Fine-Grained Evaluation for Entity Linking

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

Abstract

The Entity Linking (EL) task identifies entity mentions in a text corpus and associates them with an unambiguous identifier in a Knowledge Base. While much work has been done on the topic, we first present the results of a survey that reveal a lack of consensus in the community regarding what forms of mentions in a text and what forms of links the EL task should consider. We argue that no one definition of the Entity Linking task fits all, and rather propose a fine-grained categorization of different types of entity mentions and links. We then re-annotate three EL benchmark datasets – ACE2004, KORE50, and VoxEL – with respect to these categories. We propose a fuzzy recall metric to address the lack of consensus and conclude with fine-grained evaluation results comparing a selection of online EL systems.

1 Introduction

Entity Linking (EL) is an Information Extraction task whose goal is to identify mentions of entities in a text and to link each mention to an unambiguous identifier in a Knowledge Base (KB) such as Wikipedia, BabelNet (Moro et al., 2014), DBpedia (Lehmann et al., 2015), Freebase (Bollacker et al., 2008), Wikidata (Vrandečić and Krötzsch, 2014) or YAGO (Rebele et al., 2016) (among others). The results of this task offer a practical bridge between unstructured text and structured KBs, where EL has applications for semantic search, document classification, relation extraction, and more besides (Wu et al., 2018).

While a broad number of EL techniques and systems have been proposed in recent years (Wu et al., 2018), a number of authors have noted that there is a lack of consensus on the fundamental question of what kinds of mentions in a text an EL system should link to which identifiers in the KB (Ling et al., 2015; Waitelonis et al., 2016; Jha et al., 2017; Rosales-Méndez et al., 2018b).

A closely related task to EL is that of Named Entity Recognition (NER), which identifies mentions of (named) entities in a task, but without linking the mention with a KB identifier. The types of entities that the NER task should target were defined at the Message Understanding Conference 6 (MUC-6) (Grishman and Sundheim, 1996), specifically entities corresponding to types such as Person, Organization, Place and other Numerical/Temporal expressions. While this provided a consensus for evaluating NER systems, some authors noted that such a categorization is coarse-grained and proposed finer-grained classification mechanisms (Fleischman and Hovy, 2002).

In the context of EL, target Knowledge Bases will often contain entities from a wide variety of classes, including Movies, Products, Events, Laws, etc., not considered by the traditional NER definitions; a dataset such as Wikidata has around 50,000 entity classes, for example. As a result, class-based definitions of entities are restrictive. Hence some authors have proposed more general definitions for the EL task: as Ling et al. (2015) note, while some authors follow traditional NER definitions, others propose a looser definition that any KB identifier (e.g., Wikipedia article URL) can be the target for a link; they further note that within these definitions there is a lack of guidelines with respect to how EL datasets should be labeled and what sorts of mentions and links EL systems should (ideally) offer.

This ambiguity complicates research and applications relating to EL, as highlighted previously by various authors (Ling et al., 2015; Jha et al., 2017; Rosales-Méndez et al., 2018b). Figure 1 shows the results of a selection of popular online EL systems: Babelfy (Moro et al., 2014), DBpedia Spotlight (Mendes et al., 2011),

FRED (Gangemi et al., 2017) and TagME (Ferragina and Scaiella, 2010). We see significant differences in the annotations provided, which we argue are due not only to differing precision/recall of the systems, but also to differing policies on issues such as overlapping entities (should "Michael Jackson" be annotated within the documentary title), common/named entities (should "interview" be linked to the corresponding Wikipedia article `wiki:Interview`), and more besides. With such varying perspectives on the EL task, it is not clear how we should define gold standards that offer a fair comparison of tools (Ling et al., 2015).

The standard approach to tackle 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 labeling an EL dataset or performing evaluation. However, the appropriate policy may depend on the particular application, setting, etc. In this paper we pursue an alternative approach, which is to embrace different perspectives of the EL task, proposing a fine-grained categorization of different types of EL mentions and links, and then re-annotating three existing datasets with respect to these categories, allowing us to compare the performance of EL tools within the different categories. Specifically, our contributions are as follows (indicating also the relevant section):

§ 2 We design and present the results of a questionnaire addressed to authors of EL papers intended to understand the consensus (or lack thereof) regarding the goals of the EL task.

§ 3 We argue that no one definition of an entity mention/link fits all, and hence propose a fine-grained categorization scheme for the EL task covering details regarding base form, part of speech, overlap, and reference type.

§ 4 We relabel three existing EL datasets – ACE2004 (subset), KORE50 and VoxEL – per our novel categorization scheme, extending the set of annotations as appropriate.

