Effective Named Entity Recognition for Idiosyncratic Web Collections

Roman Prokofyev, Gianluca Demartini, and Philippe Cudré-Mauroux
eXascale Infolab
University of Fribourg
Switzerland
{firstname.lastname}@unifr.ch

ABSTRACT

Named Entity Recognition (NER) plays an important role in a variety of online information management tasks including text categorization, document clustering, and faceted search. While recent NER systems can achieve near-human performance on certain documents like news articles, they still remain highly domain-specific and thus cannot effectively identify entities such as original technical concepts in scientific documents. In this work, we propose novel approaches for NER on distinctive document collections (such as scientific articles) based on n-grams inspection and classification. We design and evaluate several entity recognition features—ranging from well-known part-of-speech tags to n-gram co-location statistics and decision trees—to classify candidates. In addition, we show how the use of external knowledge bases (either specific like DBLP or generic like DBPedia) can be leveraged to improve the effectiveness of NER for idiosyncratic collections. We evaluate our system on two test collections created from a set of Computer Science and Physics papers and compare it against state-of-the-art supervised methods. Experimental results show that a careful combination of the features we propose yield up to 85% NER accuracy over scientific collections and substantially outperforms state-of-the-art approaches such as those based on maximum entropy.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*; I.7.m [**Document and Text Processing**]: [Miscellaneous]

General Terms

Algorithms, Experimentation

Keywords

named entity recognition, text mining, term recognition.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW '14
ACM 978-1-4503-2744-2/14/04.
http://dx.doi.org/10.1145/2566486.2568013.

1. INTRODUCTION

While recent approaches to online Named Entity Recognition (NER) have become quite efficient and effective, they still do not perform equally well on all domains, leaving out some application scenarios from entity-centric information access. For highly-specialized domains such as academic literature, online information systems performing search, bookmarking, or recommendations are still organized around documents mostly. This is due to the fact that identifying entities (e.g., concepts) in specific collections such as scientific articles is more difficult than, say, in online news articles due to the *novelty* (i.e., new terms may be used which have not been previously observed in any other document/dictionary) and *specificity* (i.e., highly technical and detailed formalisms mixed with narrative examples) of the content.

While retrieving documents in an entity-centric fashion would also be beneficial for specialized domains, the difficulty of correctly extracting highly-specialized entities as well as the scarcity of semi-structured information available for specific documents are precluding such advances. As an example, the ScieceWISE portal[1] [1] is an ontology-based system for bookmarking and recommending papers for physicists. ScieceWISE is entity-centric, yet it requires human intervention to correctly extract the scientific concepts appearing in each new paper uploaded onto the system.

In this paper, we tackle the problem of NER in highly-specialized domains such as scientific disciplines. We develop new techniques to identify all relevant n-gram concepts appearing in a scientific document, based on a set of features including n-gram statistics, syntactic part-of-speech patterns, and semantic techniques based on the use of external knowledge bases. In addition, we effectively combine our various features using a state-of-the-art machine learning approach in order to get the most out of our different families of features. The results of our NER approach can then be used for many applications, including to organize data on search engine results pages, to summarize scientific documents, or to provide faceted-search capabilities for literature search.

We experimentally evaluate the effectiveness of our methods over two manually-judged collections of scientific documents: a collection of papers from SIGIR 2012 (a well-known scientific conference on Information Retrieval), and a sample of research papers retrieved from `arXiv.org`. Our

[1] `http://www.sciencewise.info`

experimental results show how semantic-aware features overcome simple text-based features and how a combination of our proposed features can reach up to 85% overall Accuracy, significantly improving over state-of-the-art domain-specific supervised approaches based on maximum entropy [9]. In summary, the main contributions of this paper are as follows:

- We tackle the problem of NER in the challenging context of idiosyncratic collections such as scientific articles.
- We describe a new, multi-step candidate selection process for named entities favoring recall (as standard techniques perform poorly in our context) and based on co-location statistics.
- We propose novel NER techniques based on semantic relations between entities as found in domain-specific or generic third-party knowledge bases.
- We extensively evaluate our approach over two different test collections covering different scientific domains and compare it against state-of-the-art NER approaches.
- We identify an effective combination of both syntactic and semantic features using decision trees and apply them on our collections, obtaining up to 85% Accuracy.

The rest of the paper is organized as follows: We start with an overview of related work in the areas of named-entity recognition, keyphrase extraction, and concept extraction below in Section 2. We describe our overall system architecture and its main features (including PDF extraction, n-gram lemmatization, part-of-speech tagging, external knowledge bases, and n-gram ranking) in Section 3. Section 4 provides definitions of our ranking features. Section 5 describes our experimental setting and presents the results of a series of experiments comparing different combinations of features. Finally, we conclude and discuss future work in Section 6.

2. RELATED WORK

Named entity recognition (NER) designates the task of correctly identifying words or phrases in textual documents that express names of entities such as persons, organizations, locations, etc. During the last decades, NER has been widely studied and the best NER approaches nowadays produce near-human recognition accuracy for generic domains such as news articles. Several prominent NER systems use either hand-coded rules or supervised learning methods such as maximum entropy [4] or conditional random fields [10]. These methods heavily rely on large corpora of hand-labeled training data, which are generally-speaking hard to produce. Besides the high costs associated to the manual annotation of the training data, this also raises the problem of domain-specificity; For instance, models trained for news articles are most likely to perform well on such documents only [24].

In that context, there has been a lot of attention given to NER applied to newswire text (mostly because of the high quality of such texts), focusing on entity types such as location, person, and company names. On the other hand, the task of NER for more domain-specific collections, e.g., for scientific or technical collections, remains largely unexplored, with a few exceptions including the biomedical domain where previous work has focused on specific entity types like genes, protein and drug names [28, 9]. In this paper we focus on semantic-based NER over such domain-specific collections.

