ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1248

Unsupervised Ranking of Knowledge Bases for Named Entity Recognition

Yassine Mrabet[1, 2] and **Halil Kilicoglu**[1] and **Dina Demner-Fushman**[1]

Abstract. With the continuous growth of freely accessible knowledge bases and the heterogeneity of textual corpora, selecting the most adequate knowledge base for named entity recognition is becoming a challenge in itself. In this paper, we propose an unsupervised method to rank knowledge bases according to their adequacy for the recognition of named entities in a given corpus. Building on a state-of-the-art, unsupervised entity linking approach, we propose several evaluation metrics to measure the lexical and structural adequacy of a knowledge base for a given corpus. We study the correlation between these metrics and three standard performance measures: precision, recall and F1 score. Our multi-domain experiments on 9 different corpora with 6 knowledge bases show that three of the proposed metrics are strong performance predictors having 0.62 to 0.76 Pearson correlation with precision and 0.96 correlation with both recall and F1 score.

1 Background

With the tremendous growth in the amount of textual data, extracting semantic information from unstructured texts has become critical in several applications such as information retrieval, marketing, content management and question answering. Named entity recognition (NER), the task of identifying and categorizing textual mentions into pre-defined semantic categories, plays a key role in such applications. It is also often a prerequisite for other text mining processes such as relation extraction, keyword identification and document clustering.

The range of named entities covered by the NER task has grown continuously since the first MUC conference[3]. Starting with a few named entity categories, such as PERSON, LOCATION, and ORGANIZATION; nowadays, the entity linking task [22] addresses linking of textual mentions to entities from a reference database or knowledge base (KB) with no semantic restrictions on the type of entities. Besides their role as reference data, KBs are also increasingly used in NER methods. For instance, they have been used in designing features for supervised learning [24], as labeling sources in constructing training corpora [10, 28], and as resources for named entity disambiguation in both unsupervised [17] and supervised [25] approaches.

With the exponential growth of domain-specific KBs and the increasing heterogeneity in open-domain KBs, it becomes important to assess which KB is suitable to extract named entities in a given text or corpus. Qualitative assessments, while useful, can be time-consuming and require domain expertise [8, 4]. On the other hand, an automatic, quantitative evaluation based on KB and corpus characteristics can assist significantly in assessing suitability. We refer to this type of automatic evaluation as *knowledge base ranking for named entity recognition* (*KB ranking*).

The term *knowledge base ranking* is often used in the literature to refer to *fact ranking* or *entity ranking*, the task of finding the facts or entities that are the most relevant for a keyword query or a structured query [5, 2]. In this paper, we do not address this task.

Our work can be situated within the wider field of ontology evaluation (see [27] for a review). However, most works in this area cover tasks that are out of the scope of our study. These tasks include, for instance, the evaluation of the general representational (or domain) adequacy of a KB [8, 4], the inner cohesion of an ontology [31], finding relevant criteria for ontology design [18, 30], or checking logical consistency [9]. These criteria are constant for a given knowledge base and do not change according to different contexts of application.

As we are primarily interested in the use of knowledge bases for NER, only a few related studies stand out. Lozano-Tello et al. [15] proposed a generic framework called *OntoMetric* to evaluate the suitability of an ontology for a given application. They defined manually 160 generic features related to ontologies and designed an interface to help users decide which ontology is more relevant to their use case. The users have to manually enter their objective and criteria according to the defined vocabulary. This manual approach is limited and can not be applied to more complex applications such as NER, where we need to take into account the corpus characteristics in addition to ontology features. More generally, manual approaches are not suited to learning relevant features when several empirical observations are needed (e.g., observations on different NER corpora).

Gangemi et al. [8] proposed a formal model to evaluate ontologies, including a component for intended use situation. In particular, they considered the use of natural language processing to evaluate ontologies according to annotated corpora. However, the goal of the comparison was to evaluate the general usefulness (or quality) of an ontology: e.g., the frequency of occurrence of an ontology concept in the annotated corpus is used to measure the importance of the concept. In the same line, they also proposed to use the hierarchy of concepts and entropy to have an estimation of the usefulness of ontology concepts.

Their objective is basically different from ours; we want to evaluate the adequacy of a KB for NER in a given corpus. For the same KB, this evaluation is expected to give different results on different corpora. More precisely, the question that we want to answer is: "*if we want to use a KB to find and disambiguate named entities in a given text, how could we know which KB will provide better results?*".

Baseline approaches to KB ranking for NER could be derived from

[1] National Library of Medicine, Bethesda, MD, United States {mrabety, kilicogluh, ddemner}@mail.nih.gov
[2] CNRS/LORIA, France, yassine.mrabet@loria.fr
[3] `http://cs.nyu.edu/cs/faculty/grishman/muc6.html`

KB indexes such as the Linked Open Vocabulary[4], by selecting the KBs that have more candidate entities for a given textual mention; however, such an approach does not allow ranking of the KBs by their suitability for NER as other aspects come into play. These aspects include, for example, the ambiguity of the text to be annotated from the KB point-of-view, and the contextual similarity between the textual context of the named entities and the KB graph linking the corresponding (or candidate) KB concepts.

To the best of our knowledge, no automatic solution has been proposed previously to rank KBs according to their suitability for NER. This can be partly explained by the lack of KB-agnostic annotation tools for NER; most of the existing tools rely on a specific combination of a learning corpus and a KB.

In this paper, we propose a novel KB ranking method based on an unsupervised NER method. We formally define several evaluation metrics related to the lexical ambiguity of textual corpora and to the structural similarity between the knowledge base and the text to be annotated.

