* (2003) 97–100.
- [44] J. P. McCrae, D. Spohr, P. Cimiano, Linking lexical resources and ontologies on the semantic web with lemon, in: *ESWC*, 2011, pp. 245–259.
- [45] J. P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, P. Cimiano, The OntoLex-Lemon Model: Development and Applications, in: *eLex*, 2017, pp. 19–21.
- [46] G. de Melo, G. Weikum, Language as a Foundation of the Semantic Web, in: *Proceedings of the Poster and Demonstration Session at ISWC*, 2008.

- [47] G. de Melo, Lexvo.org: Language-related information for the Linguistic Linked Data cloud, *Semantic Web* 6 (2015) 393–400.
- [48] H. Rosales-Méndez, A. Hogan, B. Poblete, NIFify: Towards Better Quality Entity Linking Datasets, in: *WWW Companion Volume*, 2019, pp. 815–818.
- [49] O. Etzioni, M. J. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates, Unsupervised named-entity extraction from the Web: An experimental study, *Artif. Intell.* 165 (2005) 91–134.
- [50] R. Navigli, Word sense disambiguation: A survey, *ACM Comput. Surv.* 41 (2009) 10:1–10:69.
- [51] S. Guo, M. Chang, 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.
- [52] R. Sukthankar, S. Poria, E. Cambria, R. Thirunavukarasu, Anaphora and coreference resolution: A review, *CoRR* abs/1805.11824 (2018).
- [53] P. F. Strawson, On referring, *Mind* 59 (1950) 320–344.
- [54] T. Baker, P. Vandenbussche, B. Vatant, Requirements for vocabulary preservation and governance, *Library Hi Tech* 31 (2013) 657–668.
- [55] A. Miles, S. Bechhofer, SKOS Simple Knowledge Organization System Reference, W3C Recommendation, 2009. <https://www.w3.org/2004/02/skos/>.
- [56] P. Vandenbussche, G. Atemezeng, M. Poveda-Villalón, B. Vatant, Linked open vocabularies (LOV): A gateway to reusable semantic vocabularies on the web, *Semantic Web* 8 (2017) 437–452.
- [57] K. Beck, M. Beedle, A. Van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries, et al., *Manifesto for Agile Software Development* (2001).
- [58] H. Hajishirzi, L. Zilles, D. S. Weld, L. S. Zettlemoyer, Joint coreference resolution and named-entity linking with multi-pass sieves, in: *EMNLP*, 2013, pp. 289–299.
- [59] G. Durrett, D. Klein, A joint model for entity analysis: Coreference, typing, and linking, *TACL* 2 (2014) 477–490.
- [60] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, in: *ACL*, 2014, pp. 55–60.
- [61] K. Clark, 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.
- [62] K. Clark, C. D. Manning, Deep reinforcement learning for mention-ranking coreference models, in: *EMNLP*, 2016, pp. 2256–2262.
- [63] G. A. Miller, C. Fellbaum, WordNet then and now, *Language Resources and Evaluation (LRE)* 41 (2007) 209–214.
- [64] 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.
- [65] L. Vial, B. Lecouteux, 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.
- [66] T. Miller, 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.
- [67] L. A. Zadeh, Fuzzy sets, *Information and control* 8 (1965) 338–353.
- [68] K. Lee, L. He, 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.
- [69] J. Plu, R. Prokofyev, A. Tonon, P. Cudré-Mauroux, D. E. Difallah, R. Troncy, G. Rizzo, Sanaphor++: Combining Deep Neural Networks with Semantics for Coreference Resolution, in: *International Conference on Language Resources and Evaluation (LREC)*, 2018.
- [70] M. Joshi, O. Levy, L. Zettlemoyer, D. S. Weld, BERT for Coreference Resolution: Baselines and Analysis, in: *Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 5802–5807.
- [71] L. Huang, C. Sun, X. Qiu, X. Huang, GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge, in: *Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3507–3512.
- [72] D. Loureiro, 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.