Open Information Extraction.

To address some of the above issues, researchers have recently focused on Web-scale NER (also known as Open Information Extraction) using automatic generation of training data [31], unsupervised NER based on external resources such as Wikipedia and Web n-gram corpora [18], and robust NER performance analysis across domains [26]. In this area, information extraction at scale is run over the Web to find entities and factual information to be represented in structured form [32, 6]. Instead, we focus on well-curated and highly-technical textual content. Compared to previous work in NER, we focus on effectively performing NER on domain-specific collections like technical articles. We apply state-of-the-art techniques together with specific approaches for scientific documents including the use of domain-specific knowledge bases to improve the quality of NER at a level comparable to the one achieved for news documents.

Key Term Extraction.

Another task related to this paper is *key term extraction*. Key term extraction deals with the extraction and ranking of the most important phrases in a text. This can be used, for instance, in text summarization or tagging [3]. In [15], authors address this task as a ranking problem rather than a classification task. Contrary to NER research, many approaches in the area of key term extraction deal with technical and scientific document collections. Some recent evaluation competitions such as [16] are specifically geared towards scientific articles. Although the Precision of the top-performing systems is typically around 40% for such competitions, these results can be considered as rather high due to the specificity of the terms appearing in the scientific documents and the rather subjective nature of the ground-truth in that context. At this point, we want to emphasize that key phrases extraction is different from the task we address in this paper, which aims at identifying *all* possible entities in a document to enable further entity-centric processes (e.g., in the search engine).

The candidate identification step of term extraction systems typically filters all of the possible n-grams from the documents by frequency, retaining high frequency n-grams only. Some methods use hand-coded part-of-speech tag patterns to provide additional filtering [30, 12], though hand-coded tag patterns are not always able to capture the variety of all valid entities due to tagging ambiguity (i.e., the same term may be considered either as a verb or as an adjective depending on the context). Instead, in our work we use standard frequency filtering with a re-weighting step to identify as many candidates as possible and part-of-speech tags as a feature to boost both Precision and Recall of NER.

The majority of keyphrase extraction studies use supervised models, the most commonly used approaches being naive Bayes [30, 11], decision trees [30] and support vector machines [17]. In our work, we use a decision tree-based classifier since it is able to handle easily both numerical and categorical data with little data preprocessing. Decision trees

are also simple to interpret by the end-users who are the authors of scientific papers. Specifically, we base our work on a decision tree model and ensemble methods for feature selection using extremely randomized trees [13].

We also note that the work we present in this paper actually lies in between the NER and key term extraction tasks. In standard NER, the goal is to identify all named entities mentioned in a document while in key term extraction the goal is to identify the most representative terms in a document. The task we address in this paper is rather to identify the subset of named entities that are *valid* for the given idiosyncratic documents considered (see also our examples in Section 3.1).

Entity Linking.

Some previous work successfully used Wikipedia or DBPedia to identify significant terms in textual documents [22, 20, 7]. However, such methods operate only on the entities that already exist in the knowledge bases. The task of identifying entity mentions given a background corpus of entities is also known as *Entity Linking*. On the other hand, our goal is to also discover *new* entities from scientific documents, potentially by leveraging generic-purpose or specific knowledge bases.

Also related to this paper is the task of ad-hoc object retrieval [25, 29], that is, the task of identifying an entity in a background entity corpus given its textual description in a document. The task of NER that we address in this paper is a necessary step to enable entity search, which let users find information in an entity centric fashion.

3. SYSTEM OVERVIEW

3.1 Problem Definition

The task we address is the identification of all valid entities related to a given domain in a domain-specific collection. In the context of this paper, we define a *valid entity* as an n-gram representing a relevant concept of a scientific domain and not just as any real-world object. To give a clearer understanding of what a valid entity is in our case, let us look at a few examples. Consider the n-gram "Saving Private Ryan". Usually such a string represents a valid entity referring to a popular movie, but it does not make much sense to mark this n-gram as valid in an Information Retrieval paper, where it was given as a query example. Another example illustrating the complexity of our task comes from disambiguation decisions. Consider the n-gram "large numbers"; It can be a valid entity in document is talking about *large numbers* in a pure mathematical sense, but in many other cases it is just a linguistic construction.

To assess the performance of our approach, we use a standard set of evaluation metrics: Precision, Recall, F1 score, and Accuracy, which are computed on a per document basis (i.e., each item in our test collection is represented by a pair *(document, n-gram)*). These metrics allow us to show how well an approach performs both on true positives and true negatives and to discuss the resulting trade-offs.

In this work, we exclusively focus on the identification of n-grams entities with $n > 1$ because of the high level of inherent ambiguity that unigrams have in scientific literature. Many unigrams are ambiguous and can often be used both as entities and non-entities, even when inspecting a single document. Moreover, we argue that most unigram entities are very generic and can be recognized by simple dictionary lookups[2]. Thus, techniques like *entity linking* [7, 8] are in our opinion more suitable to address unigram entity recognition.

3.2 Framework

To evaluate our proposed approach, we have built a system that takes as input a set of scientific documents in PDF format and returns as output the set of n-grams appearing in the text of the documents that represent scientific concepts. Figure 1 below gives an overview of the architecture of our system.

The first components in our pipeline extract text from the input documents and perform some automatic preprocessing (e.g., lemmatization). The following steps consist in identifying the candidate entities that are potentially relevant concepts. The candidate selection step focuses on high Recall while keeping the number of candidate n-grams orders of magnitude lower than the total number of n-grams in a document. Finally, we use a series of approaches to select the valid n-grams among the candidates (focusing on high Precision). We discuss this pipeline in more detail in the following.