We studied the relevance of the proposed evaluation metrics in ranking 6 different KBs from both the open and biomedical domains for NER in 9 different corpora. Our results show that the proposed metrics are strongly correlated with precision, recall and F1 measures and that they can be used as predictors of the adequacy of a KB for NER in a given textual corpus. This finding paves the way to automatic and fine-grained selection and combination of knowledge bases for named entity recognition.

The remainder of the paper is structured as follows. In the next section, we describe the KB-agnostic named entity recognition method that underlies our approach and the evaluation metrics in more detail. We discuss the motivation behind the different metrics. In Section 3, we present our experiments. Finally, we discuss and analyze the results and perspectives in Section 4 before giving our concluding remarks.

2 Methods

In the current work, we consider a KB to consist of a set of concepts, instances, relations and a set of labels representing natural language expressions of concepts and instances. To ensure the required portability for our approach, we built an unsupervised NER method from an existing, KB-agnostic tool for entity linking called *KODA* [17]. In this section, we present the overall NER process and the evaluation metrics proposed to rank the KBs.

2.1 Named entity recognition

KODA is a KB-agnostic entity linking tool. It exploits TF-IDF indexing of the KB labels, and KB relations to disambiguate the entities in the input text. In the course of this study, we modified and extended KODA to build a NER method and used the extended tool as the basis of our experiments. Given a text t and a knowledge base k, our NER process follows the steps outlined below.

- Split t into sentences and perform part-of-speech tagging.
- For each sentence, select textual mentions corresponding to a sequence of allowed part-of-speech tags (e.g., noun, adverb, adjective).
- Use each textual mention as a keyword query to look up KB entities based on TF-IDF search. If no exact match is found between the mention and the KB entities, select all subsequences of words as potential candidates. The mentions recognized at this step are referred to as *candidate textual mentions*.
- Disambiguate ambiguous mentions (i.e., those that have more than one corresponding KB entity with the maximum TF-IDF score). Disambiguation is performed according to *global coherence*: i.e., select the entity that has more KB relations with the entities obtained from other textual mentions in t. This step is accomplished using Integer Linear Programming [17]. The generic disambiguation process can be viewed as the *selection of the subgraph of the KB that is the most similar to the textual context being annotated*.
- Determine the semantic category of the disambiguated textual mention, by using mappings of KB concepts, as detailed below.
- Filter the entities to keep only the semantic categories considered in the corpus.
- In the case of nested entities or overlapping entities with the same type (e.g., "*San Francisco, CA*" vs. "*San Francisco*"), keep only the entity with the best TF-IDF score.

In order to classify these named entities according to the considered semantic types (e.g., PERSON, LOCATION, ORGANIZATION), we built manual mappings between the concepts of the KBs and the semantic types considered in the target corpora. As large and dense concept hierarchies might be difficult to browse, we first computed the transitive closure offline by considering only the subset of instantiation facts (e.g., *RDF type relation*) and subsumption facts (e.g., *RDFS subClassOf relation*). Next, we sorted the concepts according to their frequency in the closure and extracted manually the relevant concepts, i.e., those that can be mapped to a named entity category according to the corpus. For example, *dbpedia*:*Place* was mapped to LOCATION in the CoNLL 2003 corpus [23], and *yago*:*wordnet_illness_114061805* was mapped to DISEASE in the I2B2 corpus [32].

In the online NER step, we collect all the classes associated with the KB entities linked to the disambiguated textual mentions then use the mappings to associate these mentions with a semantic type. If one mention is associated with more than one semantic type, it is considered as ambiguous and discarded from the results of the NER process.

Figure 1 shows an example of named entities recognized by our NER method using DBpedia on a sentence from a New York Times corpus [14]. In this example, there are multiple candidates with the same (best) TF-IDF score for the term "Malone" in the KB. Global coherence led to the selection of only one candidate (*dbpedia*:*Kevin_Malone*) because it is linked with the entities *dbpedia*:*Carlos_Perez_(pitcher)* and *dbpedia*:*San_Fransisco* in DBpedia triples.

2.2 Evaluation of KB Adequacy for NER

We propose several evaluation metrics by defining and combining three elementary principles:

1. *Ambiguity*: How ambiguous is the text with respect to a KB?
2. *Coverage*: How much of the text has been annotated and disambiguated with the KB?
3. *Structure*: To what extent did KB relations participate in the disambiguation?

To define relevant metrics taking into account these 3 aspects, we make the distinction between mentions recognized lexically with the

[4] http://lov.okfn.org/dataset/lov/

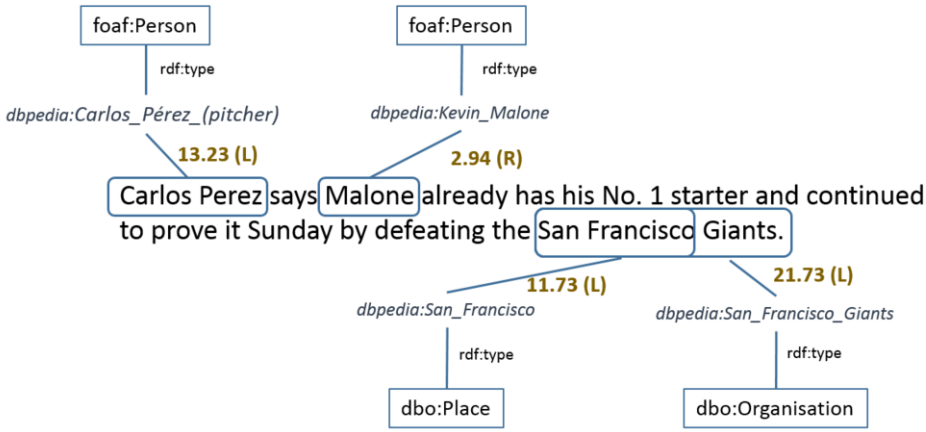

Figure 1. Example of named entity recognition using DBpedia. The numbers on the edges are TD-IDF scores. L indicates that disambiguation was performed lexically. R indicates that disambiguation was performed using KB relations.

best TF-IDF score and the ambiguous mentions disambiguated using the KB relations. We denote the set of all candidate mentions in a corpus C according to a KB k as $M_k(C)$ and the set of disambiguated mentions as $D_k(C)$. Figure 2 presents the mentions sets that are generated by our KB-agnostic recognition process.