§ 5 We conduct a fine-grained evaluation of the performance of five EL systems with respect to individual categories of EL annotations.

§ 6 Addressing the lack of consensus, we propose a fuzzy recall and F_1 measure based on a configurable membership function, presenting results for the five EL systems.

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: Output annotations made by four different systems – Babelfy (b), TagME (t), DBpedia Spotlight (d) and FRED (f) – on the same input text.

§ 7 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 Questionnaire: Lack of Consensus

To understand what consensus (or lack thereof) exists within the EL community regarding the EL task, we designed a concise questionnaire with two sentences for which we proposed a variety of EL annotations, providing the text, the annotated text mentions, and proposed links to the respective Wikipedia articles. The two sentences – shown in Figure 2 (with a summary of results that will be described presently) – were designed to exemplify the types of design choices that vary from author to author (Ling et al., 2015); specifically, we target the following questions:

1. *Entity types*: should types not typically considered under MUC-6 definitions (other than as MISC) be linked (e.g., linking Living with Michael Jackson to the corresponding Wikipedia article)?

2. *Overlapping mentions*: should mentions inside other mentions be annotated (e.g., Michael Jackson)?

3. *Common entities*: should common nouns be annotated (e.g., documentary)?

4. *Parts of speech*: should only nouns be annotated or should other parts of speech also be linked (e.g., reports or white)?

5. *Mention types*: should complex types of mentions, such as pronouns (e.g., he) or descriptive noun phrases indicating named entities, be annotated (e.g., linking he and his four siblings to `wiki:The_Jackson_5`)?

In an [interview]$^{.19}$ with [Martin Bashir]^{1} for the [2003]$^{.28}$ [documentary]$^{.28}$ [Living with {Michael Jackson}$^{.75}$]$^{.97}$, the [{King}$^{.08}$ of {Pop}$^{.33}$]$^{.94}$ [recalled]$^{.06}$ that [Joe]^{1} often [sat]$^{.08}$ with a [white]$^{.11}$ [belt]$^{.14}$ at [hand]$^{.14}$ as [{he}$^{.56}$ and {his}$^{.39}$ {four}$^{.08}$ {siblings}$^{.14}$]$^{.50}$ [rehearsed]$^{.08}$.

[Russian]$^{.61}$ [daily]$^{.14}$ [Kommersant]$^{.97}$ [reports]$^{.06}$ that [Moscow]$^{.94}$ will [supply]$^{.06}$ the [Greeks]$^{.94}$ with [gas]$^{.36}$ at [{rock}^{0} bottom {prices}$^{.19}$]$^{.28}$ as [Tsipras]$^{.92}$ [prepares]$^{.03}$ to [meet]$^{.06}$ the [{Russian}$^{.53}$ {President}$^{.12}$]$^{.97}$.

Figure 2: The sentences included in the questionnaire and the ratio of respondents who suggested to annotate the mentions. Multiple links were proposed for the mentions underlined (see Table 1).

6. *Link types*: should mentions link to the explicitly named entity (e.g., linking Moscow to `wiki:Moscow`), or should complex forms of reference such as *meronymy* (e.g., linking Moscow to `wiki:Government_of_Russia`), *hypernymy* (e.g., linking daily to `wiki:Newspaper` as the closest entity present in the KB, or linking Russian President to `wiki:Vladimir_Putin`), or *metaphor* (e.g., linking King to `wiki:King`) be considered?

For each sentence, respondents were asked to select the mentions and links that they consider an EL system should *ideally* provide; specifically, they were presented an optional multiple choice question for each mention: the option to select one (or more, in underlined cases) suggested links, and the option to not annotate the mention. In order to address this survey to the EL research community, we extracted the emails of all papers referenced in the recent survey by Wu et al. (2018) that are directly related to EL. We sent the questionnaire to 321 researchers, of which 232 requests were delivered successfully. We received a total of 36 responses. Aggregated responses are available online[1], where in Figure 2 we provide a summary of results, indicating in superscript the ratio of respondents who agreed to some link being provided for the given mention.

First we see that the only mentions that all 36 respondents agreed should be linked are Martin Bashir and Joe, which are traditional MUC types, non-overlapping, named entities, with direct mentions and links; on the other hand, there was also consensus that `rock` should not be linked. Otherwise, all other mentions showed some level of disagreement. Regarding our questions:

1. *Entity types*: per Living with Michael Jackson (0.97), the vast majority of respondents do not restrict to non-MISC MUC types.