3.3 Data Preprocessing

Our system receives PDF documents as input and transforms them into raw text using an open-source library[3]. We then perform a series of preprocessing steps; First, we lowercase all words (except acronyms) appearing at the beginning of sentences to prevent duplicate entity creation in the latter steps. At this point, we make a separate copy of the resulting text (before lemmatization) on which we apply Part-Of-Speech (POS) tagging.

The first copy of the text is then lemmatized, using the a lemmatization approach based on WordNet [21]. We have opted for lemmatization in our context since the other typical possibility, stemming, is too aggressive on scientific documents as it often conflates scientific concepts which should be kept distinct.[4] In the final step, we build an *n-gram index* from the resulting text to efficiently perform the candidate selection phase described below.

3.4 Candidate Selection

The goal of the candidate selection step is to extract as many candidate entities as possible from the scientific articles, while limiting the number of false positives. To achieve this goal, we extend techniques based on word co-locations [19]. First, we extract from the n-gram index all *bigrams* having a frequency (i.e., number of occurrences in the input document) greater than a threshold k (e.g., $k = 2$). Next, the extracted bigrams are joined together into trigrams; Two bigrams are joined if and only if it is possible to merge them to form a valid trigram (i.e., if the same word ends a bigram and starts another one). The resulting trigram frequency is then looked-up from the n-gram index.

[2]an analysis of the tags verified by the users of the ScienceWISE platform shows that 23% of the tags are unigrams, and only 5% of them (1% overall) are not found in Wikipedia.

[3]We use Apache Tika `http://tika.apache.org/` for this task.

[4]We have performed our extraction experiments without any lemmatization and found that this reduces Recall by 4%.

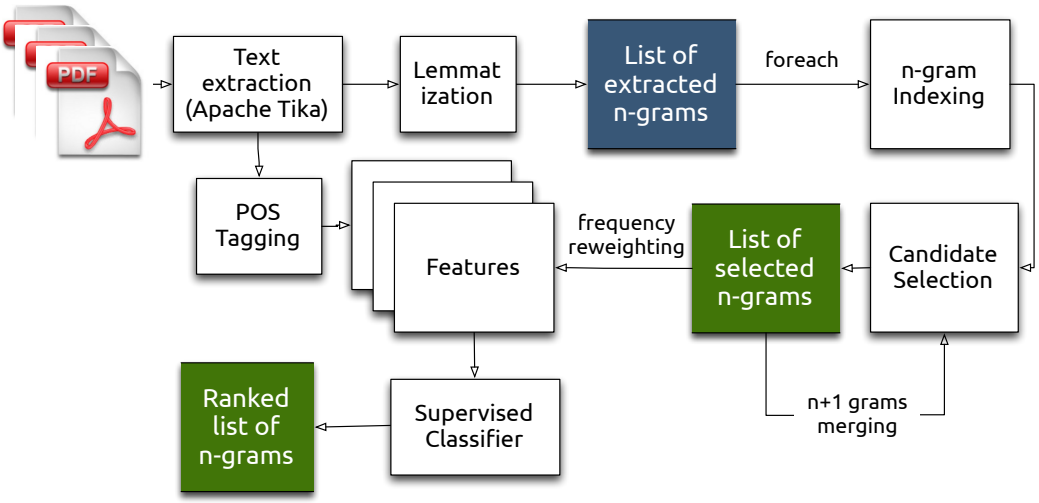

Figure 1: Processing pipeline. First, the plain text is extracted from the PDF documents. Then, the text is pre-processed using lemmatization and POS tagging. Candidate n-grams are generated and indexed. Then, n-grams are selected based on a predefined set of features (see Section 3.4). Finally, a supervised approach (e.g., decision trees) is responsible to generate a ranked list of n-grams that have been identified as valid entities in the Web documents.

This process is repeated for trigrams, up to the maximal n-gram size N considered ($N = 5$ in our experiments as for $N > 5$ we could not identify valid concepts in our test collections). The difference between simply restricting the frequency of any n-gram to k and our approach is that we can extract n-grams with a frequency lower than k: As can be seen on the graph of n-gram occurrence distribution depicted in Figure 2), there are many valid n-grams in the collection that appear just once or twice in the text, and removing them with a frequency threshold would result in a sharp decrease in Recall. Hence, after processing every document, we regroup the extracted n-grams from the entire collection and look them up again in every document. This process preserves n-grams that passed the frequency threshold k in some papers, but not in others.

This collection-wide n-gram selection approach results in an increase of Recall from 42.2% to 96.1%. Alternatively, we also tried two further approaches: using the collection-level n-gram frequencies to serve as a cutoff frequency k, and running the n-gram merging process from scratch after adding collection-wide n-grams. These approaches yielded Recall values of 87.4% and 93.2% respectively.

Removing Incomplete N-Grams.

In the last step, we apply a frequency reweighing process that takes into account the fact that some n-grams appear as part of other n-grams. We illustrate our reweighing mechanism by an example. Assume that in a document two bigrams "latent dirichlet" and "dirichlet allocation" appear both with frequency f, and that a trigram "latent dirichlet allocation" also appears with the same frequency. It is safe to say in that case that those two bigrams do not appear in the text as separate entities, but only as part of a bigger trigram. Our process hence starts from the longest n-grams (i.e., from n-grams with larger n), and proportionally decrements the frequency of the shorter n-grams that

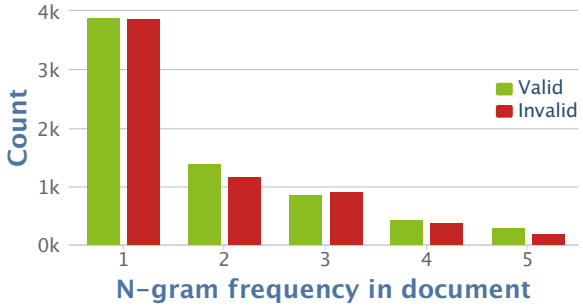

Figure 2: Valid/invalid n-gram count distribution for the SIGIR collection. Only the first 5 frequencies are shown.

are subsumed by it. At the end of this process, we eliminate all n-grams having a re-weighted frequency equal to zero.

