Evolutionary intelligence manuscript No.
(will be inserted by the editor)

A Step Further Towards a Consensus on Linking Tweets to Wikipedia

Mohamed Cherif Nait-Hamoud · Fedoua Lahfa · Abdellatif Ennaji

Received: date / Accepted: date

Abstract The study of contemporary tweet-based Entity Linking (EL) systems reveals a lack of a standard definition and a consensus on the task. Specifically, identifying what should be annotated in texts remains a recurring question. This prevents proper design and fair evaluation of EL systems. To tackle this issue, the present paper introduces a set of rules intended to define the EL task for tweets. We experimented the effectiveness of the proposed rules by developing TELS, an end-to-end supervised system that links tweets to Wikipedia. The experiments conducted on five publicly available datasets show that our system outperforms the baselines with an improvement, in terms of overall macro F1-score (micro F1-score), ranging from 25.04% (7.32%) up to 35.36% (42.03%). Moreover, feature analysis reveals that when the annotation is not limited to very few entity types, the proposed rules capture more efficiently annotators' tacit agreements from datasets. Consequently, the proposed rules constitute a step further towards a consensus on the EL task.

Keywords Information extraction · Tweet entity linking · Topic extraction · Wikification

Mohamed Cherif Nait-Hamoud
University of Aboubekr Belkaid, department of science computing, BP 13000 Tlemcen, Algeria
University of Larbi Tebessi, department of mathematics and science computing, BP 12000 Tebessa, Algeria
E-mail: mohamed-cherif.nait-hamoud@univ-tebessa.dz

Fedoua Lahfa
University of Aboubekr Belkaid, department of science computing, BP 13000 Tlemcen, Algeria
E-mail: f_didi@mail.univ-tlemcen.dz

Abdellatif Ennaji
LITIS laboratory EA-4108, University of Rouen Normandie, Rouen, France
E-mail: abdel.ennaji@univ-rouen.fr

1 Introduction

Due to the fast growing adoption of social media technology, data shared in microblogs are considered as valuable source of information for plethora of applications including user interest mining [38], events detection [15] and healthcare [6]. Entity linking (EL) is a well-known technique used also in data extraction from microblogs such as Twitter. The main goal of the EL task is to map important sequence of terms called mentions into relevant resources of a reference knowledge base such as Wikipedia, DBpedia, or Yago. The linking process seeks for knowledge base resources, called entities, whose principal topics are the extracted mentions. When Wikipedia is used as a sole reference knowledge base, the problem is commonly known as Wikification.

Nowadays, several studies are conducted to deal with tweet entity extraction and linking [36, 5]. The informality, shortness and ambiguity of microblog texts make the EL task more challenging for tweets than for well-structured and rich texts. Due to the lack of formal definition of the EL task [30], various interpretations were adopted for both datasets annotation and systems design. Effectively, the examination of annotated mentions of gold standards used in previous works revealed this issue [17]. Besides, proposed systems do not follow neither a standard definition of the EL task nor rules specifying what exactly should be linked in texts. Supervised systems such as those proposed in [21, 11, 13] rely on self-collected datasets manually annotated to fit designers' own interpretations of the EL task. Unsupervised systems such as those proposed in [8, 23, 37] use mainly relations of entity mentions in knowledge base graphs for entity extraction and linking. However, this may induce high false positives rate due to the extraction of noisy entity mentions not coherent with the context. In the other hand, proposed systems are mostly validated on manually annotated and non-public datasets. The few shared datasets include annotations that do not follow specific guidelines and highly deviate from most adopted EL task interpretations.

The foregoing discussion shows that there is a lack of a consensus on the EL task which prevents fair and appropriate comparison of EL systems. Our contributions presented in this paper to tackle this issue are:

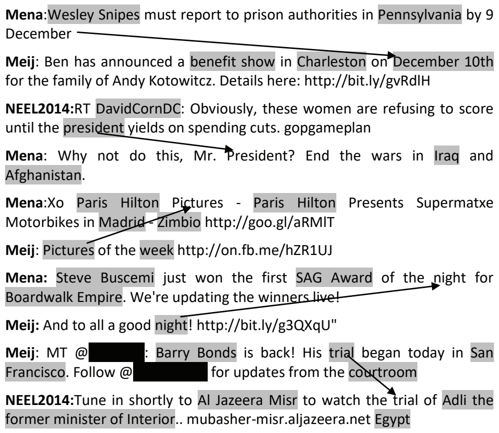
1. we introduce a set of Wikification rules specifying tweet mentions that need to be linked to Wikipedia.
2. we propose a set of guidelines for manual datasets annotation.
3. we provide a new end-to-end supervised system named TELS (for Tweet-based Entity Linking System) for Wikification of tweets. Developed in light of the proposed rules and guidelines, TELS outperforms most commonly known systems, namely: TagMe, AIDA, DBpedia Spotlight and WAT.
4. we share a revised version of Meij dataset [21] that fits the rules and guidelines proposed in this paper[1].

[1] The revised version of Meij is shared at:
https://drive.google.com/file/d/1CiGNjyK350Lyn3h5yLreOKNKq7Eb4EZo/view?usp=sharing

The remainder of this paper is organized as follows. Section 2 describes the problem statement. Section 3 presents related works on tweet-based EL systems and EL task issues. Section 4 introduces the proposed rules for Wikification of tweets. Section 5 describes the implementation details of the proposed system TELS. Section 6 presents the conducted experiments, the obtained results and discusses limitations of the proposed system TELS. Section 7 concludes the paper and gives insights on future works.

2 Problem statement

To investigate the extent of the problem, we examined available reference datasets used for the evaluation of EL systems and we experienced the behavior of some commonly known systems. Figure 1 illustrates annotated samples picked from three existing datasets namely, Meij [21], Mena [12] and Microposts NEEL 2014 [25] used in previous works [21, 13, 7] for tweet-based EL systems design and evaluation.



Mena:Wesley Snipes must report to prison authorities in Pennsylvania by 9 December

Meij: Ben has announced a benefit show in Charleston on December 10th for the family of Andy Kotowitcz. Details here: http://bit.ly/gvRdlH

NEEL2014:RT DavidCornDC: Obviously, these women are refusing to score until the president yields on spending cuts. gopgameplan

Mena: Why not do this, Mr. President? End the wars in Iraq and Afghanistan.

Mena:Xo Paris Hilton Pictures - Paris Hilton Presents Supermatxe Motorbikes in Madrid - Zimbio http://goo.gl/aRMlT

Meij: Pictures of the week http://on.fb.me/hZR1UJ

Mena: Steve Buscemi just won the first SAG Award of the night for Boardwalk Empire. We're updating the winners live!

Meij: And to all a good night! http://bit.ly/g3QXqU"

Meij: MT @[redacted]: Barry Bonds is back! His trial began today in San Francisco. Follow @[redacted] for updates from the courtroom

NEEL2014: Tune in shortly to Al Jazeera Misr to watch the trial of Adli the former minister of Interior.. mubasher-misr.aljazeera.net Egypt

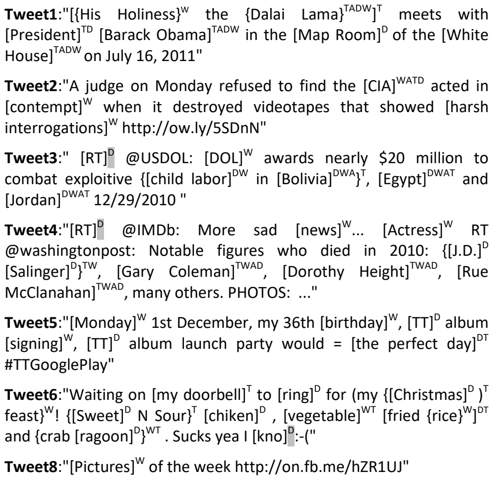
Fig. 1 Divergence of existing datasets annotation

The examination of annotated entity mentions highlighted in gray in Figure 1 reveals the following divergences:

1. mentions '*pictures*' and '*night*' were annotated in Meij but not in Mena
2. mention '*president*' was annotated in NEEL2014 but not in Mena
3. mention '*trial*' was annotated in Meij but not in NEEL2014
4. temporal mention '*December 10th*' was annotated in Meij but mention '*9 December*' was not annotated in Mena

Moreover, most previously proposed systems were evaluated on non-publicly shared datasets [8, 11, 31, 14]. Very few tweet-based datasets are shared for comparison. Available datasets include disputable annotations or are mainly designed for different tasks. Specifically, Meij dataset [21] was criticized in [11] for its extremely lenient annotation; Micropost NEEL 2016 collection [26] was initially introduced for Named Entity Recognition and Linking (NERL) which mainly differs from the EL task; Microposts NEEL 2014 datasets [25] include annotations of days of the week, dates, years and numbers that extremely deviate from the interpretation adopted by most EL systems. The two remaining datasets Brian [20] and Mena [12] include respectively 22% and 6% of non-Wikipedia links. These latter were used in [13] for linking mentions to external home pages in addition to YAGO knowledge base.