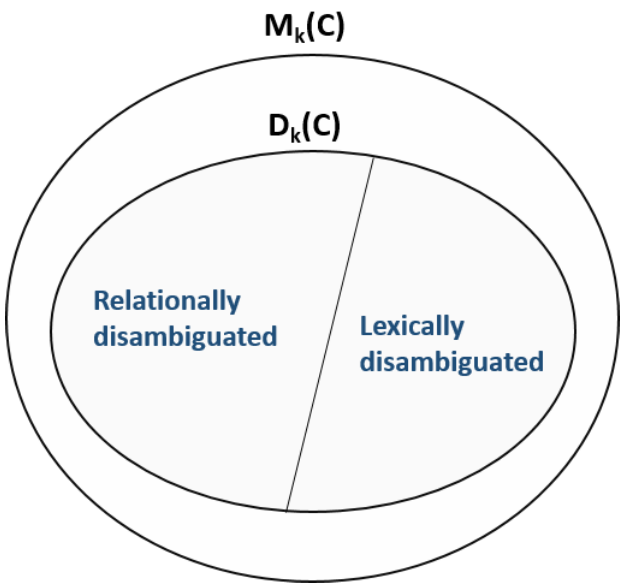

Figure 2. Named Entity Sets as recognized with Knowledge Base k in corpus C

We propose and study 10 evaluation metrics. These metrics are Coverage (V), Disambiguation Ratio (D), Lexical Disambiguation Ratio (L), Relation Disambiguation Ratio (R), Average Corpus Ambiguity (A), Average TF-IDF Score (S), Lexical Adequacy ($LEXQ$), Graph Adequacy (GQ), Weighted Quality (WQ) and Overall PERformance IndicAtor ($OPERA$). They are described below.

Coverage ($V_k(C)$): This metric indicates the percentage of corpus tokens that have been annotated and disambiguated with the knowledge base k for corpus C.

$$V_k(C) = \frac{\sum_{m \in D_k(C)} |tokens(m)|}{|tokens(C)|} \quad (1)$$

Disambiguation Ratio ($d_k(C)$): This metric indicates the ratio of textual mentions that have been disambiguated among the set of detected (annotated) mentions.

$$d_k(C) = \frac{|D_k(C)|}{|M_k(C)|} \quad (2)$$

Lexical Disambiguation Ratio ($L_k(C)$): This metric indicates the ratio of mentions that are disambiguated using only their TF-IDF score. Low $L_k(C)$ values indicate a bigger disambiguation problem. For a given mention $m \in M_k(C)$ it is computed as follows:

$$L_k(C) = \frac{|\{m \in M_k(C)\ s.t.\ N_{max}(m,k) = 1\}|}{|M_k(C)|} \quad (3)$$

Where $N_{max}(m,k)$ is the number of entities in k that share the maximum TF-IDF score for the mention m.

Relation Disambiguation Ratio ($R_k(C)$): This feature indicates the ratio of textual mentions from $D_k(C)$ that have been disambiguated using KB relations (cf. section2.1). Higher values of $R_k(C)$ indicate a stronger participation of the KB graph in disambiguation.

$$R_k(C) = \frac{|\{m \in D_k(C) s.t. N_{max}(m,k) > 1\}|}{|M_k(C)|} = 1 - L_k(C) \quad (4)$$

Average Score (S_k): This is the average best TF-IDF score for mentions in $M_k(C)$. A high TF-IDF average would indicate that the corpus targets specific subsets of the KB that use highly informative terms. S_k is computed as follows:

$$S_k(C) = \frac{\sum_{m \in M_k(C)} score(m,k)}{|M_k(C)|} \quad (5)$$

Average Corpus Ambiguity ($A_{k,T}(C)$): This metric represents the average ambiguity level in a corpus C according to a KB k (cf. equation 7). It is the average of the **Lexical Ambiguity** of each mention

m in $M_k(C)$. Highly ambiguous mentions are likely to be unclassifiable or wrongly classified by the KB. The formula for Lexical Ambiguity ($a_{k,T}(m)$) is presented in equation 6 below. T is the observation threshold, i.e., $a_{k,T}(m)$ is computed against the first T search results and $r \geq N_{max}(m,k)$ is the actual number of results from the KB index.

$$a_{k,T}(m) = \frac{Min(N_{max}(m,k),T)}{Min(r,T)} \quad (6)$$

$$A_{k,T}(C) = \frac{\sum_{m \in M_k(C)} a_{k,T}(m)}{|M_k(C)|} \quad (7)$$

From these elementary metrics, we derive several composite evaluation metrics to predict both the quantity and the quality of the disambiguation provided by one KB for for a given corpus.