3.5 Supervised N-Gram Selection

Rather than simply weighting different features in order to determine whether an n-gram represents a correct concept or not, one can use machine learning approaches to learn to identify correct entities. In this paper, we construct a feature space consisting of the features presented in Section 4 and use a manually extracted set of entities appearing in scientific documents as training data for a decision tree classifier [13]. Once trained, the classifier is then able to take as input a new document and—thanks to the processing pipeline depicted in Figure 1—to effectively select all valid scientific concepts from the document.

4. FEATURES FOR NER

In this section, we describe the five different families of features used by our system to detect named entities in scientific Web documents. We propose different families of fea-

tures ranging from simple syntactic POS patterns to features using third-party resources such as external knowledge bases and structured repositories like DBLP[5]. We also propose to combine our features using machine learning approaches. More specifically, we use decision trees to decide which n-grams correspond to valid concepts in the documents. This also allows us to understand which features are the most valuable in our context based on a hierarchy generated by our learning component.

4.1 Part-of-Speech Tags

Part-Of-Speech (POS) tags have often been considered as an important discriminative feature for term identification. Many works on key term identification apply either fixed or regular expression POS tag patterns to improve their effectiveness. Nonetheless, POS tags alone cannot produce high-quality results. As can be seen from the overall POS tag distribution graph extracted from one of our collections (see Figure 3), many of the most frequent tag patterns (e.g., $JJ\,NN$ tagging adjectives and nouns[6]) are far from yielding perfect results.

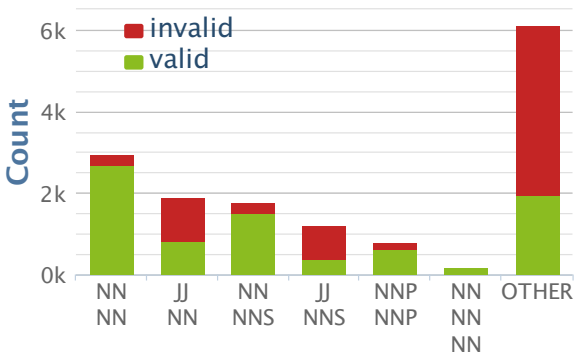

Figure 3: Top 6 most frequent part-of-speech tag patterns of the SIGIR collection, where JJ stands for adjectives, NN and NNS for singular and plural nouns, and NNP for proper nouns.

Given those results, we designed several features based on POS tags that might perform better than predefined POS patterns. First, we consider raw POS tags where each POS tag pattern represents a separate binary feature. Though raw POS tags can provide a good baseline in some settings, we do not expect them to perform well in our case because of the large variety of POS tag patterns in both collections, many of which can be overly specific.

A more appealing choice is to group (or *compress*) several related POS tag patterns into one aggregated pattern. We use two grouping techniques: Compressing all POS tag patterns by only taking into account i) the first or ii) the last POS tag in the pattern. Using the compressed POS tag versions, we significantly reduce the feature space, which is the key to achieve higher performance and allows for model generalization. We discuss those two schemes in more detail in Section 5.2. To perform POS tagging, we used a standard approach based on maximum entropy [27].

[5] `http://dblp.dagstuhl.de/`

[6] see `http://www.cis.upenn.edu/~treebank/` for an explanation on POS tags

4.2 Near n-Gram Punctuation

Another potentially interesting set of features closely related to POS tags is punctuation. Punctuation marks can provide important linguistic information about the n-grams without resorting to any deep syntactic analysis of the phrase structure. For example, the n-gram *"new summarization approach based"*, which does not represent any valid entity, has a very low probability of being followed by a dot or comma, while the n-gram *"automatic music genre classification"*, which is indeed a valid entity, often appears either at the beginning or at the end of a sentence.

The contingency tables given in Table 1 and Table 2 illustrate this: The *+punctuation* and *-punctuation* rows show, respectively, the counts of the n-grams that have at least one punctuation mark in any of its occurrences and the counts of the n-grams that have no punctuation mark in all their occurrences. From the tables, we observe that the presence of punctuation marks (+punctuation) either before or after an n-gram occurs twice as often for the n-grams that are valid entities compared to the invalid ones. We also observe that the absence of punctuation marks after an n-gram happens less frequently for the valid n-grams than for the invalid ones.

Table 1: Contingency table for punctuation marks appearing *immediately before* the n-grams.

	Valid	Invalid	Total
+punctuation	1622	847	2469
−punctuation	6523	6065	12588
Totals	8145	6912	15057

Table 2: Contingency table for punctuation marks appearing *immediately after* the n-grams.

	Valid	Invalid	Total
+punctuation	4887	2374	7261
−punctuation	3258	4538	7796
Totals	8145	6912	15057

Thus, both directly preceding and following punctuation marks are able to provide relevant information on the validity of the n-grams and can be used as binary features for NER.

4.3 Domain-Specific Knowledge Bases: DBLP Keywords and Physics Concepts

DBLP is a website that tracks and maintains bibliographic references for the majority of computer science journals and conference proceedings. The structured meta-data of its records include high quality keywords that authors assign to their papers.

Author-assigned keywords represent a very reliable source of named entities for documents related to this specific domain. In fact, the overall Precision of n-grams from author-assigned keywords for our computer science dataset is 95.5% (with 27.4% Recall), and hence can be used as a highly discriminative feature.

While DBLP provides high quality annotations for computer science documents, there is no such knowledge base

for our physics collection. Thus, we decided to perform a similar matching using the concepts from one of the largest physics ontology available—the ScienceWISE ontology[7]. All the concepts in this ontology represent valid named entities which, as for DBLP, can be used as a highly discriminative feature.