Furthermore, existing systems adopt different interpretations of the EL task. To show this fact, Figure 2 presents eight tweet samples annotated using TagMe [8], WAT [23], DBpedia Spotlight [4] and AIDA [37]. The mentions of different entities within the tweet samples are shown between different kind of brackets indicating the list of systems that annotate each mention.



Tweet1:"[{His Holiness}^{W} the {Dalai Lama}^{TADW}]^{T} meets with [President]^{TD} [Barack Obama]^{TADW} in the [Map Room]^{D} of the [White House]^{TADW} on July 16, 2011"

Tweet2:"A judge on Monday refused to find the [CIA]^{WATD} acted in [contempt]^{W} when it destroyed videotapes that showed [harsh interrogations]^{W} http://ow.ly/5SDnN"

Tweet3:" [RT]^{D} @USDOL: [DOL]^{W} awards nearly $20 million to combat exploitive {[child labor]^{DW} in [Bolivia]^{DWA}}^{T}, [Egypt]^{DWAT} and [Jordan]^{DWAT} 12/29/2010 "

Tweet4:"[RT]^{D} @IMDb: More sad [news]^{W}... [Actress]^{W} RT @washingtonpost: Notable figures who died in 2010: {[J.D.]^{D} [Salinger]^{D}}^{TW}, [Gary Coleman]^{TWAD}, [Dorothy Height]^{TWAD}, [Rue McClanahan]^{TWAD}, many others. PHOTOS: ..."

Tweet5:"[Monday]^{W} 1st December, my 36th [birthday]^{W}, [TT]^{D} album [signing]^{W}, [TT]^{D} album launch party would = [the perfect day]^{DT} #TTGooglePlay"

Tweet6:"Waiting on [my doorbell]^{T} to [ring]^{D} for (my {[Christmas]^{D})^{T} feast}^{W}! {[Sweet]^{D} N Sour}^{T} [chiken]^{D} , [vegetable]^{WT} [fried {rice}^{W}]^{DT} and {crab [ragoon]^{D}}^{WT} . Sucks yea I [kno]^{D}:-("

Tweet8:"[Pictures]^{W} of the week http://on.fb.me/hZR1UJ"

Fig. 2 EL task interpretations. T: TagMe, W: WAT, D: DBpedia Spotlight, A: AIDA

It can be shown from Figure 2 that while some systems such as WAT and DBpedia Spotlight link common words such as "*pictures*" and "*ring*", AIDA restricts to dominant entities such as "*CIA*" and "*Barack Obama*". The illustrated samples show also how systems deal with overlapping entities. Specifically, TagMe annotates the longest description of entity mentions such as: "*His Holiness Dalai Lama*" in **Tweet1**, "*child labor in Bolivia*" in **Tweet3**, "*Sweet N Sour*" and "*fried rice*" in **Tweet6**; while some systems annotate merely chunks of these mentions.

We believe that the adoption of rules for designing EL systems and guidelines for datasets annotation will alleviate the aforementioned problems. For instance, if we adopt annotating longest and coherent entities within tweet texts as a rule, the annotation of **Tweet6** of Figure 2 leads to the results shown in Table 1. The proposed rule excludes linking common words "*my doorbell*" and "*ring*" and discards shorter mentions such as "*ragoon*", "*Christmas*", "*sweet*" and "*rice*". This motivates examining further rules for standardizing the EL task and the adoption of guidelines for datasets annotation.

Table 1 Annotated mentions based on a hypothetical rule

Mention	Entity (Wikipedia page ID)	Corresponding Wikipedia title
Christmas feast	3712168	Christmas dinner
sweet n sour	1017309	Sweet and sour
vegetable	5791492	Vegetable
crab rangoon	47775352	Crab Rangoon
chicken	5741239	Chicken as food
fried rice	509468	Fried rice

3 Related works

The foregoing discussion suggests that the EL task is not well-defined but subject to various interpretations. An in-depth examination of previous works is needed to identify the EL issues and weaknesses of contemporary EL systems.

3.1 EL task issues

Previous studies raised the issue of EL systems replication noticing the use of invaluable resources like obsolete reference knowledge bases and no longer available reference datasets. In this perspective, Hassibi et al. [16] conducted a replication study on TagMe [8] that they qualified as the most popular entity linking system. Hassibi et al. [16] examined TagMe according to requirements introduced in SIGIR 2015 Workshop, namely: *Repeatability*, *Reproducibility*, and *Generalizability*. Repeatability informs about the possibility of repeating results under the same conditions. Whereas, reproducibility is concerned with reproducing results when the conditions are different but comparable. However, the version of the knowledge base and test datasets used in [8] are no longer available. In addition, current version of TagMe API presents deviations from the original implementation but the differences are not documented. Hassibi et al. stated that the resources provided by TagMe designers are invaluable for replicating their approach. The issue of benchmarking was also raised by Cornolti et al. [3] and Usbeck et al. [32] where some benchmark datasets and protocols are proposed for evaluating system performances. Unfortunately, the authors in [3, 32] did not propose any requirement regarding the quality of annotation which prevents the use of this benchmark for fair and appropriate comparison. Besides, these proposed benchmarks include collections dealing with two different scopes, regardless of Meij [21] and Microposts NEEL 2014 datasets [25] introduced to annotate tweets, the remaining collections deal with plain text annotation.

Finally, recent works raised an issue of a great importance related to the different interpretations of the entity linking problem [19, 33, 35]. Rosales-Mendez et al. [27] highlighted the importance of formalizing the concept of entity and the benefits for the construction of gold standards. Jha et al. [17] proposed a set of rules for document annotation for EL and NERL tasks and proposed to adopt entity types, introduced previously in MUC-6 conference [10], to define what should be linked. In addition, they derived a semi-automatic tool named EAGLET for gold standards annotation checking. Rosales-Mendez et al. in [28] proposed a tool named NIFify to explore the quality of gold standards and the evaluation of EL systems. In a latter study, Rosales-Mendez et al. [29] stated that defined entity types fail to fit all the interpretations of the EL task. To tackle this problem, the authors conducted a questionnaire involving 37 authors to establish a consensus about what could be linked. They suggested the development of a fine-grained categorization of the different types of entity mentions and links. Consequently, they re-annotated some EL datasets with respect to these categories and proposed a fuzzy recall metric to address the lack of a consensus. However, the experiment conducted by the authors has two main limitations: (1) the two used short texts in the questionnaire are not representative; they do not highlight the choice of the participants regarding marking same mentions within different contexts; (2) the experiment is based on voting and does not justify the choices made by the participants; in other words, the experiment does not indicate what strategy is used by each participant to identify entities.

3.2 Tweet-based EL systems

Several works proposed tweet-based entity linking approaches and systems [36, 5]. These endeavors use mainly either supervised or unsupervised techniques for spotting and linking entities. In both cases, a set of predefined features and criteria were used. These features such as *Link Probability* [8, 21], *Commonness* [21, 22, 8], *Relatedness* [21, 22, 8] and *Coherence* [8, 11] are mostly derived from seminal works. Existing works use all or subsets of those features with potential adaptation of their definitions. DBpedia Spotlight [4] is an unsupervised system that does not exploit relatedness for the extraction of entities. Entity extraction in DBpedia Spotlight is carried out with a spotting algorithm. This latter performs text tokenization and substring matching using the Aho-Corasick algorithm and prefix tree. A generative probabilistic model is used in DBpedia Spotlight to associate relevant entities to mentions.

EL systems that exploit relatedness and coherence for the extraction of entities are called collective linking systems. TagMe [8] is an unsupervised system based on collective linking. TagMe starts by extracting candidate entities based on a vote that combines Commonness and Relatedness, then, entities are filtered using Coherence. Other unsupervised systems such as AIDA [37], KAURI [31] and WAT [23] are graph-based systems. AIDA extracts entities using Stanford CRF [9], then a graph is built and entity linking is carried out considering dense regions of the graph. WAT combines Commonness and Relatedness using a graph for entity extraction and linking, similar to TagMe. KAURI incorporates user interest to improve entity disambiguation. Han et al. [14] proposed to incorporate the degree of direct references between candidates to improve KAURI. Ran et al. [24] explored topical coherence, they used attention mechanism inspired from attention-based models and incorporated user interest in a factor graph model. Feng et al. [7] proposed to mine entities from Twitter instead of knowledge bases to consider the influence of time on entity relevance. Instead of using knowledge bases, they propose to extract dominant candidate entities from a Twitter dataset using a community detection algorithm. Table 2 shows the drawbacks of unsupervised systems.