Lexical Adequacy ($LEXQ_k(C)$): This metric represents the absolute ratio of named entities that have been disambiguated with TF-IDF search.

$$LEXQ_k(C) = L_k(C) \times V_k(C) \quad (8)$$

Graph Adequacy ($GQ_k(C)$): This metric indicates how useful the KB graph is in disambiguating named entities in a given corpus. Coverage is used as a coefficient to take into account the discrepancies in size and coverage between different KBs.

$$GQ_k(C) = R_k(C) \times V_k(C) \quad (9)$$

Weighted Quality ($WQ_k(C)$): This metric is a weighted combination of:

- A quality indicator for lexical disambiguation ($\frac{Norm(S_k(C))}{1+A_{k,T}(C)}$), which uses the normalized value of average TF-IDF score ($Norm(S_k(C))$), and the average corpus ambiguity ($A_{k,T}(C)$). The motivation here is (i) that a high average of TF-IDF values w.r.t. other KBs indicate that the textual mentions in the corpus are using (highly) informative terms from the KB and (ii) the more ambiguous the mentions are, the riskier is the selection of the one mention with best TF-IDF score.
- A quality indicator for relational disambiguation, consisting in the average corpus ambiguity $A_{k,T}(C)$. The motivation here is that having more candidate KB entities to chose from increases the odds of finding relations between the good candidates. However, if ambiguity is too high, it can lead to relations between the wrong KB entities. In practice, such high-ambiguity threshold would depend (i) on the considered knowledge base, (ii) on the targeted corpus, and (iii) on the observation threshold T used to compute $a_{k,T}(m)$. Therefore, for relational disambiguation, finding a balanced estimation between the positive and negative impact of ambiguity is not straightforward. In this paper, we chose to consider only the positive aspect of ambiguity for relational disambiguation and to analyze the impact of this choice in our experiments.

We use the contribution of lexical disambiguation $L_k(C)$ to weight the quality indicator for lexical disambiguation, and the contribution of relational disambiguation $R_k(C)$ to weight the quality indicator for relation-based disambiguation (we have from equation 4 that $L_k(C) = 1 - R_k(C)$). The final formula for the overall quality indicator $WQ_k(C)$ is:

$$WQ_k(C) = (L_k(C) \times \frac{Norm(S_k(C))}{1 + A_{k,T}(C)}) + (R_k(C) \times A_{k,T}(C)) \quad (10)$$

Overall PERformAnce Predictor ($OPERA_k(C)$): This metric combines quality metrics (weighted quality) with quantitative measures (coverage) to account for the size of the knowledge bases and the amount of lexical matches between the corpus and the KB. It also uses $d_k(C)$ as a factor indicating how successful the KB was in disambiguating the automatically detected mentions.

$$OPERA_k(C) = WQ_k(C) \times d_k(C) \times V_k(C) \quad (11)$$

For comparison, we defined two baseline metrics. The first metric uses the average TF-IDF score ($S_k(C)$). The second baseline metric is $S_k(C) \times d_k(C) \times V_k(C)$, which takes into account the coverage of the knowledge base to combine both basic quality and quantity factors.

3 Experiments

We applied our unsupervised named entity recognition method to extract named entities from 4 open-domain corpora and 5 biomedical corpora using 3 open-domain knowledge bases and 3 biomedical knowledge bases. The corpora used in the experiments are described below.

- **TREC** corpus [21] consists of sentences extracted from TREC documents. Similar to the CoNLL03 corpus, this corpus includes the following entity types: PERSON, ORGANIZATION, LOCATION, and OTHER.
- **NYT** corpus [14] consists of 8,000 named entities (PERSON, ORGANIZATION, LOCATION) from a random subset of New York Times articles (1998-2000) in the TREC corpus [26]. The documents were pre-annotated with a named entity tagger and then manually corrected by two annotators.
- **WikiNER** corpus [1] was created by manual annotation of the body text of 145 Wikipedia articles describing various named entities, with a roughly equal proportion of article topics from each of the four CoNLL03 entity types. Initial annotation was performed using a fine-grained inventory of 96 entity types (e.g., CITY, COMPANY) which were then mapped to CoNLL03 classes. Three annotators were involved in the annotation task and inter-annotator agreement was measured on a portion of the corpus.
- **CoNLL03** named entity corpus [23] consists of English and German documents. The English portion of the corpus, used in our experiments, is taken from the Reuters Corpus[5] and consists of news stories from August 1996 to August 1997. The corpus was manually annotated, mostly following the MUC guidelines. In addition to MUC named entity types (PERSON, ORGANIZATION, LOCATION), an additional category (MISC) was also annotated. The annotated entities are non-overlapping and non-nested.
- **AZDC** (Arizona Disease Corpus) [12] consists of 2,783 sentences from 793 PubMed abstracts annotated with disease mentions. One annotator performed the annotation. A textual mention was annotated if it could be mapped to a unique concept with a relevant semantic type (e.g., *Disease of Syndrome*, *Neoplastic Process*) in the UMLS Metathesaurus [3]. Acronyms and negated/hedged mentions were annotated, while symptoms, general disease classes (e.g., *infection*) and overlapping mentions were ignored.
- **i2b2** corpus [32] consists of discharge summaries contributed by Partners Healthcare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh Medical Center as well as progress reports from University of Pittsburgh Medical Center. The named

[5] http://trec.nist.gov/data/reuters/reuters.html

entity annotation was performed manually and focuses on three categories: PROBLEM, TEST, TREATMENT. All documents were de-identified.