4.4 Wikipedia/DBPedia Relation Graphs

Wikipedia is by far the largest general-purpose knowledge-base currently available. In the context of our task, Wikipedia exhibits the following valuable features[8]:

- The majority of pages in Wikipedia represent valid named entities.
- Pages are interconnected with each other through links appearing in the page body and through their categories.
- Many pages have alternative spellings which are encoded by a special "redirects" property.

We base our Wikipedia features on collection statistics. Specifically, we use a machine-processable version of Wikipedia called DBPedia[9], which contains all entities in Wikipedia described in a structured format and interconnected to other datasets. We start by computing the Precision and Recall values when matching Wikipedia pages with the n-grams from our collections. Table 3 shows the resulting values for two cases: i) exact string matching with page title and ii) matching allowing variants based on the "redirects" property. As expected, we observe that allowing flexible matching with redirects results in a significant growth in Recall, with some loss in Precision[10].

Furthermore, taking into consideration the relatively low Precision of exact Wikipedia matchings, one can try to improve the above technique by finding further methods to separate the valid entities from the invalid ones. Hulpus *et al.* [14] recently observed that interlinked Wikipedia pages are much more likely to form a connected component in the Wikipedia category graph than random pages. Given that finding, we use the size of the connected component a Wikipedia page belongs to as an additional feature for valid concepts.

Following the approach in [14], we construct the neighboring page graph by following relationships in DBPedia of types {*broader,subject,related*} for up to two hops in both directions. The two hops threshold was chosen based on previous research from [14], which claimed that bigger distances result in much larger graphs and introduce noise. The Wikipedia administrative categories and pages referring to etymology (e.g., "English phrases") are excluded using an existing list of stop URIs[11].

Figure 4 shows how often the connected component of a given size contains more valid than invalid entities, while Figure 5 shows the average percentages of valid and invalid entities in a component of a given size. We observe that larger connected components tend indeed to contain more valid entities than smaller ones.

Based on the analysis made above, we construct the following set of NER features using relation graphs:

- *is wiki*: whether a candidate n-gram can be exactly matched to a Wikipedia page title,
- *is redirect*: whether a candidate n-gram can be matched using an alternative spelling of a Wikipedia page,
- *component size*: the size of the connected components an n-gram belongs to, constructed with and without the redirect property,
- *component+DBLP*: a binary feature, equal to 1 when an n-gram appears in the same connected component with at least one DBLP keyword, and to 0 otherwise;
- *wikilinks*: the number of outgoing links in the Wikipedia page body to other Wikipedia pages.

4.5 Syntactic Features

In addition to the features described above, we also test a series of more common syntactic features that are often used by other NER classifiers, including:

- the *n-gram length* in words,
- whether the n-gram is uppercased
- the number of other n-grams the given n-gram is part of in the document.

5. EXPERIMENTAL EVALUATION

5.1 Experimental Setting

In this section, we empirically evaluate the NER techniques proposed above. We evaluate the quality of our features as well as how to best combine them over two distinct test collections for which ground truth entity annotations have been manually created by domain experts from two specific domains: Computer Science and Physics.

Dataset Description.

Our first dataset contains 100 randomly selected papers taken from the SIGIR 2012 conference proceedings, while our second dataset contains the same number of recent (2012) articles taken from the High Energy Physics (hep-ph) section from the `arXiv.org` pre-print repository.

Our system extracted 21,531 candidate n-grams in total from the first dataset, of which 8,814 n-grams were unique. Overall, 15,057 n-grams were judged, of which 8,145 were labeled as valid and 6,912 as invalid.

In the second dataset, our system extracted 18,129 candidate n-grams, of which 7,880 n-grams were unique. Overall, 11,421 n-grams were judged, of which 5,747 were labeled as valid and 5,674 as invalid[12].

The judgments were performed on a per-document basis, meaning that an n-gram was considered as a relevant scientific concept if it represented a valid entity in the scope of a

[7] `http://data.sciencewise.info`

[8] Every feature in the above list is freely accessible through the Wikipedia API at `http://en.wikipedia.org/w/api.php`.

[9] `http://dbpedia.org/`

[10] Though Precision for the physics collection actually goes up, most likely because of the very low number of n-grams exactly matching—only about 60 cases.

[11] `http://uimr.deri.ie/sites/StopUris`

[12] Both datasets and ground truth data are made available for online exploration and download at `http://exascale.info/iNER`.

Table 3: Precision/Recall values for Wikipedia features.

	SIGIR		Physics	
	Precision	Recall	Precision	Recall
String matching	0.9045	0.2394	0.7063	0.0155
Matching with redirects	0.8457	0.4229	0.7768	0.5843

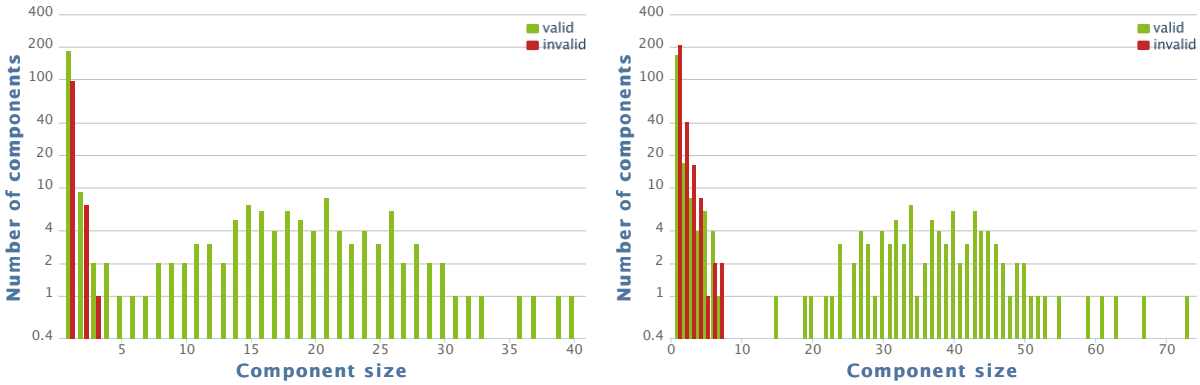

Figure 4: DBPedia connected component sizes for valid/invalid n-grams without (left) and with (right) the use of Wikipedia's redirect property.

particular document from the collection. Thus, each judgment in the collection is connected to the source document ID (document title for the first collection and arXiv.org ID for the second). All judgments have been made by one or more experts from the given scientific field.