2. *Overlapping mentions*: per Michael Jackson (0.75), most respondents also allow mentions contained in other mentions.

3. *Common entities*: most respondents do not consider common entities, with the highest response appearing for documentary (0.28).

4. *Parts of speech*: the consensus was mostly that EL should focus on nouns, with his (0.39) and white (0.11) being the most popularly selected non-nouns.

5. *Mention types*: here there was a notable split, with he (0.56) and he and his four siblings (0.5)[2] being selected for annotation by roughly half of the respondents.

6. *Link types*: for this we introduce Table 1, where we see how respondents preferred different types of reference (respondents could select multiple options); from this we conclude that although respondents preferred to use a country directly to represent nationality, they also preferred to resolve complex types of reference, such as the meronymic use of Moscow to refer to the government rather than the city, and the use of Putin's title to refer to him rather than the title itself.

So who is correct? We argue that there is no "correct" answer here. Common entities may, for example, rather be considered the target of a separate Word Sense Disambiguation task (Navigli, 2009), while pro-forms may be considered the target of a separate Coreference/Anaphora Resolution task (Sukthanker et al., 2018); these are matters of convention. In more practical terms, the

[1] `https://users.dcc.uchile.cl/~hrosales/questionnaire`

[2] One respondent noted that it was not certain that this referred to The Jackson 5.

Table 1: Links for mentions with multiple choices in Figure 2 and the ratio of respondents selecting that link

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

importance of each individual EL annotation may vary from application to application; for example, for relation extraction, it may be important to identify all repeated mentions of an entity (as each such mention may lie within a distinct relation), while for semantic search having repeated mentions may be less critical (a subset of mentions may suffice to know the document is relevant for a given entity). We propose that no one definition of the EL task fits all, and rather propose a categorization that covers different perspectives.

3 Categorization Scheme

Inspired by the results of the questionnaire and related discussion by Ling et al. (2015), among other authors, in Figure 3 we propose a categorization scheme aiming to capture and organize these diverse perspectives on the EL task. This scheme makes explicit the types of *annotations* – mention–link pairs – that can be considered when annotating EL datasets, or when developing, evaluating and applying EL systems. The categorization scheme has four dimensions: *Base Form*, *Part of Speech*, *Overlap* and *Reference*. We propose that each EL annotation be labeled with precisely one leaf-node from each dimension. Our objective with this categorization is not to imply that all types of annotations *should* be considered as part of the EL task, but rather to map out the types of annotations that *could* be considered for EL.

3.1 Base Form

The *Base Form* considers the type of mention: is it a name, a common noun, a number, a date, etc.

We separate *Proper Noun* into more specific categories that deal with the difference between mentions and KB labels: *Full Name*, *Short Name*, *Extended Name* and *Alias*. The first three should be tagged when the mentions are equal (Michael Jackson), shorter (Jackson) or longer (Michael Joseph Jackson), respectively, than the primary label of their corresponding KB-entity (wiki:Michael_Jackson). On the other hand, *Alias* is used for mentions that vary from the primary label of the KB (King of Pop).

Numeric/Temporal are mentions that refer to a number or a given temporal concept (e.g., 1, 1900, first, October, July 07, Tuesday, 2018/01/32, etc.). Such mentions were included in the MUC-6 definition and some such expressions have corresponding Wikipedia articles. The next category is *Pro-form*, which includes any pronoun (*he*, *his*, etc.) referring to a named entity elsewhere. The final category is *Common Form*, which refers to a common word, such as documentary, sat, etc., or a noun-phrase referring to a named entity with a common head term, e.g., he and his four siblings.

3.2 Part of Speech

Part of Speech denotes the grammatical function of a word in a sentence, where we include four high-level classes for which we have found instances that could be linked to the Wikipedia KB (without referring to their own syntactic form). As we have already discussed, the most common target for EL is *Noun Phrases*; we divide these into *Singular* and *Plural* for additional granularity. However, we also include *Verb*s (e.g., divorcing, metastasized), *Adjective*s (e.g., anaerobic, French) and *Adverb*s (e.g., polynomially, Socratically) for linking to the Knowledge Base.

3.3 Overlap

The *Overlap* dimension, as its name suggests, captures the nature of overlap between mentions. As an example, for the mention New York City Police Museum, the mention New York would have *Minimal* overlap (assuming York is not incorrectly identified), New York City Police would have *Intermediate* overlap, and New York City Police Museum would have *Maximal* overlap.