- **NCBI** disease corpus [6] consists of 793 PubMed abstracts also used in AZDC; however, in this corpus, all sentences in these abstracts were annotated. Pre-annotations from an automatic classifier were used as the basis of annotation. The annotations guidelines were similar to those of AZDC. Nested and non-continuous mentions were not annotated.
- **CDR** corpus [29] consists of 1500 PubMed abstracts discussing chemical-induced diseases and side effects, annotated for DISEASE and CHEMICAL categories. The corpus was manually annotated by the CTD (Comparative Toxicogenomics Database)[6] staff.
- **Berkeley04** corpus [20] consists of the first 100 titles and the first 40 abstracts from 59 MEDLINE 2001 data files. No keywords were used to retrieve the documents. Named entities of PROBLEM and TREATMENT categories were annotated by a single annotator.

In the open-domain experiments, we considered only PERSON, ORGANIZATION, and LOCATION as semantic types and discarded MISC and OTHER, as they are too ambiguous from a knowledge-base perspective and strongly biased according to different corpora.

We tested our approach in both the open-domain and biomedical domain using 6 knowledge bases: DBpedia, Yago, OpenCyc, UMLS, Snomed-CT and MeSH, described below.

- **DBpedia**[13] is a community-curated RDF knowledge base constructed semi-automatically from Wikipedia. Each Wikipedia article is interpreted as an entity in DBpedia and articles' infoboxes are used to extract automatically raw RDF triples, using the article entity as subject, the first column of the infobox as predicate and the second column as object. Manual mappings are then performed by the DBpedia community to reconcile the predicate names with the RDF properties in the reference DBpedia schema. DBpedia entities are also linked to other datasets in the Linked Open Data cloud[7] such as YAGO or Freebase. In the scope of our experiments we used the English DBpedia 2014 version. After indexing, the DBpedia database consisted of 6,921,894 entities, 25,864,784 relations and 12,782,266 terms.
- **YAGO3** [16] is a large open-domain knowledge base built from Wikipedia, WordNet and GeoNames. It describes more than 10 million entities described by more than 120 million facts. However, most facts are type statements and *RDFS subClassOf* links. After indexing the Yago3 database consisted of 19,081,230 terms, 5,216,294 relations and 5,327,864 entities.
- **OpenCyc**[8] is an open-domain knowledge base. It is a freely available version of the Cyc database. The 4.0 release of OpenCyc used in our experiments includes 800K terms as lexical descriptions of 240K concepts. Overall, the knowledge base contains about 1 million triples.
- **UMLS**[3] Metatesaurus 2015-AA consists of more than 100 biomedical vocabularies and contains more than 800K biomedical concepts with millions of relations between them. These relations are mostly lexical relations (e.g., synonymy, meronymy) and not domain relations[9]. Each concept in the UMLS Metathesaurus is associated with a set of semantic types from the UMLS semantic network. In the scope of our approach, we need to have domain relations; we consider the semantic network classes (e.g. *Disease or Syndrome*, *Drug*) as entities instead of concepts, and the semantic network relations (e.g., *causes*, *treats*) as potential facts between the concepts. By extending the potential relations using the classes hierarchy (e.g., considering the potential link *type 2 diabetes, treats, antibiotics* from the general link *Drug, treats, Disease or Syndrome*, we obtain a set of 2,408 potential links that we use as knowledge base relations. After indexing, the UMLS database consists of 133 entities, 2,408 relations and 3,693,095 terms.
- **MeSH**[10] (Medical Subject Headings) is a hierarchically-organized terminology designed mainly for indexing biomedical information. We use the 2016 RDF version of MeSH[11] as a knowledge base for biomedical entities. After indexing, the MeSH database contains 792,775 terms for 348,278 concepts, and 953,640 relations.
- **Snomed-CT**[12] (Systematized Nomenclature of Medicine – Clinical Terms) is a standardized clinical vocabulary used by health professionals for the exchange of clinical health information. It encompasses 806,831 terms, 421,308 concept and 1,836,908 relations.

The statistics from each knowledge base are presented in Table 1[13].

We used relaxed position-based matching to evaluate the performance of our unsupervised NER method. Different corpora often adopt different criteria for named entity boundaries; for example, some may include adjectives and determiners, while others ignore them. These variations make exact named entity boundaries unsuitable for correlation studies. Table 2 presents the precision, recall and F1 score based on relaxed position-matching (values for exact matching F1 scores are 5% to 21% lower for individual corpora).

DBpedia outperformed the other open-domain knowledge bases on all open-domain corpora. This can be explained by the fact that DBpedia benefits from Wikipedia disambiguation pages and redirections, and consequently has a richer set of domain relations than YAGO and richer lexicalization than OpenCyc.

On the biomedical corpora, UMLS obtained the best recall but lower precision than DBpedia. The fact that UMLS has a better recall was expected due to its broader coverage for medical terms. The fact that DBpedia obtatined better precision than UMLS can be explained by the fact that DBpedia has lower ambiguity for medical terms, which enhances the quality of both TF-IDF based search and relational disambiguation. Aside from these two general behaviours, there are no noticeable regularities where some KBs do consistently better than others on different corpora.

We study the correlation between the proposed unsupervised evaluation metrics and the standard performance measures for NER, namely, Precision, Recall and F1 score. We use the Pearson's correlation factor, ρ, to study the correlation between the metrics and the performance measures. More precisely, $\rho_C(\mu, \alpha)$ is expressed as:

$$\frac{N\sum_i \mu_i(C)\times\alpha_i(C) - (\sum_i \mu_i(C)\sum_i \alpha_i(C))}{\sqrt{N\sum_i \mu_i(C)^2 - (\sum_i \mu_i(C))^2}\sqrt{N\sum_i \alpha_i(C)^2 - (\sum_i \alpha_i(C))^2}} \quad (12)$$