Table 2 Drawbacks and main differences of unsupervised EL systems

Methods	EL interpretation	Drawbacks/Remarks
Spotlight [4]	Based on topic pertinence, contextual ambiguity and disambiguation confidence	– Spotting entities is performed separately from disambiguation (not an end-to-end system)
AIDA [37]	Based on links of the knowledge base graph	– Spotting entities are performed separately from disambiguation (not an end-to-end system)
KAURI [31], Han et al. [14], Ran et al. [24]	Based on links of the knowledge base graph	– Link all named entities of tweets published by one user (a variant of the EL Task) – Need additional information about user topics of interest – Deal only with tweets and could not be used for short texts
Feng et al. [7]	Based on dominance of entities in a collected dataset of tweets	– Treats only disambiguation without any restriction on detected entities which promotes the detection of noisy entity mentions and induces more false positives.
TagMe [8], WAT [23]	Based on links of the knowledge base graph	– Use Commonness and a knowledge base graph without restrictions on entities which promotes the detection of common words and induces more false positives.

Some other systems [13, 11] used supervised or hybrid techniques for linking tweets to other knowledge base sources such as Yago and DBPedia. Guo et al. [11] made use of a structural SVM along with popular features such as page views, information about the context of tweets and features derived from Relatedness and Coherence. Meij et al. in [21] generated a ranked list of candidate pairs (mention, entity) using an n-gram language model. Thereafter, they used Random Forest (RF) classifier for entities' extraction and linking. It is the most similar study to our work. The three main differences with our work are: (1) we used a different set of features associated to rules and guidelines that define a clear interpretation of the EL task (2) the contextual similarity is not carried out as a separated step as in [21], instead, it is incorporated as features used to build the model, and (3) the coherence between entities omitted in [21] has been taken into account in our proposed system. Habib et al. [13] worked on linking entities to Wikipedia knowledge base and also to external home pages. The authors adopted an approach based on Support Vector Machines (SVM) and Conditional Random Fields (CRF) models. First, candidate named entities with the highest recall are extracted using a CRF model. Then, a ranked list of entities is obtained using a language model similar to the work of Meij et al. [21]. Finally, the SVM classifier is used to filter ranked entities. The main drawbacks and differences of EL supervised systems are highlighted in Table 3.

Table 3 Drawbacks and main differences of supervised EL systems

Methods	EL task interpretation	Drawbacks/Remarks
Meij [21]	Task definition based on a self-annotated and shared dataset	– Dataset used to define the EL task is criticized for its extremely lenient annotation – Coherence of entities regarding tweet contents is not considered
Mena [13]	Task definition based on a self-annotated shared dataset	– Use Google search API which is no longer available for free – Link to Yago and external home pages – Coherence of entities regarding tweet contents is not considered
Guo et al. [11]	Type-based annotation	– Restricts to 6 entities types : Person, Location, Organization, TV Show, Book/Magazine, Movie

Our investigation revealed the following two issues that motivated the work presented in this paper.

1. weakness of Tweet-based EL systems and the existing attempts to establish a consensus on the EL task.

2. absence of benchmark datasets for tweets based upon common annotation guidelines.

4 Rules for proper Wikification of tweet contents

To tackle the problem of lack of a consensus on the EL task, we propose a set of rules to identify and link mentions to Wikipedia. The definition of the task requires introducing the basic concepts for the sake of clarity.

Definition 1 (Mention). Each occurrence of a term or sequence of terms in a text represents a mention m. Each mention is defined by a pair (p_m, l_m), where p_m is its starting position in the text and l_m its length.

Definition 2 (Entity). Excluding Wikipedia redirect and disambiguation pages, each Wikipedia page represents an entity identified by a unique page ID.

Definition 3 (Annotation). An annotation $< m, E >$ is the linking of a mention m, located in a tweet, to an entity E.

Definition 4 (Entity dominance). The dominance of an entity E corresponds to the extent to which E is popular and central in a given source of common knowledge.

Definition 5 (Entity relevance). The relevance of an entity E to a given mention m, located in a tweet t, is proportional to:

1. topical coherence of E with the context of t
2. dominance of E with respect to a recent common knowledge source

Intuitively, given a mention m and an entity E that corresponds to a proper Wikipedia page (see Definition 2), a pair $< m, E >$ is a valid annotation if the mention m is the principal topic of the entity E, the topic of E is the most similar to the context of the tweet and E is dominant according to human common knowledge. Dominance of entities is used mainly when the context is lacking to capture important entities and to discard common words. More formally, $< m, E >$ is a valid annotation if and only if the following conditions hold:

R1. m is not a Twitter user mention[2]
R2. E is a proper entity according to Definition 2
R3. m is the primary topic of the entity E
R4. m is the longest description of E that occurs in the tweet
R5. E is relevant according to Definition 5

[2] A twitter user mention is a user defined nickname, it occurs in tweets preceded by the symbol @

Proposed rules state that, Twitter user mentions are user-defined nicknames that do not actually depend on the context of tweets. Consequently, **R1** prevents linking this type of mentions. Rule **R2** prevents linking mentions to Wikipedia redirect and disambiguation pages. Redirect pages do not carry any contextual information but are rather shortcuts to other pages, where disambiguation pages are used in Wikipedia to manage conflicts occurring when several articles have the same title and so often address different topics. Hence, disambiguation pages have not primary topics which discord with **R3**. Rule **R3** prevents linking mentions to inappropriate entities by taking into account merely the primary topic of each entity. Rule **R4** promotes linking longest mentions found in tweets. For instance, in the passage "*Harry Potter and the Deathly Hallows Part II broke box office records*", the mention "*Harry Potter and the Deathly Hallows Part II*" is likely to be selected and linked instead of the short and general mention "*Harry Potter*". Finally, **R5** allows linking mentions to relevant entities. Relevance of an entity to a given mention is identified based on (1) the coherence of the entity with the context of the mention and (2) the dominance of the entity regarding a recent common knowledge source. Due to the noisy content of tweets where the coherence of entities with the context of mentions cannot be determined effectively, the dominance of entities improves the annotation decision and discards common words. The dominance of an entity depends on the given common knowledge source, the more the source is recent the more the dominance is accurate. Consider, for example, the short passage "*The president George Bush*", recent knowledge allow to link "*George Bush*" to the Wikipedia page of George Bush the son instead of that of George Bush the father.

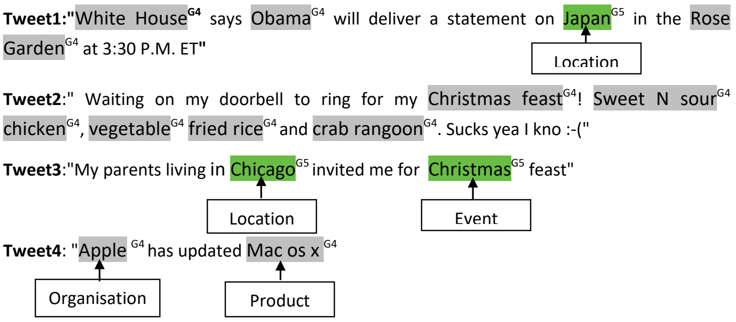
In fact, what causes the lack of a consensus in the EL community is the ambiguity around the relevance of entities (see Definition 5). More precisely, the trade-off between the dominance of an entity and its coherence with the context of tweets is the key point. Unless this trade-off is quantified, a trade-off could be learned from provided manually annotated datasets that follow a precise definition of the EL task. Making such datasets available requires the introduction of easy-to-use guidelines for the annotation of reference datasets. To this purpose, we introduce five guidelines that include rules **R1-3** discussed earlier; and guidelines **G4** and **G5** defined as follows:

G4. Annotate longest coherent entities in tweets.

G5. Annotate from the remaining mentions in the text dominant entities of types: Person, Organization, Location, Character, Events, Products, and Thing proposed in the taxonomy of the NEEL 2016 challenge [26].

Guidelines **G4** and **G5**, associated to rules **R4-5**, allow to solve the trade-off between entities coherence with the context and their dominance. During the manual annotation process, **G4** and **G5** need to be used sequentially. Starting with **G4** enables getting clear relevant entities by considering coherence only. By **G5**, remaining dominant entities are identified using a type-based annotation rule. The choice of NEEL 2016 challenge taxonomy of entity types is based on: (1) the findings of Kwak et al. [18], where it was shown that 85% of tweet topics are headline news, and (2) our assumption that those types cover the most dominant topics of news headlines.

To show how the proposed guidelines can be used, consider the hypothetical dataset presented in Figure 3. For the proper annotation of this dataset, **G4** is firstly used. Therefore, mentions highlighted in gray are annotated and linked to entities coherent with the context of their respective tweets. Then, **G5** is used for annotation of eventual remaining entity mentions. Therefore, mentions fitting the adopted list of types are annotated. Identified mentions in each sample of Figure 3 are highlighted in green with an indication of their types. Eventhough mentions '*Apple*' and '*Mac OS x*' of *Tweet4* are dominant entities, they are annotated following the guideline **G4**. This is due to the application order of the guidelines that gives the priority to the coherence of entities with the topic.