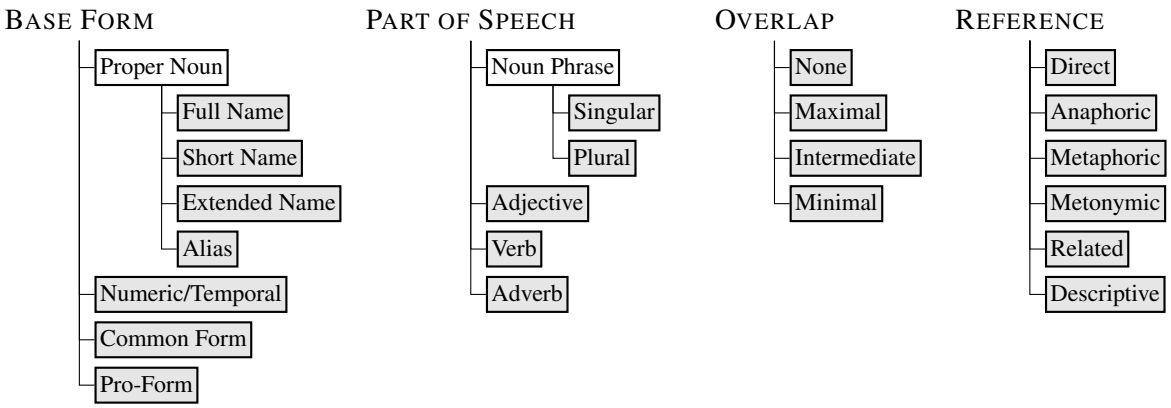

Figure 3: EL categorization scheme with concrete alternatives (leaf-nodes) shaded for each dimension

3.4 Reference

The *Reference* dimension considers the relation between the mention and the linked entity. The topic of *reference* is a complex one that has been the subject of much attention over many decades (Strawson, 1950); here we propose a pragmatic but comprehensive set of options.

The *Direct* category considers a direct explicit mention of an entity based on a known KB label/alias, such as M. Jackson or the King of Pop for `wiki:Michael_Jackson` or divorcing for `wiki:Divorce`. *Anaphoric* denotes a reference to an antecedent (or postcedent) for pro-forms such as he or her that are coreferent with a named entity. *Metaphoric* captures figurative references to entities whose characteristics are referred to, such as He is a pool [shark] linking to `wiki:Shark`. *Metonymic* indicates reference by common association, such as using Moscow to refer to `wiki:Government_of_Russia` or using Peru to refer to `wiki:Peru_national_football_team`. *Related* is used when only a near-synonym, hypernym or hyponym is available for a mention in the KB, such as the Russian [daily] being linked to `wiki:Newspaper`.[3] Finally, *Descriptive* is used for (non-pro, non-proper) referring noun-phrases, such as his father referring to `wiki:Joe_Jackson`; unlike similar *Anaphoric* references, *Descriptive* references do not necessarily rely on an antecedent to be disambiguated, as per the case of Hendix's band referring to `wiki:The_Jimi_Hendrix_Experience` or Fiji's capital referring to `wiki:Suva` without requiring further context to disambiguate.

[3] Wikipedia will often redirect from a term to a related article; for example, `wiki:Daily_newspaper` currently redirects to `wiki:Newspaper`.

4 Fine-Grained Datasets

A wide variety of EL datasets have been proposed (e.g., ACE2004 (Ratinov et al., 2011), KORE50 (Hoffart et al., 2012), MEANTIME (Minard et al., 2016), DBpedia Spotlight (Mendes et al., 2011), VoxEL (Rosales-Méndez et al., 2018a)) to support EL quality measurement. However, often the guidelines followed in the annotation process are left implicit, such as the inclusion/exclusion of overlapping mentions, common entities, etc. Furthermore, different types of entities are not distinguished.

To put our categorization into practice, we select three existing datasets – KORE50, ACE2004, and VoxEL – and categorize each annotation. We further add novel annotations not considered in the original labeling process with the goal of capturing as many potential (categorized) links to Wikipedia articles as possible. The process of annotation was done manually by three authors (with the help of the NIFify tool (Rosales-Méndez et al., 2019) for labeling overlaps, parts of speech, as well as validation). The annotation process was iterative. The first author began an initial annotation based on a strict and relaxed notion of "entity", with strict referring to classical definitions of entities, and relaxed referring to any entity mention linkable with Wikipedia (following the literature). This process raised ambiguities regarding metonymic references, descriptive noun phrases, etc. Hence the authors defined fine-grained categories to address ambiguous cases and designed the questionnaire to better understand the community consensus. The first author relabeled the data per the fine-grained categorization, with semi-automated verification. The other authors then revised these annotations in detail; there were significant dif-

ferences, leading to further discussion and refinements in the categorization. Ambiguities and disagreements were iteratively resolved by the authors through discussion and modification of the categories. The resulting datasets reflect the consensus of the authors. The categorization scheme shown in Figure 3 is also a result of this process, where we iteratively extended and refined these categories as we encountered specific cases in these datasets; as a result, these categories suffice to cover all cases of possible mentions of Wikipedia entities in the datasets.