[6] http://ctdbase.org
[7] http://linkeddata.org
[8] http://sw.opencyc.org/
[9] http://www.ncbi.nlm.nih.gov/books/NBK9684/\#ch02.sec2.4
[10] https://www.nlm.nih.gov/mesh/
[11] https://id.nlm.nih.gov/mesh/
[12] https://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
[13] The number of relations here is considered to be the number of distinct subject-object pairs

KB	Domain	Entities	Relations	Labels	Density
DBpedia2014 [13]	Open	6,921,894	25,864,784	12,782,266	$5.39\ 10^{-7}$
YAGO3 [16]	Open	5,327,864	5,216,294	19,081,230	$1.83\ 10^{-7}$
OpenCyc [14]	Open	238,443	754,792	829,203	$1.32\ 10^{-5}$
UMLS Lite [3]	Biomedical	133	2,408	3,693,095	0.13
Snomed CT [15]	Biomedical	421,308	1,836,908	806,831	$1.03\ 10^{-5}$
MeSH [16]	Biomedical	348,278	953,640	792,775	$7.86\ 10^{-6}$

Table 1. Knowledge Bases

Corpus	DBpedia			YAGO			OpenCyc			UMLS			MeSH			Snomed-CT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TREC	**82.9**	**49.0**	**61.6**	80.3	29.9	43.5	66.7	30.3	41.6	–	–	–	–	–	–	–	–	–
NYT	**69.7**	**42.8**	**53.1**	68.5	27.9	39.7	48.9	20.5	28.9	–	–	–	–	–	–	–	–	–
WikiNER	**82.1**	**49.0**	**61.4**	81.3	31.7	45.6	63.2	26.5	37.3	–	–	–	–	–	–	–	–	–
CoNLL 03	**69.1**	**49.3**	**57.6**	65.7	31.4	42.5	48.5	31.4	38.1	–	–	–	–	–	–	–	–	–
AZDC	74.3	65.5	**69.6**	69.7	50.0	58.2	**78.4**	34.3	47.8	67.0	**70.2**	68.6	76.8	61.0	68.0	70.2	54.2	61.2
I2B2	65.3	31.5	42.5	56.8	16.0	24.9	51.1	23.7	32.3	60.5	**35.8**	**45.0**	**74.7**	20.6	32.3	53.6	15.3	23.8
NCBI	73.5	70.7	**72.0**	67.9	55.4	61.0	**77.9**	40.8	53.6	62.8	**72.4**	67.2	75.9	62.2	68.4	71.5	59.5	65.0
CDR	**74.1**	64.2	**68.8**	70.7	19.1	30.1	64.3	38.4	48.1	66.3	**64.9**	65.6	71.5	25.8	37.9	60.7	36.7	45.8
Berkeley04	71.3	65.1	**68.1**	64.5	38.1	47.9	59.3	39.9	47.7	62.6	**69.8**	66.0	**78.6**	50.2	61.3	67.0	39.9	50.0

Table 2. Precision (P), Recall (R) and F1 score for unsupervised NER. Best results are highlighted per row (corpus)

Where N is the number of knowledge bases, $\mu_i(C)$ represents the value of a performance measure (i.e., precision, recall, F1) when using knowledge base k_i to extract named entities from corpus C. $\alpha_i(C)$ represents the value of an annotation metric, α (e.g., graph adequacy, lexical adequacy) when knowledge base k_i is used to extract named entities from corpus C.

The values of the Pearson factor range from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating the absence of correlation. The Pearson factor is the most relevant for our study as it relies on the actual values of the variables, instead of their rank. A strong Pearson correlation would therefore suggest the portability of our method to additional corpora and knowledge bases. In contrast, the Spearman's rank correlation factor will not use the actual values of precision, recall and F1, but rather an integer rank value (e.g., a difference of 50% in F1 between 2 KBs could become equivalent to a difference of 1%), which does not allow assessing the scalability of the approach. Table 3 presents the Pearson correlation values for each metric.

4 Findings and Discussion

Our results show strong correlations between the proposed metrics and Recall, Precision and F1. Recall was naturally correlated with tokens coverage with 0.96 average Pearson factor on all corpora, 0.92 minimum correlation and 0.001 variance.

Precision was positively correlated with our Weighted Quality (WQ) metric with 0.59 average Pearson factor on all corpora, and 0.14 variance. One outlier behavior was observed for 2 corpora out of 9 (NCBI Disease corpus and Arizona Disease corpus), where WQ was not correlated with precision. These two corpora use the same document set. The NCBI corpus annotation extends AZDC annotations by annotating all sentences. When we studied these two corpora more closely with an error analysis, we observed that a disease name is often annotated only once in an abstract even if it occurs multiple time, which leads to a random behaviour for precision. From this perspective, our method was able to detect the bias of the manual annotation. If these two corpora are not included, the correlation of WQ with precision reaches an average of 0.76 with a variance of 0.001.

F1 scores were strongly correlated with the $OPERA$ metric ($WQ \times D \times V$) on all corpora: average correlation of 0.96, minimum 0.88 and variance of 0.001, which follows the observations on the elementary metrics; i.e., WQ is a strong predictor of precision and V is a strong predictor of recall.