Tweet1:"White House^{G4} says Obama^{G4} will deliver a statement on Japan^{G5} in the Rose Garden^{G4} at 3:30 P.M. ET"

Tweet2:" Waiting on my doorbell to ring for my Christmas feast^{G4}! Sweet N sour^{G4} chicken^{G4}, vegetable^{G4} fried rice^{G4} and crab rangoon^{G4}. Sucks yea I kno :-("

Tweet3:"My parents living in Chicago^{G5} invited me for Christmas^{G5} feast"

Tweet4: "Apple ^{G4} has updated Mac os x ^{G4}

Fig. 3 Applicability of the proposed guidelines

Due to the genericity of the proposed rules, prior to systems implementation some design choices should be specified. These latter include the choice of the common knowledge source that allows to evaluate the dominance of entities, the set of criteria that define entities topical coherence and dominance, and contexts associated to each entity to enable the evaluation of entities topical coherence with tweet contexts.

In the following, we introduce a new EL system developed in light of the proposed rules to validate their effectivectiveness.

5 Proposed EL system

The proposed system TELS is designed for linking a tweet to Wikipedia pages representing its topical entities. As shown in Figure 4, the overall process of TELS is based on six main steps (1) Inverted index and context dataset creation, (2) tweet text preprocessing, (3) generation of candidate pairs (mention, entity), (4) feature extraction, (5) entity extraction and disambiguation, and (6) entity longest description filtering.

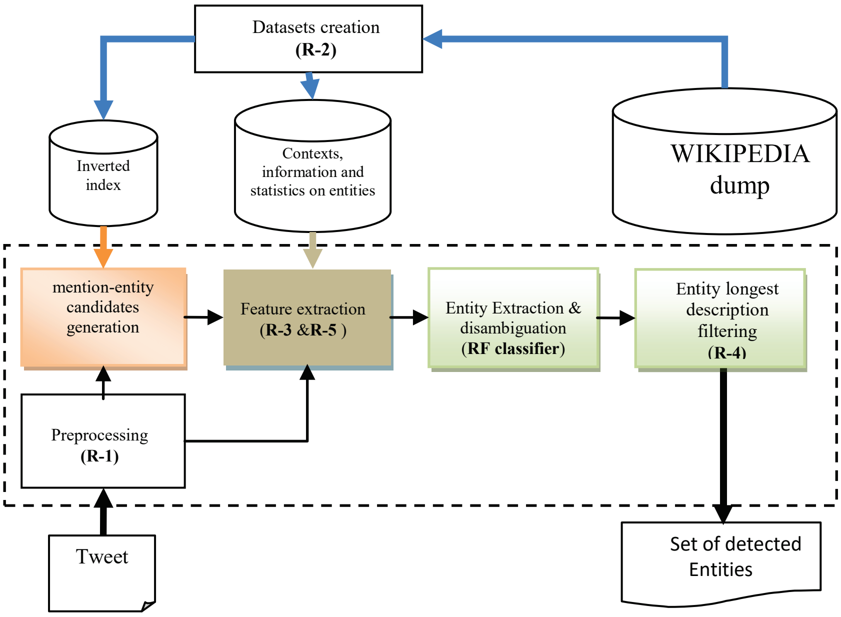

Datasets creation
(R-2)
Inverted index
Contexts, information and statistics on entities
WIKIPEDIA dump
mention-entity candidates generation
Feature extraction
(R-3 &R-5)
Entity Extraction & disambiguation
(RF classifier)
Entity longest description filtering
(R-4)
Preprocessing
(R-1)
Tweet
Set of detected Entities

Fig. 4 TELS architecture

After a pre-processing step, to capture all significant mentions in texts, TELS extracts all sequences of n consecutive terms located in the input tweet (a.k.a., n-grams); with n ranging from 1 to the total number of terms in the tweet. Thereafter, each n-gram is matched to the entries of a pre-built dictionary that includes all anchors used in the Wikipedia dump. If an n-gram matches an anchor, then all Wikipedia pages referenced by the anchor are associated to the n-gram to form candidate pairs (n-gram, entity). Afterwards, a representative feature vector is extracted from each candidate pair and fed into Random Forest classifier for entity spotting and linking in one shot. Finally, obtained overlapping pairs are filtered to keep the pair (n-gram, entity) with the longest entity mention as recommended by rule **R4**.

Basically, TELS differs from the other systems in that TELS follows a set of rules that provide a clear and precise definition of the EL task. The proposed rules are implemented in the different components of TELS as detailed in the following sections.

5.1 Preprocessing and candidate generation

Due to the noisy content of tweets, a preprocessing step is necessary for all the process of TELS to clean tweet texts. For this sake, First, Twitter user mentions are removed according to **R1**. Second, as camelcases are often used in tweets to incorporate topics of bloggers, hashtags are expanded to extract each term starting with a capital letter. For instance, the hashtag "#The-ForceAwakens" is expanded to the following sequence of terms "*The Force Awakens*". After the removal of URLs, tweet texts are split into phrases ending at punctuation symbols. This choice is adopted to avoid the extraction of mentions formed from terms originally separated by a punctuation symbol. For instance, in the passage of text *" ... free. Fight for human rights"*, the mention *"free Fight"* should not be extracted. For acronyms, point symbols are replaced by a special character and restored after text splitting. Afterwards, all possible sequence of n consecutive terms (i.e., *n-grams*) are extracted from each phrase. Each n-gram is identified by its phrase number, its position in the phrase and its length. Then, n-grams of each phrase are matched to the entries of the dictionary of anchors. If an n-gram matches an anchor, candidate pairs (n-gram, entity) are generated for each Wikipedia page (entity) referenced by this anchor.

5.2 Feature extraction

Once the candidate pairs generated, TELS uses preprocessed texts to extract a feature vector for each pair (n-gram, entity). Basically, the proposed features are strucutred in two subsets thought to fulfill the requirements of rules **R3** and **R5**. As discussed earlier, rule **R5** ensures that extracted entities from a tweet are either coherent with the context or dominant entities. Features associated to rule **R5**, described in Table 4 below, can be classified in two types: dominance features and topical coherence features. For each entity E, dominance features include the page rank (r_E) as a centrality marker from Wikipedia knowledge base graph; and the number of page views during the year 2019 (v_E) as a popularity marker from Wikimedia project[3]. Topical coherence features are mainly obtained from the overlap of an input tweet text with three proposed contexts ($L_E, C_E, \mathscr{C}_E$). The context L_E stands for the set of all references (i.e., anchors or redirect page titles used in Wikipedia) that link to E. The context C_E refers to the categories section of E extracted from its textual content. Finally, $\mathscr{C}_E$ corresponds to the set of anchors and page titles of Wikipedia pages referenced in the abstract of E. The context $\mathscr{C}_E$ is used to capture the degree of direct cross-references between entities.

Features of the second subset, described in Table 5, are related to rule **R3** introduced to ensure that the n-gram is the principal topic of the entity. In addition to Commonness (CMNS) and Term Frequency-Inverse Document Frequency ($TF - IDF$) that consider both the n-gram and the entity, we

[3] https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews

Table 4 Features associated to R5 selected and designed for topical coherence and dominance

Features	Description
Topical coherence features	
$L_E Overplap(t, E) = \|t \cap L_E\| \setminus \{n - gram\}$	Overlap between a preprocessed tweet t and the L_E context excluding the current n-gram
$InCategory(n\text{-}gram, C_E)$	$\begin{cases} 1, \text{ if n-gram occurs in the categories of } E) \\ 0, \text{ otherwise} \end{cases}$
$C_E Overlap(n\text{-}gram, t, E) = \|t \cap C_E\| \setminus \{n\text{-}gram\}$	Overlap between a preprocessed tweet t and the C_E context excluding the current n-gram
$\mathscr{C}_E Overlap(n\text{-}gram, t, E) = \|t \cap \mathscr{C}_E\|$	Overlap between a preprocessed tweet t and the $\mathscr{C}_E$ context
Dominance features	
r_E	Page rank of *E*
v_E	Number of visits of the Wikipedia page E during the year 2019

include two types of features: mention-centric features and entity-centric features. The first type of features exploits only information about the n-gram, where the second considers only the entity. Entity-centric features include TFL that indicates the frequency of entity title in the context L_E, and PTT that represents the overlap of the entity title with the tweet.