The labeling process was extremely time consuming (taking place over six months) due to the large number of annotations generated, particularly for common-form mentions; for this reason, in the case of ACE2004, we only annotate 20 documents (from a total of 57) wherein, for example, the number of annotations increases from 108 in the original data to 3,351 in our fine-grained version[4]. Per MUC-6, we also include emerging/not-in-lexicon entities and entity types in our annotations. Additionally, we label some mentions with multiple alternatives (per Table 1).

In Table 2 we summarize details of the reconstructed version of these three datasets, including the number of documents, sentences and annotations. Except for the case of ACE2004, our re-annotated datasets maintain the same text collection as in the original version. For each category, we also show the number of annotations tagged with it. The most frequent annotations are those that belong to *Common Form*, *Singular Noun*, *Non-Overlapping* and *Direct* categories.

In the final iteration, we perform validation checking for erroneous links avoiding invalid, redirect and disambiguation pages; erroneous categorizations, e.g., multiple tags in the same category; erroneous mentions, such as trailing spaces; etc. During this process, we also encountered and resolved numerous issues with the original datasets; of note, in ACE2004 we found various spelling errors of entity names, e.g., Stewart Talbot as a misspelling of Strobe Talbott, Coral Islands as a misspelling of Kuril Islands, etc.

5 Fine-Grained Evaluation

Our fine-grained datasets allow us to understand the performance of EL systems in more detail regarding different types of EL annotations. In Table 3, we present the Precision (**P**), Recall (**R**) and F_1 score ($\mathbf{F}_1$) for five popular EL systems with online APIs: Babelfy (**B**), TagME (**T**), DBpedia Spotlight (**D**), AIDA (**A**) and FREME (**F**). In the case of Babelfy, we consider both settings: *strict*, which excludes common entities ($\mathbf{B}_s$); and *relaxed*, which includes common entities ($\mathbf{B}_r$). Results are shown for subsets of annotations corresponding to a particular category (A), where we also present the number of unique mentions for annotations of that category ($|A|$). Recall that our gold standard may have multiple annotations for a single mention in the gold standard, listing different possible links for an individual mention (see, e.g., Table 1); on the other hand, evaluated systems predict a single link for each mention. We thus evaluate annotations on a mention-by-mention basis.[5] We consider a predicted mention to be a *true positive* if it is included in A and the predicted link

Table 2: Content of relabeled datasets

	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

[4] https://github.com/henryrosalesmendez/categorized_EMNLP_datasets

[5] While a mention can only appear in one PART-OF-SPEECH and OVERLAP category, it can appear in multiple BASE FORM or REFERENCE categories for alternative links.

Table 3: Results for Babelfy (strict/relaxed), TagME, DBpedia Spotlight, AIDA and FREME on the unified dataset.

		B_s			B_r			T			D			A			F		
	$\|A\|$	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	0.3	0.37	0.5	0.36	0.42	0.39	0.28	0.33
Extended Mention	9	1.0	0.56	0.71	0.83	0.56	0.67	1.0	0.44	0.62	1.0	0.44	0.62	1.0	0.44	0.62	0.8	0.44	0.57
Alias	112	0.56	0.16	0.25	0.33	0.21	0.25	0.52	0.32	0.4	0.67	0.38	0.48	0.6	0.29	0.4	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	0.0	0.0	0	0.0	0.0	0	0.0	0.0
Common Form	2,452	0.21	0.0	0.01	0.66	0.33	0.44	0.49	0.28	0.35	0.88	0.04	0.08	0.43	0.0	0.0	0.56	0.0	0.01
Pro-form	153	0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	0	0.0	0.0	0	0.0	0.0
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	0.2	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	0.7	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	0.6	0.14	0.22
Verb	334	0	0.0	0.0	0.86	0.02	0.04	0.37	0.17	0.23	1.0	0.0	0.01	0	0.0	0.0	0	0.0	0.0
Adverb	12	0	0.0	0.0	0	0.0	0.0	0.56	0.42	0.48	0	0.0	0.0	0	0.0	0.0	0	0.0	0.0
Non-Overlapping	2,871	0.75	0.12	0.2	0.67	0.33	0.45	0.58	0.38	0.46	0.84	0.19	0.32	0.78	0.19	0.3	0.71	0.17	0.27
Maximal Overlap	464	0.87	0.17	0.29	0.85	0.36	0.5	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	0.3	0.71	0.52	0.6	0.57	0.3	0.39	0.56	0.13	0.21	0.54	0.1	0.17	0.78	0.1	0.17
Minimal Overlap	825	0.82	0.04	0.09	0.61	0.37	0.46	0.5	0.15	0.23	0.8	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	0.3	0.7	0.17	0.27
Anaphoric	153	0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	0	0.0	0.0	0	0.0	0.0
Metaphoric	69	0.0	0.0	0.0	0.57	0.29	0.38	0.43	0.35	0.38	0.91	0.14	0.25	0	0.0	0.0	0	0.0	0.0
Metonymic	73	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Related	829	0.64	0.06	0.11	0.36	0.14	0.2	0.39	0.24	0.3	0.76	0.1	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	0.6	0.02	0.03	0	0.0	0.0	0.6	0.02	0.03
All	4,231	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