Our evaluation metric outperformed the TF-IDF baselines by +0.09 Pearson value for F1 and recall and +0.21 Pearson value for precision. The TF-IDF baseline had also a high variance of 0.22 for an average correlation value of 0.41, which shows that the quality of lexical matches does not provide a reliable indicator of precision. Our Lexical Disambiguation Ratio (L) had the best correlation for precision (0.62) with a relatively low variance of 0.107. L indicates the difficulty of the disambiguation problem as it represents the number of non-ambiguous mentions. This also shows that ambiguity, which is derived from TF-IDF scores, is more useful than the raw TF-IDF

	Precision			Recall			F1		
Correlation	Average	Range	Variance	Average	Range	Variance	Average	Range	Variance
Baselines									
TF-IDF	+ 0.41	[-0.43, +0.86]	0.217	+ 0.180	[-0.12, +0.62]	0.060	+ 0.26	[-0.01, +0.60]	0.04
$TF - IDF \times V$	+ 0.32	[-0.71, +0.99]	0.409	+0.87	[+0.76, +0.94]	0.002	+ 0.87	[+0.76, +0.96]	0.004
Composite Metrics									
$OPERA$ ($WQ \times D \times V$)	+ 0.27	[-0.52, +0.81]	0.196	0.96	[+0.90, +0.99]	0.001	**+ 0.96**	**[+0.88, +0.99]**	**0.001**
Weighted Quality (WQ)	+ 0.59	[+0.00, +0.99]	0.146	+ 0.22	[-0.23, +0.59]	0.063	+ 0.32	[-0.10, +0.75]	0.071
Disambiguation Coverage ($D \times V$)	- 0.08	[-0.73, +0.53]	0.170	+ 0.88	[+0.70, +0.95]	0.005	+ 0.80	[+0.53, +0.93]	0.014
Lexical Adequacy ($L \times V$)	+ 0.30	[-0.49, +0.85]	0.221	+0.94	[+0.81, +0.99]	0.004	+ 0.94	[+0.79, +0.99]	0.004
Graph Adequacy ($R \times V$)	-0.51	[-0.99, +0.56]	0.224	+ 0.23	[-0.79, +0.95]	0.31	+ 0.10	[-0.85, +0.90]	0.31
Elementary Metrics									
Tokens Coverage V	+ 0.06	[-0.80, +0.73]	0.29	**+ 0.96**	**[+0.92, +0.99]**	$7e^{-4}$	+ 0.90	[+0.78, +0.99]	0.005
Average Ambiguity	- 0.61	[-0.99, +0.58]	0.229	-0.02	[-0.66, +0.96]	0.227	- 0.13	[-0.85, +0.91]	0.23
Disambiguation Ratio (D)	-0.20	[-0.54, +0.26]	0.04	+ 0.65	[+0.31, +0.84]	0.02	+ 0.57	[+0.10, +0.84]	0.03
Lexical Ratio (L)	**+0.62**	**[-0.05, +0.99]**	**0.107**	+0.21	[-0.66, +0.95]	0.26	+ 0.31	[-0.55, +0.98]	0.23

Table 3. Range, Variance and Average Pearson correlation factors between annotation metrics and Precision, Recall and F1 score. Best results are highlighted, second best are underlined.

values.

Including the Disambiguation Ratio (d) in the formula of $OPERA$ ($WQ \times d \times V$) led to a better correlation for F1 score. We also observe that Disambiguation Coverage ($d \times V$) is less correlated for recall than V (0.90 vs 0.80 Pearson values, respectively) . From additional experiments, we also found that $WQ \times V$ had an average correlation of 0.88 only (compared to 0.96 with $WQ \times d \times V$). Therefore, we can conclude that the elementary metric d had a positive impact for the prediction of precision values.

Our study is not exhaustive with regards to the number of metrics that might be considered. However, our results show that the proposed, general annotation metrics can predict, to a large extent, the adequacy of a KB for annotating named entities in a given corpus.

We limited our named entity recognition approach to commonly used methods. TF-IDF scores and global coherence maximization with KB relations are used in many related studies, including supervised classification approaches [19, 7, 11]. Therefore, we think that our observations can benefit other named entity recognition methods, provided that they use these two general principles.

We have no evidence at the current stage that our evaluation metrics would be relevant for recognition methods that do not rely on global coherence and TF-IDF scores. This includes token classification methods such as conditional random fields, which rely primarily on annotated corpora. A potential future direction is to use our unsupervised annotations to provide KB-derived semantic features at token level to study the performance of supervised classifiers.

Another potential future direction is to extend our method to evaluation of training corpora for supervised classification. In this setting, a training corpus can be seen as a knowledge base where the manual annotations are knowledge base entities and the co-occurrences of two annotations in the same sentence or context indicate the knowledge base relations.

5 Conclusions

We presented a new ranking approach to assess the suitability of knowledge bases for named entity recognition in a given corpus. More precisely, we proposed several unsupervised annotation metrics and studied their correlation with performance measures such as precision, recall and F1 score. Our results show that these metrics can be strong predictors of NER performance and that they significantly improve ranking relevance when compared to TF-IDF baselines. With the important increase in scope of named entities, our approach can play a key role in large-scale NER as it allows selecting relevant knowledge bases for a given textual context. It can also be applied to the selection of sub-graphs from the same (large) knowledge base. Our short-term goal is to deploy a web service that takes natural language texts as input and ranks all indexed knowledge bases according to the performance predictors proposed in this paper. This also includes the integration of other KBs deemed to be of sufficient interest. We also plan to study the performance of supervised classifiers according to these metrics when they use unsupervised knowledge base annotations as training features.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

REFERENCES

[1] Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran, 'Named entity recognition in wikipedia', in *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 10–18. Association for Computational Linguistics, (2009).

[2] Krisztian Balog, Edgar Meij, and Maarten De Rijke, 'Entity search: building bridges between two worlds', in *Proceedings of the 3rd International Semantic Search Workshop*, p. 9. ACM, (2010).

[3] Olivier Bodenreider, 'The unified medical language system (umls): integrating biomedical terminology', *Nucleic acids research*, **32**(suppl 1), D267–D270, (2004).

[4] Janez Brank, Marko Grobelnik, and Dunja Mladenic, 'A survey of ontology evaluation techniques', in *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, pp. 166–170, (2005).