Proposed mention-centric features include MRF that stands for the modified relative frequency of the n-gram. Instead of considering only the exact occurrence of the n-gram in the context L_E, we extend the relative frequency used in [21] to take into account occurrences of anchors containing the n-gram. However, we notice that relative frequency and MRF are less significant if the n-gram corresponds to a redirect page title. Effectively, contrary to anchors, redirect page titles may occur at most once in the context L_E. To this purpose, we include the feature TOR that indicates if the n-gram corresponds to an anchor or to a redirect page title. The modified relative frequency MRF is also less significant when all references of an entity are different. To this purpose, we include the feature $DIST$ to capture the distribution of the context L_E. If all references of the context L_E are different $DIST$ tends to 1, otherwise it tends to $1/|L_E|$. Mainly, Wikipedia anchors are short and concise descriptions of page titles. Unfortunately, this is not always the case which causes long an-

Table 5 Features associated to **R3** selected and designed for relevance of pairs (n-gram, entity)

Features	Remarks
$CMNS(\textit{n-gram}, E) = \frac{\#(\text{occurrences of n-gram in } L_E)}{\sum'_E (\#\text{occurrences of n-gram in } L'_E)}$	Probability of E being a target of a link that corresponds to the n-gram
$MRF(\textit{n-gram}, E) = \frac{\lvert l, l \in L_E \text{ s.t. n-gram in l}\rvert}{\lvert L_E \rvert}$	Frequency of the occurrence of n-gram in L_E including its occurrences as a chunk of other anchors.
$TOR(\textit{n-gram}, E)$	$\begin{cases} 1, & \text{if n-gram is a redirect} \\ 0, & \text{if n-gram is an anchor} \\ 2, & \text{if n-gram is both} \end{cases}$
$DIST(E) = \frac{\#\text{ distinct elements of } L_E}{\lvert L_E \rvert}$	Distribution of anchors. It captures the distribution of the context L_E; it tends to 1 if all the references are different and to $1/\lvert L_E \rvert$ in the opposite case.
$PNT(\textit{n-gram}, E) = \frac{\#\text{ tokens of n-gram in title of E}}{\#\text{ of tokens of the title of E}}$	Proportion of n-gram in the preprocessed title of E
$TFIDF(\textit{n-gram}, E) = R_E * \log_{10} \frac{\lvert W \rvert}{\lvert R_{\textit{n-gram}} \rvert + 1}$	Term Frequency-Inverse Document Frequency of n-gram. R_E is the number of times n-gram referenced E. $R_{\textit{n-gram}}$ is the number of pages referenced by n-gram. $\lvert W \rvert$ represents the number of Wikipedia pages.
$\textit{Pureness}(\textit{n-gram}, E) = \frac{\lvert l, l \in L_E \text{ s.t.n-gram} = l \rvert}{\lvert l, l \in L_E \text{ s.t.n-gram in } l \rvert}$	Proportion of the exact occurrence of n-gram in L_E
$\textit{Absorption}(\textit{n-gram}, E) = \frac{\lvert l, l \in L_E \text{ s.t. l in n-gram}\rvert}{\lvert L_E \rvert}$	Proportion of elements of L_E contained in n-gram
$PTT(t, E) = \frac{\#\text{ tokens of cleaned title in tweet t}}{\#\text{ tokens of title}}$	Proportion of E's title tokens in a tweet t
$TFL(E) = \frac{\#\text{ occurrences of the E title in } L_E}{\lvert L_E \rvert}$	Title frequency in L_E

chors to be likely less frequent in the context L_E of an entity E. To alleviate this issue we add two new features, *Absorption* and *Pureness*. The former is used to indicate the proportion of anchors contained in the n-gram in the context L_E, where the latter represents the ratio of the n-gram occurrences to the number of anchors containing the n-gram in L_E.

The collected features for each candidate pair are used for entity extraction and disambiguation phases based on a supervised learning technique.

5.3 Entity extraction and disambiguation

Our system TELS is an end-to-end system, it performs entity extraction (spotting) and linking in one step. Effectively, the two steps are mapped to a pairwise classification using Random Forest (RF) that predicts if a candidate pair (*n-gram, entity*) is a valid annotation according to **R3** and **R5**. The choice of RF is based on the findings of Meij et al. [21] and Speck et al. [34], where it was shown that RF is the best classifier for the EL task. The RF classifier is an ensemble learning technique based on bagging that prevents overfitting. It consists of generating N-decision trees from N-iterations on random subsets of samples and retaining the class that corresponds to the mode of all classes. For each node of a decision tree, Random Forest selects randomly m features to obtain the best split.

The disambiguation phase depends on the extraction of candidate pairs and their corresponding features. These two phases exploit two datases built from a Wikipedia dump.

5.4 Prebuilt datasets

The proposed system TELS uses two prebuilt databases: a dictionary of anchors inlcuding redirect page titles (i.e., inverted index) and a database of contexts. These databases are created from the English Wikipedia dump of January 2019. The dictionary of anchors is a key-value store that associates to each entry (Wikipedia anchor or redirect page title) the set of Wikipedia pages it refers to, along with a score representing the number of times the entry was used as cross-reference to each page. The database of contexts is also a key-value store that includes the following information for each Wikipedia page:

1. page title
2. list of anchors and redirect page titles that link to this page
3. categories of the page included in the Wikipedia categories section
4. list of anchors and titles of Wikipedia pages referenced in the abstract
5. number of times the page was visited during the year 2019
6. rank of the page in the Wikipedia knowledge base graph

To built the above datasets, the Wikipedia dump is firstly preprocessed. After parsing the Wikipedia dump and removing disambiguation pages, 12.404.539 pages were retained. Thereafter, 4.803.670 redirect pages and 2.560.791 non-Wikipedia English articles were removed ending up with 5.040.078 pages. The collected pages are used to build the databases according to **R2**.

To implement the two databases, LMDB (Lightening-Mapped Database) storage engine [2] was used due to its high read performances. Effectively, LMDB read transactions are very cheap, the keys are indexed in a B+ tree. Besides, LMDB uses mapped memory with zero copy lookup and exploits the operating system's buffer cache. The comparison of storage engine performances[4] shows that LMDB is the best alternative. Moreover, to improve serialization, Google Protocol Buffers[5] were used due to their best performances compared with XML, JSON and other existing formats[6].

6 Experiments

As there is lack of benchmark datasets for tweet-based system evaluation, some shared datasets were collected and adapted to be used as gold standards.

6.1 Gold standards

For the evaluation of our proposed system, we used five datasets: Mena [12], Microposts NEEL 2016 Train and Dev [26], Microposts NEEL 2014 Train [25] and Brian [20]. The choice of these datasets is based on two main criteria. First, they present interesting and various properties needed for comparison: (1) multi-domain datasets, (2) datasets of various sizes including different proportions of event (tweets likely to contain entities) and non-event tweets. Second, annotations of these datasets adopt various interpretations of the EL task. Microposts NEEL 2016 datasets were introduced in NEEL (Named Entity rEcognition and Linking) 2016 challenge. The task consists of detecting entities and their types (person, location, organization, etc.). Entity mentions are linked to the English version of the DBpedia resources if they exist and NIL identifier is used otherwise. They were built considering event tweets and non-event tweets to evaluate performances in avoiding false positives. Microposts 2016 Train and Dev collections include Twitter mention annotations and links to redirect and disambiguation pages. Brian and Mena entity mentions are linked to YAGO and to external website home pages. Mena dataset includes a total of 33 links (6%) to non-Wikipedia pages. In addition, 32 links are not proper articles; but Wikipedia redirect pages. Brian dataset includes 352 (22%) links to non-Wikipedia pages and 55 links to Wikipedia redirects. For appropriate use of the above datasets to compare systems, the datasets are adapted to fit the rules **R1-2** introduced in Section 4. Table 6 below describes general information about the content of the datasets after revision. The revisions performed are as follows:

– links to redirects are replaced with their target pages

[4] http://www.lmdb.tech/bench/inmem/

[5] https://developers.google.com/protocol-buffers/docs/overview

[6] https://labs.criteo.com/2017/05/serialization/

- links to disambiguation pages are replaced or removed
- Twitter mentions are removed
- NIL mentions are removed
- non-Wikipedia links are removed

Table 6 Properties of the datasets revised to respect the proposed rules. [a]: events including death of Amy Winehouse, London Riots, Oslo bombing and Westgate shopping Mall terrorist attack; [b]: Events on UCI Cyclo-cross, Star Wars and Force Awakens Premiere; [c]: events including the death of Amy Wine-house, the London Riots and the Oslo bombing

Revised collection	# tweets	# annotations	% non-event tweets	Domains
Micro. NEEL 2016 Train	6000	5606	47%	– Worthy events from 2011 to 2013[a] – Events from 2014 to 2015[b]
Micro. NEEL 2016 Dev	100	214	4%	– US primary election – Star Wars The force awakens premiere
Micro. NEEL 2014 Train	2339	3815	32.19%	– Notable events from 15th July 2011 to 15th August 2011[c]
Brian	1603	1232	51,5%	– Economic recession – Australian Bushfires, a large regional event – Gaz explosion in Bozeman, a very quick local event
Mena	162	477	0%	Sports, politics, movie stars

The remaing available collection Meij [21] is selected for training. This choice is motivated by the fact that Meij tweets are of general interest domains, contrarily to the other datasets that were collected from specific domains.

6.2 Training dataset

A revised version of Meij dataset [21] was used to train our proposed system. As the raw version was criticized in [11], we inspected its annotations and carried out a deep revision to fulfill the guidelines introduced in section 4 for datasets annotation. Specifically, we revised the annotations to fit topical coherence and dominance requirements. In addition, we noted that the selected tweets do not include examples of ambiguous entities. To this end, we augmented the dataset with 53 short texts of less than 140 characters (14% of the whole dataset) to add representative examples of ambiguous entities. We ended up with a dataset including 428 tweets with a proportion of 16.12% of non-event tweets.