corresponds to any of the possible links given by A for that mention. We consider a predicted mention to be a *false positive* if the mention is given by A, but the predicted link is not given by A for that mention; in other words, when restricting A by category, system annotations on mentions outside of A are ignored.[6] Finally we consider a mention to be a *false negative* if it is given by A but either the mention is not predicted by the system or the system predicts a link not given by A for tht mention. We also show the overall result in the final row considering the full gold standard. The results consider a unified dataset that concatenates all three datasets. Better results (closer to 1) are shaded darker to aid with visual comparison.

Comparing the different types of annotations, on a high-level, we can see that recall, in particular, varies widely across categories; unsurprisingly perhaps, systems, in general, exhibit higher recall for proper nouns (particularly full and extended names) with direct references. Looking at categories with poor results, we see that none of the evaluated systems consider anaphoric references, perhaps considered as a task distinct from EL. Interestingly, no system captures metonymic references, though as previously seen the results of our questionnaire (see Table 1) indicate that respondents prefer such types of links over their literal counterparts (e.g., linking Moscow in the given context with `wiki:Government_of_Russia` rather than `wiki:Moscow`). Comparing systems, Babelfy (relaxed) and TagME achieve much higher recall for common forms than the other systems: we attribute this to these systems making the design choice to additionally support common entities.

While the previous results consider dimensions independently, there are $7 \times 5 \times 4 \times 6 = 840$ possible combinations across the four dimensions; not all of these can occur (for example, a *Pro-form* mention requires *Anaphoric* reference). In the unified dataset, we found 123 combinations to have at least one annotation. In order to understand in more detail how the systems perform for annotations in combined categories, for the six system configurations, Figure 4 presents a best-first accumulative progression of Precision, Recall and F_1: we start with the combined category in which each system performs best ($x = 1$), adding tags in the next-best combined category until all annotations in the gold standard are considered. We see that although precision remains relatively high throughout the gold standard, recall drops considerably as combined categories on which there is less consensus are added. The recall and F_1 measures, in particular, present a clear division in the systems: Babelfy (relaxed) and TagME maintain a higher recall as more combined categories are considered due to their inclusion of common entities (but suffer from reduced precision as a result).

[6] If a system predicts a link present in the gold standard for the mention but with a different category than tested, it will thus be considered a false positive for the tested category.

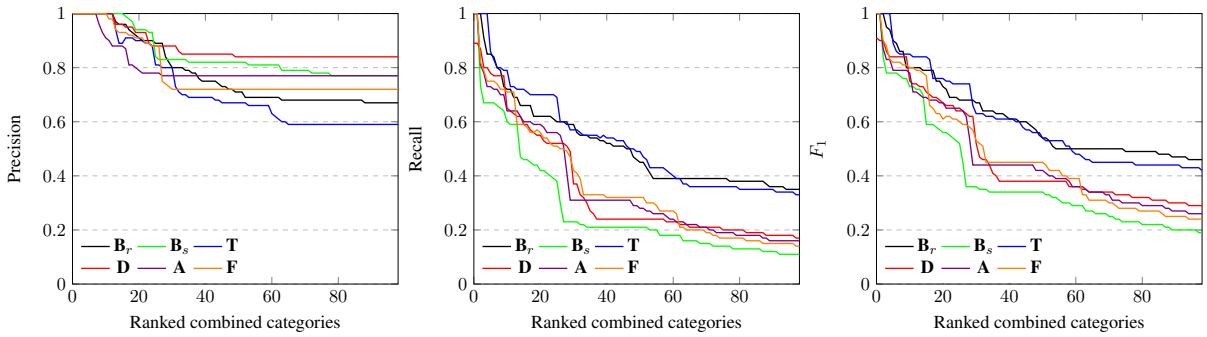

Figure 4: Cumulative results for Babelfy (relaxed/strict), TagME, DBpedia Spotlight, AIDA and FREME for the unified dataset over ranked combinations of categories