[5] Marc Bron, Krisztian Balog, and Maarten De Rijke, 'Example based entity search in the web of data', in *Advances in Information Retrieval*, 392–403, Springer, (2013).

[6] Rezarta Islamaj Doğan and Zhiyong Lu, 'An improved corpus of disease mentions in pubmed citations', in *Proceedings of the 2012 workshop on biomedical natural language processing*, pp. 91–99. Association for Computational Linguistics, (2012).

[7] Paolo Ferragina and Ugo Scaiella, '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*, pp. 1625–1628. ACM, (2010).

[8] Aldo Gangemi, Carola Catenacci, Massimiliano Ciaramita, and Jos Lehmann, 'A theoretical framework for ontology evaluation and validation', in *SWAP*, volume 166. Citeseer, (2005).

[9] Asunción Gómez-Pérez, 'Evaluation of ontologies', *International Journal of intelligent systems*, **16**(3), 391–409, (2001).

[10] Younggyun Hahm, Jungyeul Park, Kyungtae Lim, Youngsik Kim, Dosam Hwang, and Key-Sun Choi, 'Named entity corpus construction using wikipedia and dbpedia ontology.', in *LREC*, pp. 2565–2569, (2014).

[11] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti, 'Collective annotation of wikipedia entities in web text', in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 457–466. ACM, (2009).

[12] Robert Leaman, Christopher Miller, and G Gonzalez, 'Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark', in *Proceedings of the 2009 Symposium on Languages in Biology and Medicine*, volume 82, (2009).

[13] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, et al., 'Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia', *Semantic Web Journal*, **5**, 1–29, (2014).

[14] Xin Li, Paul Morie, and Dan Roth, 'Identification and tracing of ambiguous names: Discriminative and generative approaches', in *Proceedings of the National Conference on Artificial Intelligence*, pp. 419–424. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, (2004).

[15] Adolfo Lozano-Tello and Asunción Gómez-Pérez, 'Ontometric: A method to choose the appropriate ontology', *Journal of database management*, **2**(15), 1–18, (2004).

[16] Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek, 'Yago3: A knowledge base from multilingual wikipedias', in *7th Biennial Conference on Innovative Data Systems Research*. CIDR Conference, (2014).

[17] Yassine Mrabet, Claire Gardent, Muriel Foulonneau, Elena Simperl, and Eric Ras, 'Towards knowledge-driven annotation', in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pp. 2425–2431, (2015).

[18] Fabian Neuhaus, Amanda Vizedom, Ken Baclawski, Mike Bennett, Mike Dean, Michael Denny, Michael Grüninger, Ali Hashemi, Terry Longstreth, Leo Obrst, et al., 'Towards ontology evaluation across the life cycle', *Applied Ontology*, **8**(3), 179–194, (2013).

[19] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson, 'Local and global algorithms for disambiguation to wikipedia', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 1375–1384. Association for Computational Linguistics, (2011).

[20] Barbara Rosario and Marti A Hearst, 'Classifying semantic relations in bioscience texts', in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 430. Association for Computational Linguistics, (2004).

[21] Dan Roth and Wen-tau Yih, 'Global inference for entity and relation identification via a linear programming formulation', *Introduction to statistical relational learning*, 553–580, (2007).

[22] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum, 'Large-scale cross-document coreference using distributed inference and hierarchical models', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 793–803. Association for Computational Linguistics, (2011).

[23] Erik F. Tjong Kim Sang and Fien De Meulder, 'Introduction to the conll-2003 shared task: Language-independent named entity recognition', in *Proceedings of CoNLL-2003*, eds., Walter Daelemans and Miles Osborne, pp. 142–147. Edmonton, Canada, (2003).

[24] Maksim Tkachenko and Andrey Simanovsky, 'Named entity recognition: Exploring features.', in *KONVENS*, pp. 118–127, (2012).

[25] Felix Tristram, Sebastian Walter, Philipp Cimiano, and Christina Unger, 'Weasel: a machine learning based approach to entity linking combining different features', in *Proceedings of 3th International Workshop on NLP and DBpedia, co-located with the 14th International Semantic Web Conference (ISWC 2015), October 11-15, USA*, (2015).

[26] Ellen M. Voorhees, 'Overview of the TREC 2002 question answering track', in *Proceedings of The Eleventh Text REtrieval Conference, TREC 2002, Gaithersburg, Maryland, USA, November 19-22, 2002*, eds., Ellen M. Voorhees and Lori P. Buckland, volume Special Publication 500-251. National Institute of Standards and Technology (NIST), (2002).

[27] Denny Vrandečić, *Ontology evaluation*, Springer, 2009.

[28] Cristofer Weber and Renata Vieira, 'Building a corpus for named entity recognition using portuguese wikipedia and dbpedia', in *I Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish*, pp. 9–15, (2014).

[29] Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wiegers, and Zhiyong Lu, 'Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task', *Database*, **2016**, (2016).

[30] Christopher A Welty, Ruchi Mahindru, and Jennifer Chu-Carroll, 'Evaluating ontological analysis', in *Semantic Integration Workshop (SI-2003)*, p. 92. Citeseer, (2003).

[31] Haining Yao, Anthony Mark Orme, and Letha Etzkorn, 'Cohesion metrics for ontology design and application', *Journal of Computer science*, **1**(1), 107–113, (2005).

[32] Uzuner , South BR, Shen S, and DuVall SL, '2010 i2b2/va challenge on concepts, assertions, and relations in clinical text', in *Journal of the American Medical Informatics Association: JAMIA*, volume 18, pp. 552–556, (2011).