6.3 TELS settings

TELS system is based on Random Forest (RF) classifier which mainly depends on two hyperparameters: (1) the number of iterations N and (2) the feature size m. While the former indicates the number of trees in the forest, the latter represents the maximum number of features that are randomly selected at each node to determine the best split. For the selection of best settings, we performed a grid search to fine-tune these two parameters. Therefore, we varied the value of m from 4 to the total number of features which is 16. Besides, we varied the number of trees N in the forest from 30 to 100 by a step of 10 and from 100 to 1000 by a step of 100. We evaluated the models using the Out-Of-Bag (OOB) technique. Accordingly, the full Meij dataset is used for training from which OOB samples are used to estimate the prediction performances. The OOB method has been proven to be as accurate as testing performances on a dataset of the same size of the training dataset [1]. Figure 5 below depicts the effects of parameters selection on performance in terms of F1-score.

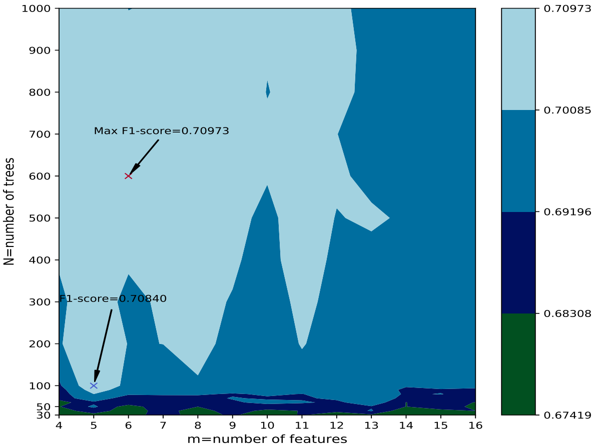


Fig. 5 Effects of parameters' settings on TELS performances

The performances when $N \geq 100$ lie in a narrow interval ranging from 0.6919 to 0.7097, regardless of the values of m as shown in Figure 5. The best performance is obtained with the settings $m = 6$ and $N = 600$, while comparable performance (with a decrease of only 0.18%) is obtained with $m = 5$ and $N = 100$. The blue and the purple curves in Figure 6 indicate, repectively, the maximal and minimal performances obtained when fixing a parameter while varying the other one alternately. The green curve in Figure 6(b) indicates that when m is set to 5, the performances are comparable with the maximal performances obtained when varying m for different settings of N.

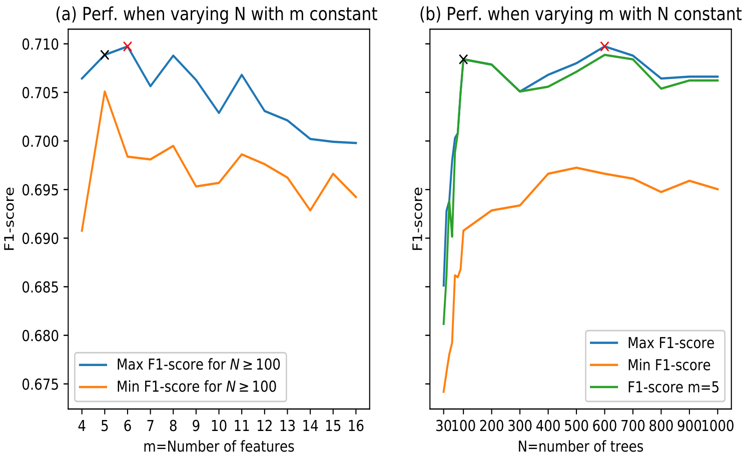


Fig. 6 Sensitivity of TELS to parameters settings. The red cross indicates the best settings, the black cross indicates the performance with $m = 5$ and $N = 100$

The present experiment shows that with a number of trees $N \geq 100$, TELS is stable and relatively not sensitive to variation of the parameter m. Hence, we retained for TELS the settings $m = 5$ and $N = 100$ representing the best trade-off between performance and complexity in terms of number of trees and number of features.

6.4 Baselines settings

For the evaluation of our proposed system, the most known and leading state-of-the-art baselines with shared implementations were retained: TagMe[7], WAT[8], AIDA[9] and DBpedia Spotlight[10]. To obtain the results of the state-of-the-art systems, we used their corresponding RESTful APIs. However, despite the importance of settings, most previous works omit explicit specification of the settings adopted for baseline systems. We noticed that the evaluation will not be fair, when using default settings. As a demonstrative example, Figure 7 shows the evaluation of TagMe and WAT on Microposts Dev 2016, Mena, Microposts NEEL 2014 Train and Brian datasets when considering both default

[7] TagMe website: `https://tagme.di.unipi.it/tagmehelp.html`
[8] WAT website: `https://sobigdata.d4science.org/web/tagme/wat-api`
[9] AIDA website: `https://www.mpi-inf.mpg.de/yago-naga/aida/`
[10] DBpedia Spotlight website: `https://www.dbpedia-spotlight.org/`

and recommended settings. The results show a significant increase in systems performances and a change in their ranking. Therefore, for fair and appropriate comparison, system settings should be taken into consideration.

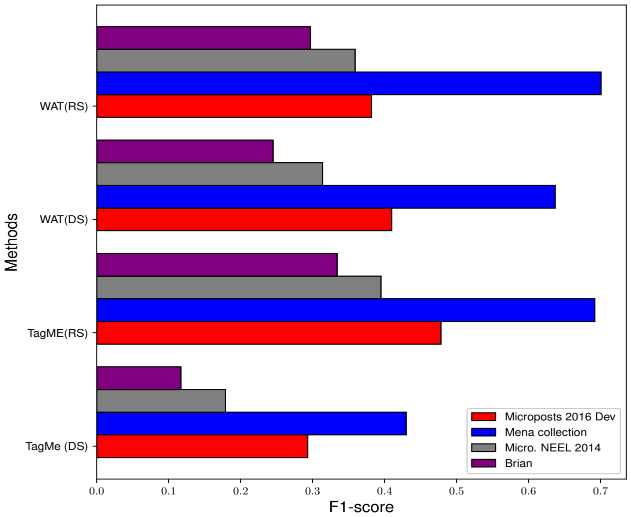


Fig. 7 Comparison of baselines settings: RS:Recommended; DS:Default Settings

In the following experiments, default settings are used for AIDA due to the absence of any recommendation, DBpedia Spotlight confidence parameter is set to 0.5, as recommended in [4], TagMe pruning threshold is set to 0.2 as recommended in [8] and WAT ρ (rho) parameter is set to 0.2.

6.5 Results

To assess the overall performances of TELS and the baselines, we used two metrics: micro F1-score and macro F1-score. As the gold standards are different in terms of samples size (see Section 6.1), using micro F1-score amounts to considering the five gold standards as one dataset. Besides, macro F1-score shows how systems perform overall cross the different datasets. Table 7 below shows that TELS outperforms the baselines with an overall increase, in terms of macro F1-score, from 25.04% up to 35.3%. In terms of micro F1-score, the increase performed by TELS ranges from 7.32% up to 42.03%.

Table 8 illustrates the performances of TELS and the baselines obtained on individual datasets in terms of recall, precision and F1-score. The results show that TELS outperforms all the baselines in terms of F1-score on four datasets and performs as well as AIDA on Brian dataset. These results show the effectiveness of TELS and its robustness against different dataset types.

Table 7 Overall performances of the baselines and TELS with settings (m=5, N=100)

	Overall micro F1-score	Overall macro F1-score
TagMe	0.3514	0.400
WAT	0.330	0.4211
DBpedia Spotlight	0.390	0.4196
AIDA	0.4367	0.4328
TELS	0.4687↗	0.5412↗

Table 8 Results of systems comparison with (m=5, N=100) as settings of TELS

		N2016Train	N2016Dev	Mena	Brian	N2014Train
TagMe	Recall	0.514	0.396	0.733	0.455	0.397
	Precision	0.233	0.293	0.490	0.264	0.392
	F1-score	0.320	0.337	0.587	0.334	0.395
WAT	Recall	0.479	0.353	0.765	0.510	0.434
	Precision	0.219	0.415	0.646	0.209	0.306
	F1-score	0.301	0.381	0.700	0.297	0.359
DBpedia Spotlight	Recall	0.502	0.333	0.704	0.426	0.397
	Precision	0.302	0.348	0.583	0.232	0.467
	F1-score	0.377	0.340	0.638	0.301	0.429
AIDA	Recall	0.362	0.057	0.597	0.293	0.276
	Precision	0.617	0.631	0.844	0.633	0.662
	F1-score	0.457	0.106	0.699	**0.400**	0.390
TELS	Recall	0.446	0.532	0.673	0.377	0.334
	Precision	0.512	0.775	0.790	0.428	0.619
	F1-score	0.4772↗	0.6312↗	0.7270↗	**0.401**	0.4427↗

6.6 Rules effectiveness and features analysis

The effectiveness of the proposed rules and guidelines for datasets annotation can be highlighted through the analysis of features importance. Table 9 shows the importance of features associated to **R3** and **R5** (see Table 5 and Table 4, respectively).