Relevance Judgments.

Deciding whether or not a given n-gram represents a valid scientific entity can be subject to discussion. Therefore, the guidelines we have given to the assessors stipulate that an n-gram should be considered as a valid entity if it belongs to the domain of the document and satisfies any one of the two following conditions:

- it would make sense to take the n-gram and create a thesaurus/encyclopedia entry about it, or
- the n-gram could be used by an expert to search/filter the papers according to domain-specific (e.g., scientific or technical) criteria.

5.2 Experimental Results

Individual Features.

Table 4 presents the effectiveness of our individual feature families over the Physics test collection, while Table 5 presents similar results for the SIGIR collection. We observe that well-performing features on the Physics collection are based on POS tags or on the connected components obtained from *redirect* information in Wikipedia. We also evaluate our set of basic syntactic features (see Section 4.5) for comparison. On the SIGIR collection, we observe that the best performing features are based on POS tags both in terms of F1 and Accuracy. In terms of Precision, the best approach is the one using the graph connected components.

Feature Comparison.

To find effective feature combinations, we use a decision tree classifier with default parameters [23]. To prevent the classifier from over-fitting the training data, we restrict the minimum number of samples in the leaves to 100 and the maximum depth of the tree to 5. All the results presented below are the mean values resulting from a 10-fold cross-validation of our supervised approach.

We compare the effectiveness between pairs of competing features: compressed and uncompressed POS tags on one hand (see Section 4.1), and building DBPedia connected components with and without the "redirects" property on the other hand (see Section 4.4).

Tables 6 and 7 show the Precision, Recall, F1, and Accuracy values over both collections for different combinations of compressed and uncompressed POS tags features and DBPedia category graph features with and without the *redirect* property. We observe that adding Wikipedia redirects allows to significantly improve Recall in most cases without a significant loss in Precision. Improved Recall is somewhat expected since the *redirect* property allows to match many more Wikipedia concepts. More importantly and as mentioned earlier, this Recall growth does not produce any major loss in Precision, which results in a consistent growth in Accuracy.

Another important result here is that compressed POS tags produce roughly the same Precision values as uncompressed ones with a much smaller number of features. The reason is that the uncompressed POS tag pattern space is much richer than the one of the compressed patterns, which in theory could allow classifiers to yield better performance at the price of possible over-fitting. However, by using a smaller feature space we observe a minor decrease in Precision on both collections with a higher F1 score on the SIGIR

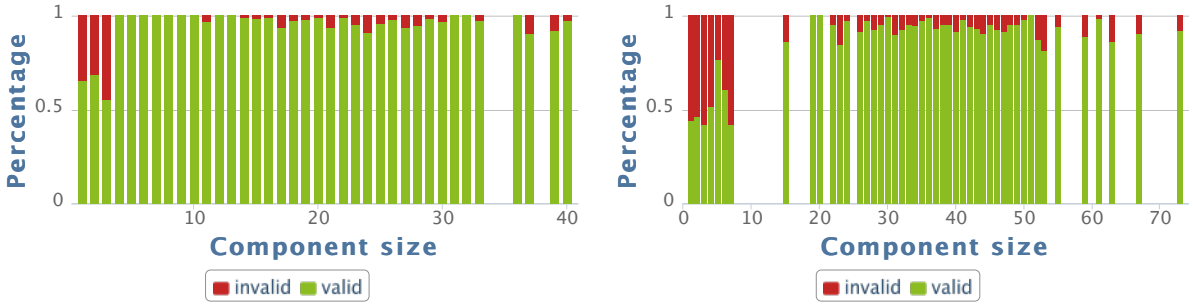

Figure 5: DBPedia connected component size percentage distribution for valid/invalid n-grams without (left) and with (right) using Wikipedia redirect properties.

Table 4: Empirical results for individual feature families on the Physics collection.

	Precision	Recall	F1 score	Accuracy
Compressed POS tags	0.5742	0.9511	**0.7160**	0.6198
Component	0.5039	**1.0**	0.6702	0.5039
Component+Redirects	**0.8116**	0.5572	0.6605	**0.7117**
Punctuation	0.5039	**1.0**	0.6702	0.5039
Syntactic	0.5940	0.1771	0.2728	0.5243

Table 5: Evaluation results for individual feature families on the SIGIR collection.

	Precision	Recall	F1 score	Accuracy
Compressed POS tags	0.8183	0.7307	**0.7715**	**0.7772**
Component	**0.8981**	0.2280	0.3635	0.5888
Component+Redirects	0.8883	0.3869	0.5388	0.6588
Punctuation	0.6414	**0.9450**	0.7642	0.6820
Syntactic	0.6819	0.2124	0.3236	0.5429

collection. Hence, we conclude that compressing POS tags is a better choice since it allows for better model generalization.

Feature Selection.

Table 8 shows the NER features we propose ranked by the score they yield when combined using randomized trees as suggested by [13] on the SIGIR collection. As we can see, the simple techniques based on POS patterns is highly discriminative. However, POS tags are by themselves not sufficient; Other top features include the ones that look at external knowledge bases such as DBLP and the structure connecting the DBPedia entities mentioned in the document.