6 Fuzzy Recall and F_1 Measures

While the previous results provide insights into different design choices taken by different systems with respect to different types of EL annotations, still, per Table 1, the results are perhaps *too* fine-grained. For the purposes of comparing the performance of systems, we may rather prefer a measure that aggregates performance across different categories. However, this presents a conceptual problem: for some applications or definitions of the EL task, certain categories may be considered more important than others. For instance, should we penalize a system equally for missing the annotations gas and Tspiras in the second sentence of Figure 2? In most EL settings, the former is arguably less important than the latter, but nonetheless it may depend on the setting.

Reflecting the lack of consensus for the EL task, we rather propose a measure inspired by Fuzzy Set Theory (Zadeh, 1965): given a universe of elements U and a particular element $x \in U$, rather than considering traditional *crisp* sets A with binary membership (where $x \in A$ or $x \notin A$), a fuzzy set A^* is associated with a membership function $\mu_{A^*} : U \to [0, 1]$, which denotes the degree to which an element x is a member of A^* (given by $\mu_{A^*}(x)$). Given that a crisp set A can be defined as a fuzzy set with membership function $\mu_A : U \to \{0, 1\}$, fuzzy sets are a generalization of crisp sets. Intuitively, we can then consider a gold standard with a fuzzy set of annotations A^*, where (e.g.) annotations forming part of the core consensus of EL have a higher membership degree than those for which consensus does not exist; different membership degrees can also be applied for evaluation in different application settings.

Formally, an annotation is a triple $a = (o, o', l)$, where o and o' denotes the start and end offset of a mention in a text ($o < o'$), and l denotes a link (a KB identifier or a not-in-lexicon string). For a given text, a gold standard G is a set of annotations, as is the result of a system 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 this case, however, while we still consider S to be a crisp set, we allow a fuzzy version of the gold standard G^* with $\mu_{G^*} : G \to [0, 1]$.[7] In practice, for a given annotation $a \in G$, we propose that $\mu_{G^*}(a)$ is a function of the categorization for a; for example, with reference to Figure 2, we may consider that common forms have a lower degree of membership than proper forms. We are left to define Precision, Recall and F_1 measures for S with respect to G^*.

For a given system result S, gold standard G and its fuzzy version G^*, 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. On the other hand, a gold standard annotation *not* proposed by the system may be due to different design choices; we hence propose to use a fuzzy recall measure with respect to G^*, namely $R^* = \frac{\sum_{a \in S} \mu_{G^*}(a)}{\sum_{a \in G} \mu_{G^*}(a)}$, thus applying different costs for missing annotations (type II errors) depending on the annotation in question. We then define the fuzzy F_1 measure in the natural way: $F_1^* = \frac{2 \cdot P \cdot R^*}{P + R^*}$. The following properties can be verified for R^* and F_1^*:

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

[7] For annotations $a \notin G$, we assume $\mu_{G^*}(a) = 0$.

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

- PROP3: missing annotations with higher membership degree are penalized more in R^* and F_1^* than those with lower degree.

The definition of the membership function μ_{G^*} is then dependent on the setting. Here we define an instance of μ_{G^*} based on the questionnaire results shown in Figure 2. Specifically, we consider annotations with *Proper Forms*, *Noun Phrases*, *No Overlap* and *Direct Reference* – which consistently score greater than 0.9 in Figure 2 – to have a membership degree of 1; we consider these to be *strict* annotations. We assign all other annotations – which we call *relaxed* annotations – a constant membership degree of α, where higher values of α place more importance on achieving relaxed annotations; when $\alpha = 0$, relaxed false negatives are not punished; when $\alpha = 1$, both strict and relaxed false negatives are weighted equally.

Finally, the gold standard may offer multiple alternative links for a mention while the evaluated systems predict one link per mention. We apply the same procedure outlined previously: checking for each mention that the predicted link matches one of the alternatives in the gold standard. In the case of R^*, the membership score 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^*.