Table 9 Analysis of the importance of features grouped by rule

Datasets	R3	R5	
		Dominance	Topical coherence
Meij raw	0.702	0.154	0.141
Meij revised	0.634	0.120	0.246
Brian	0.752	0.103	0.145
Mena	0.703	0.116	0.181
NEEL 2016 DEV	0.404	0.118	0.478
NEEL 2016 Train	0.613	0.207	0.180
NEEL 2014 Train	0.597	0.185	0.218

The analysis indicates that our revised version of Meij dataset induces an increase in topical coherence importance, a decrease in dominance and **R3** importances relatively to raw Meij annotation. This suggests that topical coherence was less considered in raw Meij annotation. The results show also that features importance depends and varies according to the annotation performed for each dataset. Specifically, **R3** features are found more important than other features in all the datasets except for NEEL 2016 Dev where topical coherence features are found more important. In NEEL 2016 Train, dominance features are found more important than topical coherence features which can be explained by the restriction of annotation to eight specific entity types.

Table 10 shows the performances of TELS when rules are included incrementally. The results show an incremental increase of performances when incorporating features of the different proposed rules. Exceptionally, the performance of TELS on Brian dataset decreases when adding features associated to **R5**. A deep investigation shows that the decrease is due to the restrictive annotation, in the Brian dataset, to only three specific entity types namely: ”*Person*”, ”*Organization*” and ”*Location*” [20]. Due to this restriction, only the location ” *Victoria*” is annotated in the passage ” *Victoria’s bushfire disaster*”. Therefore, the event ”*Victoria’s bushfire*” of the previous passage and events such as ”*economic recession*” and ”*bushfires*” that occur frequently in Brian dataset are ignored and not annotated in the dataset which mainly affects the effectiveness of **R5**. Consequently, when datasets annotations do not restrict to very few entity types, the proposed rules capture efficiently tacit annotators’ agreements from datasets.

Table 10 Effects of incremental insertion of rules on performances, (*) NEEL 2014 Train with removal of numbers and temporal entities from annotation

Datasets	Rules		
	R3	R3+R4	R3+R5+R4
Mena	0.6689	0.6796↗	0.7270↗
Brian	0.4366	0.4388↗	0.401↘
NEEL 2016 DEV	0.5134	0.5269↗	0.6312↗
NEEL 2016 Train	0.4062	0.41306↗	0.4772↗
NEEL 2014 Train	0.3821	0.3862↗	0.4427↗
NEEL 2014 Train*	0.4151	0.4206↗	0.4871↗

Finally, to show the impact of using proper guidelines for datasets annotation, we evaluated the performances of baselines on our revised version of Meij dataset and its raw version. Since the revised Meij dataset was used for training TELS, we exclude it from this experiment. The results depicted in Figure 8(a) show an increase, highlighted in yellow, in the performances of all the baselines regarding all metrics.

Besides, Figure 8(b) illustrates the performance of TELS and the baselines on raw NEEL 2014 Train and a revised version of this dataset. The revision of this later follows rules **R1-2** for annotation and consists of excluding entities not coherent with the context from annotation such as numbers, days of the week, months, years, etc. The results of Figure 8(b) show an increase in recall

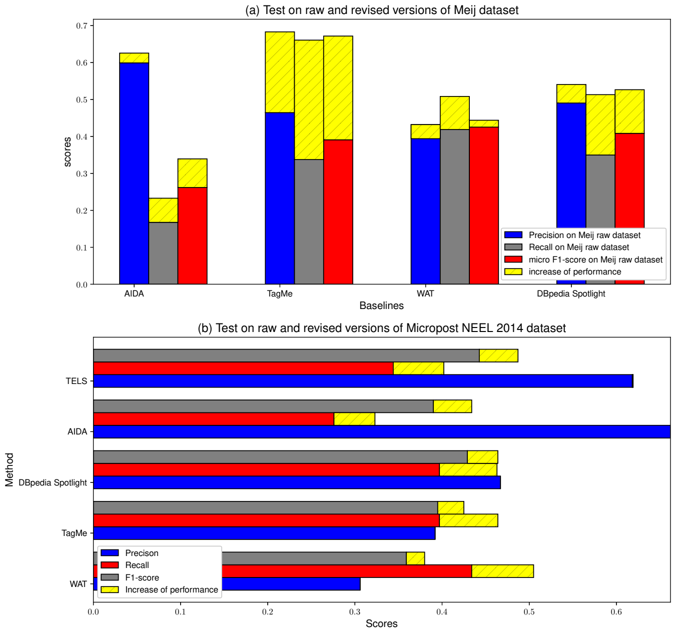


Fig. 8 Effects of proposed rules and guidelines on the performances of systems

and F1-score for TELS and all the baselines compared with their evaluation on the raw dataset.

6.7 Discussion

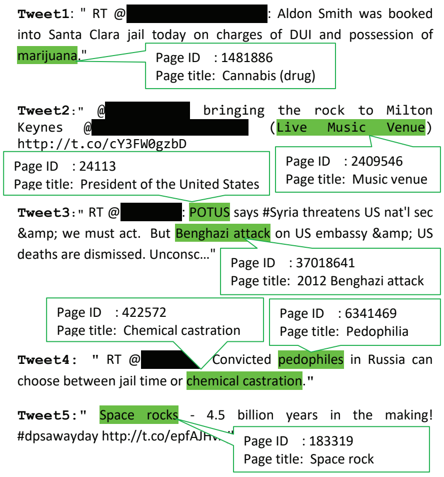
The largest performance obtained by TELS compared with the baselines is performed on Microposts NEEL 2016 DEV, with at least 65.66% of F1-score improvement. To explain this high performance, we analyzed the importance of features in terms of information gain (I.G) for all the experimented datasets. Table 11 shows that Commonness (CMNS) is the most important feature for all the datasets, except for Microposts NEEL DEV 2016 where CMNS is found less important than the overlap of texts with entity anchors ($L_E Overplap$). Moreover, when using alternately $L_E Overplap$ and $CMNS$ as unique features, the results show that the performance of TELS on Microposts NEEL DEV 2016 is more important when using the feature $L_E Overplap$. Hence, systems that do not exploit this feature or rely only on Commonness for relevant candidate pairs (n-gram, entity) extraction will have weaker performances in such case.

The evaluation on the five datasets in terms of overall performances shows that TELS provides the best trade-off between precision and recall. In fact, these results are essentially due to the proposed rules that allow to capture annotators' tacit agreements from previously annotated datasets.

Table 11 Features analysis

	$L_E Overplap$			$CMNS$		
	Importance (I.G)	Rank	F1-score	Importance (I.G)	Rank	F1-score
NEEL2016DEV	0.449	1	0.426	0.150	2	0.259
NEEL2016Train	0.094	3	0.115	0.184	1	0.239
Brian	0.058	5	0.064	0.214	1	0.293
Mena	0.123	2	0.212	0.125	1	0.300
NEEL2014Train	0.087	3	0.127	0.202	1	0.210

Despite the encouraging results of the proposed system, false positives still produced by TELS need further examination. To evaluate the behavior of TELS trained on the revised version of Meij collection, we analyzed some false positives from Microposts NEEL 2016 Train dataset. Figure 9 below shows some examples of false positives highlighted in green. For instance, the annotation "*Space rocks*" in *Tweet5* is a false positive since it was linked to an entity dealing with rock music instead of rocks present in space. Effectively, when there is a lack of contexts or only a single entity occurs in a tweet, coherence and similarity become useless. In this case, TELS relies on the dominance of entities which paradoxically may generate false positives. However, false positives detected in *Tweet1* to *Tweet4* are mostly due to the fact that TELS annotations are also based on coherence of entities with the context, while the annotations of NEEL2016 Train dataset include only entities of types Person, Location, Organization, Event, Character, Product and Thing.



Tweet1: " RT @[redacted]: Aldon Smith was booked into Santa Clara jail today on charges of DUI and possession of marijuana."

Page ID : 1481886
Page title: Cannabis (drug)

Tweet2:" @[redacted] bringing the rock to Milton Keynes @[redacted] (Live Music Venue) http://t.co/cY3FW0gzbD

Page ID : 24113
Page title: President of the United States

Page ID : 2409546
Page title: Music venue

Tweet3:" RT @[redacted]: POTUS says #Syria threatens US nat'l sec &amp; we must act. But Benghazi attack on US embassy &amp; US deaths are dismissed. Unconsc..."

Page ID : 37018641
Page title: 2012 Benghazi attack

Page ID : 422572
Page title: Chemical castration

Page ID : 6341469
Page title: Pedophilia

Tweet4: " RT @[redacted] Convicted pedophiles in Russia can choose between jail time or chemical castration."