Table 9 shows the feature ranking based on randomized trees for the Physics collection. In this case, we observe that the most indicative features are the ones based on external ontologies and knowledge bases. In this case, we believe that such features are most distinctive due to the highly technical terms used in Physics and due to the somewhat slower churn of new terminology as compared to the IR field, which is a much younger research area.

In conclusion, we observe that the use of domain-specific knowledge-bases is an effective feature for NER on technical collections.

Table 8: Ranked list of feature importance scores on the SIGIR collection. Selected number of features: 7

Feature name	Importance score
NN STARTS	0.3091
DBLP	0.1442
Component+DBLP	0.1125
Component	0.0798
VB ENDS	0.0386
NN ENDS	0.038
JJ STARTS	0.0364

Feature Ablation Analysis.

Finally, we evaluate the contribution of the individual features to the overall feature combination by a hold-out experiment: We learn a new model by removing each time a feature family to measure the impact of that feature on the overall best possible combination of the features (85% Accuracy on SIGIR and 77% Accuracy on Physics).

Table 10 shows the effectiveness obtained by discarding one feature family for the Physics collection. As we can see, the highest loss in effectiveness (-24% F1 score) is obtained

Table 6: Evaluation results for different feature combinations on the SIGIR collection. The symbols * and $^{*}*$ indicate a statistical significant difference (t-test $p < 0.05$ and $p < 0.01$ respectively) as compared to the best approach (in bold font).

All features	Precision	Recall	F1 score	Accuracy	N features
+ Uncompressed POS + Component	**0.8794**	0.8058^{**}	0.8409^{**}	0.8429^{*}	54
+ Compressed POS + Component	0.8475^{**}	0.8524^{**}	0.8499^{**}	0.8448^{**}	9
+ Uncompressed POS + Component+Redirects	0.8678^{**}	0.8305^{**}	0.8487^{*}	0.8473	50
+ Compressed POS + Component+Redirects	0.8406^{**}	**0.8769**	**0.8584**	**0.8509**	7

Table 7: Evaluation results for different feature combinations on the Physics collection. The symbol * and ** indicate a statistically significant difference (t-test $p < 0.05$ and $p < 0.01$ respectively) as compared to the best approach (in bold).

All features	Precision	Recall	F1 score	Accuracy	N features
+ uncompressed POS + Component	0.8253^{*}	0.6567^{*}	0.7311^{**}	0.7567	53
+ compressed POS + Component	0.7941^{**}	**0.6781**	0.7315^{**}	0.7492^{**}	4
+ uncompressed POS + Component+Redirects	0.8339	0.6674^{*}	**0.7412**	**0.7653**	50
+ compressed POS + Component+Redirects	**0.8375**	0.6479^{**}	0.7305^{**}	0.7592^{*}	6

Table 9: Ranked list of feature importance scores on the Physics collection. Selected number of features: 6

Feature name	Importance score
ScienceWISE	0.2870
Component+ScienceWISE	0.1948
Wikipedia Redirect	0.1104
Component	0.1093
Wikilinks	0.0439
Participation count	0.0370

when removing the background ontology of scientific terms. For the SIGIR collection (see Table 11), we observe that the biggest loss is due to the removal of POS tags (-19% F1 score) confirming the results of feature selection based on randomized trees.

Generally speaking, we see the importance of using domain-specific knowledge bases as well as linguistic properties.

Maximum Entropy Classifier Baseline.

As a method to compare to, we chose the state-of-the-art Maximum Entropy Classifier (MaxEnt) for Named Entity Recognition [2].

In contrast to our approach depicted in Figure 1, this classifier receives the *full text* of the document extracted from the PDF file together with a training set of manually labeled scientific concepts appearing in it. After training the model, the classifier is able to detect unseen scientific concepts given the full text of a new document.

To evaluate the MaxEnt NER approach, we trained it on 80% of SIGIR data and used the rest 20% as a test dataset[13].

During the experiment, 3,380 new n-grams were extracted, out of which 346 new valid entities were discovered.

For a fair comparison, we evaluate our top-performing supervised method on the same data. The results of this experiment are presented in Table 12. As can be observed, the decision tree-based method outperforms the state-of-the-art MaxEnt approach by roughly 15% both in Precision and Recall [14].

[13]The parameters of the tagger were estimated using the *generalized iterative scaling* [5] method.

Table 12: Evaluation results for maximum entropy tagger on SIGIR collection.

	Precision	Recall	F1 score
MaxEnt NER Baseline	0.6566	0.7196	0.6867
Our Approach (using Decision Trees)	0.8121	0.8742	0.8420

5.3 Results Discussion

Based on the experimental results described above, we first observe that the NER approach we propose in this paper for idiosyncratic Web collections substantially outperforms state-of-the-art supervised NER approaches such as MaxEnt. As an example, our best supervised approach yields a F1 score of 84% on the SIGIR collections, compared to 69% for MaxEnt.

We also note that the most effective features among the ones we propose vary depending on the test collection. However, we observe that both the feature family based on the entity-graph structure and the family based on external domain-specific knowledge bases are key to enhance NER effectiveness for idiosyncratic collections.

Finally, while comparing the two test collections, we note that the Physics collection lead to overall lower effectiveness scores. This may be explained by the more formal terminology used in that scientific domain, which makes the identification of valid scientific concepts more challenging as compared to Computer Science academic documents.

[14]Accuracy score is not shown in the table since the notion of *true negative* is not valid for the MaxEnt method, where literally every non-positive n-gram can be considered as negative.

Table 10: Effectiveness values for different feature combinations on the Physics collection. The symbols * and ** indicate a statistically significant difference (t-test $p < 0.05$ and $p < 0.01$ respectively) as compared to the approach using all features.