The F_1^* results are shown in Figure 5, where we again can distinguish the systems that link common entities – Babelfy (relaxed) and TagME – from those that do not; the former group of systems performs worse for stricter definitions of EL annotations, but outperform other systems as the definition is relaxed.[8] The raw data for these experiments can be found online.[9]

7 Conclusions

We have (i) presented the results of a questionnaire that assesses consensus on the goals of the EL task, (ii) proposed a fine-grained categorization scheme for EL annotations, (iii) comprehensively (re-)labeled three datasets accordingly to this scheme, (iv) presented results for five EL systems with respect to annotations in different categories, and (v) proposed fuzzy recall/F_1 measures to address the lack of consensus, presenting results for the five systems on a strict/relaxed spectrum.

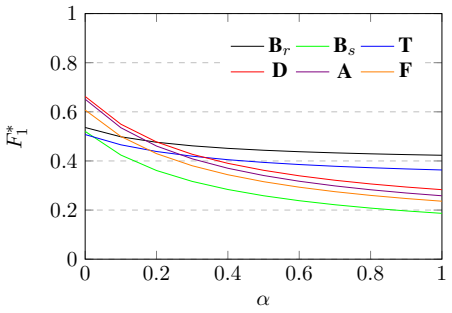

Figure 5: α-based fuzzy F_1 scores

Our main conclusions are as follows:

- Though there is consensus on some EL annotations, opinions differ for common entities, pro-forms, descriptive references, etc.

- The EL systems tested offer little or no support for pro-forms, meronymic references, referencing noun phrases, etc., despite there being considerable support for such EL annotations in the questionnaire.

- Our fine-grained evaluation distinguishes two groups of EL systems: one group targeting common entities and named entities, the other group focused on named entities.

The results of our questionnaire and system evaluation suggest the need for future work on supporting complex forms of reference within EL systems; the datasets we provide can be used to evaluate such approaches. Another important direction is to either reach a consensus on the EL task (perhaps in a similar style to MUC-6 for NER), or define protocols for evaluation in the absence of such a consensus; our fuzzy recall/F_1 metrics are concrete steps in this direction.

Acknowledgements The work of Henry Rosales-Méndez was supported by CONICYT-PCHA/Doctorado Nacional/2016-21160017. The work was also supported by the Millennium Institute for Foundational Research on Data (IMFD) and by Fondecyt Grant No. 1181896.

[8] We remark that P is agnostic to α and would result in a straight line; R^* thus follows the same trend as F_1^*.

[9] https://github.com/henryrosalesmendez/EL_exp

References

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In SIGMOD, pages 1247–1250.

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In CIKM, pages 1625–1628.

Michael Fleischman and Eduard H. Hovy. 2002. Fine Grained Classification of Named Entities. In COLING.

Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic Web Machine Reading with FRED. Semantic Web, 8(6):873–893.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A Brief History. In COLING, pages 466–471.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In CIKM, pages 545–554.

Kunal Jha, Michael Röder, and Axel-Cyrille Ngonga Ngomo. 2017. All that Glitters Is Not Gold - Rule-Based Curation of Reference Datasets for Named Entity Recognition and Entity Linking. In ESWC, pages 305–320.

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

Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design challenges for entity linking. TACL, 3:315–328.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: shedding light on the web of documents. In I-SEMANTICS, pages 1–8.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In LREC.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. TACL, 2:231–244.

Roberto Navigli. 2009. Word sense disambiguation: A survey. ACM Comput. Surv., 41(2):10:1–10:69.

Lev-Arie Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In ACL, pages 1375–1384.

Thomas Rebele, Fabian M. Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. 2016. YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. In ISWC, pages 177–185.

Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2018a. VoxEL: A Benchmark Dataset for Multilingual Entity Linking. In ISWC, pages 170–186.

Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2019. NIFify: Towards Better Quality Entity Linking Datasets. In WWW Companion Volume, pages 815–818.

Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2018b. What should Entity Linking link? In AMW.

Peter F. Strawson. 1950. On referring. Mind, 59(235):320–344.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2018. Anaphora and coreference resolution: A review. CoRR, abs/1805.11824.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. Commun. ACM, 57(10):78–85.

Jörg Waitelonis, Henrik Jürges, and Harald Sack. 2016. Don't compare apples to oranges: Extending GERBIL for a fine grained NEL evaluation. In SEMANTICS, pages 65–72.

Gong-Qing Wu, Ying He, and Xuegang Hu. 2018. Entity linking: An issue to extract corresponding entity with knowledge base. IEEE Access, 6:6220–6231.

Lotfi A Zadeh. 1965. Fuzzy sets. Information and control, 8(3):338–353.