Tweet5:" Space rocks - 4.5 billion years in the making! #dpsawayday http://t.co/epfAJHv."

Page ID : 183319
Page title: Space rock

Fig. 9 Analysis of our system behavior

It is worth noting that TELS is designed for tweets and short texts; but, not for long documents. This particularity is due to our design choices for the two steps: candidates generation and feature extraction. The technique used for coherence evaluation in the feature extraction phase is not suitable for long documents. Effectively, these features are based on text overlap and involve a look up in the inverted index for each token extracted from the text. This technique is suitable for very short texts, but becomes time and space consuming when the number of tokens is large; which is the case of long documents. In addition, assessing long documents similarity needs more sophisticated techniques. In the other hand, the extraction of mention-entity candidates consists of identifying all the n-grams located in input texts and matching them to the inverted index. As the number n ranges from 1 to the total number of terms in phrases, the absence of punctuation leads to a larger number of n-grams which affects the complexity of the proposed system.

7 Conclusion and future works

In this paper, we introduced a set of rules and guidelines for (1) the design of tweet-based EL systems and (2) the annotation of reference datasets. Given the obtained results, the proposed rules bring a step further towards a consensus on the EL task. Moreover, we proposed TELS as an end-to-end EL system developed in light of the proposed rules. The evaluation process showed that TELS performs an increase of 25.04% better than the baselines in terms of overall macro F1-score that represents the best trade-off between precision and recall. The proposed rules enabled TELS to capture efficiently annotators' tacit agreements from annotated datasets.

As further contribution in the domain, we intend to annotate the gold standard datasets following the proposed rules in this paper. Therefore, revised datasets can be used by the EL community as benchmark for tweet-based EL task. In addition, we intend to improve TELS performances by (1) incorporating information about user profiles and topics of surrounding posts to enrich the context of entities, (2) gathering external information about entities from other sources such as news headlines. The ultimate objective of this work is to standardize the EL task and design a tool for automatic checking of datasets annotation.

References

1. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
2. Chu, H. (2011). MDB: A Memory-Mapped Database and Backend for OpenLDAP. In *Proceedings of the 3rd International Conference on LDAP*, Heidelberg, Germany, pp. 35–47.
3. Cornolti, M., Ferragina, P., and Ciaramita, M. (2013). A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd In-*

ternational Conference on World Wide Web, Rio de Janeiro, Brazil, pp. 249–259.

4. Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, Graz, Austria, pp. 121–124.
5. Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing and Management*, 51(2):32–49.
6. Serban, O., Thapen, N., Maginnis, B., Hankin, C., and Foot, V. (2019). Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing and Management*, 56(3):1166–1184.
7. Feng, Y., Zarrinkalam, F., Bagheri, E., Fani, H., and Al-Obeidat, F. (2018). Entity linking of tweets based on dominant entity candidates. *Social Network Analysis and Mining*, 8(46):1–16
8. Ferragina, P. and Scaiella, U. (2010). TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, Toronto, Canada, pp. 1625–1628.
9. Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, USA, pp. 363–370.
10. Grishman, R. and Sundheim, B. (1996). Message Understanding Conference-6, *Proceedings of the 16th conference on Computational linguistics*, USA, pp. 466471.
11. Guo, S., Chang, M. W., and Kiciman, E. (2013). To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, pp. 1020–1030.
12. Habib, M. B. and Van Keulen, M. (2012). Unsupervised improvement of named entity extraction in short informal context using disambiguation clues. In *Proceedings of the Workshop of Semantic Web and Information Extraction*, Galway City, Ireland, pp. 1–9.
13. Habib, M. B. and Van Keulen, M. (2016). TwitterNEED: A hybrid approach for named entity extraction and disambiguation for tweet. *Natural Language Engineering*, 22(3):423–456.
14. Han, H., Viriyothai, P., Lim, S. J., Lameter, D., and Mussell, B. (2019). Yet Another Framework for Tweet Entity Linking (YAFTEL). In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval*, San Jose, CA, USA, pp. 258–263.
15. Hasan, M., Orgun, M. A., and Schwitter, R. (2019). Real-time event detection from the Twitter data stream using the TwitterNews+ Framework.

Information Processing and Management, 56(3):1146–1165.

16. Hasibi, F., Balog, K., and Bratsberg, S. E. (2016). On the reproducibility of the TAGME entity linking system. In *Ferro N. et al. (eds) Advances in Information Retrieval, ERIC2016*, LNCS, vol. 9626, Springer, Cham, pp. 436–449.
17. Jha, K., Röder, M., and Ngomo, A. C. N. (2017). All that glitters is not gold Rule-based curation of reference datasets for named entity recognition and entity linking. In *Blomqvist E., Maynard D., Gangemi A., Hoekstra R., Hitzler P., Hartig O. (eds) The Semantic Web. ESWC 2017*, LNCS vol. 10249, Springer, Cham, pp. 305–320.
18. Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, Raleigh North Carolina, USA, pp. 591–600.
19. Ling, X., Singh, S., and Weld, D. S. (2015). Design Challenges for Entity Linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.
20. Locke, B. and Martin, J. (2009). Named Entity Recognition: Adapting to Microblogging. *University of Colorado UG Thesis*, pages 1–12.
21. Meij, E., Weerkamp, W., and De Rijke, M. (2012). Adding semantics to microblog posts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, Seattle, Washington, USA, pp. 563–572.
22. Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, Napa Valley, California, USA, pp. 509–518.
23. Piccinno, F. and Ferragina, P. (2014). From Tagme to WAT: A new entity annotator. In *Proceedings of the 1st ACM International Workshop on Entity Recognition and Disambiguation, Co-located with SIGIR 2014*, USA, pp. 55–61.
24. Ran, C., Shen, W., and Wang, J. (2018). An attention factor graph model for tweet entity linking. In *Proceedings of the 2018 World Wide Web Conference*, Lyon, France, pp. 1135–1144.
25. Rizzo, G., Cano, A. E., Pereira, B., and Varga, A. (2015). Making sense of microposts (#Microposts2015) named entity recognition and linking challenge. In *Proceedings of the 5th Workshop on Making Sense of Microposts*, Florence, Italy, pp. 44–53.
26. Rizzo, G., Van Erp, M., Plu, J., and Troncy, R. (2016). Making sense of microposts (#Microposts2016) named entity recognition and linking (NEEL) challenge. In *Proceedings of the 6th Workshop on Making Sense of Microposts*, Montreal, Canada, pp. 50–59.
27. Rosales-Méndez, H., Hogan, A., and Poblete, B. (2018a). VoxEL: A benchmark dataset for multilingual entity linking. In *Vrandečić D. et al. (eds) The Semantic Web ISWC 2018. ISWC 2018*, LNCS, vol. 11137, LNCS, Springer, Cham, pp. 170–186.
28. Rosales-Méndez, H., Hogan, A., and Poblete, B. (2019). Nifify: Towards better quality entity linking datasets. In *Companion Proceedings of The 2019 World Wide Web Conference*, San Francisco, USA, pp. 815–818.

29. Rosales-Méndez, H., Hogan, A., and Poblete, B. (2020). Fine-grained evaluation for entity linking. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 718–727.
30. Rosales-Méndez, H., Poblete, B., and Hogan, A. (2018b). What should entity linking link? In *CEUR Workshop Proceedings*, vol. 2100, pp. 1–5.
31. Shen, W., Wang, J., Luo, P., and Wang, M. (2013). Linking named entities in tweets with knowledge base via user interest modeling. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, Chicago, Illinois, USA, pp. 68–76.
32. Usbeck, R., Röder, M., Ngomo, A. C. N., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Waitelonis, J., and Wesemann, L. (2015). GERBIL - General entity annotator benchmarking framework. In *Proceedings of the 24th international conference on World Wide Web*, Florence, Italy, pp. 1133–1143.
33. Van Erp, M., Mendes, P. N., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., and Waitelonis, J. (2016). Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia, pp. 4373–4379.
34. Speck, R., Ngonga, A.C.N. (2014). Ensemble Learning for Named Entity Recognition. In *Mika P. et al. (eds) The Semantic Web ISWC 2014. ISWC 2014*, LNCS, vol. 8796, Springer, Cham, pp. 293–308.
35. Weichselbraun, A., Braoveanu, A. M., Kuntschik, P., and Nixon, L. J. (2019). Improving named entity linking corpora quality. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Varna, Bulgaria, pp. 1328–1337.
36. Wu, G., He, Y., and Hu, X. (2018). Entity Linking: An Issue to Extract Corresponding Entity with Knowledge Base. *IEEE Access*, 6:6220–6231.
37. Yosef, M. A., Hoffart, J., Bordino, I., Spaniol, M., and Weikum, G. (2011). AIDA: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment*, 4(12):1450–1457.
38. Zarrinkalam, F., Kahani, M., and Bagheri, E. (2018). Mining user interests over active topics on social networks. *Information Processing and Management*, 54(2):339–357.