Feature set	Precision	Recall	F1 score	Accuracy
All Features	0.8375	0.6479	0.7305	0.7592
−ScienceWISE (SW)	0.7861^{**}	0.6072^{**}	0.6850^{**}	0.7187^{**}
−Component+SW	0.8375	0.6479	0.7305	0.7592
−Wikipedia Redirect	0.8368	0.6483	0.7305	0.7590
−Component	0.8354	0.6391	0.7241^{*}	0.7547

Table 11: Effectiveness values for different feature combinations on the SIGIR collection. All differences with respect to the use of all features are statistically significant with $p < 0.01$.

Feature set	Precision	Recall	F1 score	Accuracy
All Features	0.8406	0.8769	0.8584	0.8509
−POS tags	0.9186	0.5370	0.6776	0.7368
−DBLP	0.8330	0.8397	0.8362	0.8305
−Component+DBLP	0.8181	0.8855	0.8505	0.8395
−Component	0.8212	0.8739	0.8467	0.8369

6. CONCLUSIONS

Being able to identify entities in textual documents is known to be beneficial for many tasks, including document search, integration, classification, or summarization. While supervised methods are often used for NER in Web documents such as news articles, novel approaches are needed to perform NER over more specific domains such as for scientific papers.

In this paper, we addressed the task of NER for domain-specific collections by taking advantages of n-gram-based features. We proposed and experimentally validated over two different test collections novel NER features and their combinations using decision trees trained over data created by domain experts. More specifically, our novel features for domain-specific NER include the analysis of entity-graph components as well as the use of external domain-dependent knowledge bases such as DBLP for Computer Science or the ScienceWISE ontology for Physics.

Our results show that the analysis of entity-graph structures and the use of external knowledge bases yield significantly better results in our context. For the two collections we considered, the best performance was obtained by our combined method, yielding up to 85% Accuracy.

As a possible extension of our approach, one could use additional components in the processing pipeline. For example, entity linking approaches allowing to disambiguate entities identified in the text could be exploited. In this work, we directly matched n-grams to Wikipedia entries, though it might be more effective to perform disambiguation first. Further improvements could be obtained by enhancing other components of our system pipeline. For example, advanced PDF extraction approaches could be used to detect bibliographic sections, or to identify titles and emphasized text, which may both allow to improve candidate selection and construct new feature sets. Such approaches providing more structured input would probably yield higher effectiveness values for the task we consider.

7. ACKNOWLEDGMENT

This work is supported by the Swiss National Science Foundation under grant number PP00P2_128459.

8. REFERENCES

[1] K. Aberer, A. Boyarsky, P. Cudré-Mauroux, G. Demartini, and O. Ruchayskiy. Sciencewise: A web-based interactive semantic platform for scientific collaboration. In *10th International Semantic Web Conference (ISWC 2011-Demo), Bonn, Germany*, 2011.

[2] O. Bender, F. J. Och, and H. Ney. Maximum entropy models for named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 148–151, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[3] A. L. Berger and V. O. Mittal. Ocelot: a system for summarizing web pages. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 144–151, New York, NY, USA, 2000. ACM.

[4] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proceedings of the 6th Workshop on Very Large Corpora*, pages 152–160, 1998.

[5] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):pp. 1470–1480, 1972.

[6] L. Del Corro and R. Gemulla. ClausIE: Clause-Based Open Information Extraction. In *Proceedings of the 22nd International World Wide Web Conference (WWW 2013)*, Rio do Janeiro, Brazil, 2013.

International World Wide Web Conferences Steering Committee (IW3C2), ACM.

[7] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 469–478, New York, NY, USA, 2012. ACM.

[8] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *The VLDB Journal*, 22(5):665–687, 2013.

[9] Y. feng Lin, T. han Tsai, W. chi Chou, K. pin Wu, T. yi Sung, and W. lian Hsu. A maximum entropy approach to biomedical named entity recognition. In *Proceedings of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics*, pages 56–61, 2004.

[10] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[11] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and et al. Domain-specific keyphrase extraction. In *Proceedings of the 16th international joint conference on Artificial Intelligence*, pages 668–673. Morgan Kaufmann Publishers, 1999.

[12] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

[13] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, Apr. 2006.

[14] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the 6th ACM international conference on Web Search and Data Mining*, WSDM '13, pages 465–474, New York, NY, USA, 2013. ACM.

[15] X. Jiang, Y. Hu, and H. Li. A ranking approach to keyphrase extraction. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 756–757, New York, NY, USA, 2009. ACM.

[16] S. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin. Automatic keyphrase extraction from scientific articles. *Language Resources and Evaluation*, pages 1–20, 2012.

[17] M. Krapivin, M. Autayeu, M. Marchese, E. Blanzieri, and N. Segata. Improving machine learning approaches for keyphrases extraction from scientific documents with natural language knowledge. In *Proceedings of the joint JCDL/ICADL international digital libraries conference*, pages 102–111, 2010.

[18] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 721–730, New York, NY, USA, 2012. ACM.

[19] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[20] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM.

[21] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[22] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] T. Poibeau and L. Kosseim. Proper name extraction from non-journalistic texts. In *Computational Linguistics in the Netherlands*, pages 144–157, 2001.

[25] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 771–780, New York, NY, USA, 2010. ACM.

[26] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th conference on Computational Natural Language Learning (CONLL)*, pages 147–155, 2009.

[27] A. Ratnaparkhi et al. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142, 1996.

[28] B. Settles. ABNER: An open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

[29] A. Tonon, G. Demartini, and P. Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 125–134, New York, NY, USA, 2012. ACM.

[30] P. D. Turney. Learning algorithms for keyphrase extraction. *Inf. Retr.*, 2(4):303–336, May 2000.

[31] C. Whitelaw, A. Kehlenbeck, N. Petrovic, and L. Ungar. Web-scale named entity recognition. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 123–132, New York, NY, USA, 2008. ACM.

[32] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland. Textrunner: open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, NAACL-Demonstrations '07, pages 25–26, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.