ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

A Context-Aware Relation Extraction Method for Relation Completion

B.Sivaranjani, Meena Selvaraj

Assistant Professor, Dept. of Computer Science, Dr. N.G.P Arts and Science College, Coimbatore, India

M.Phil Scholar, Dept. of Computer Science, Dr.N.G.P Arts and Science College ,Coimbatore, India

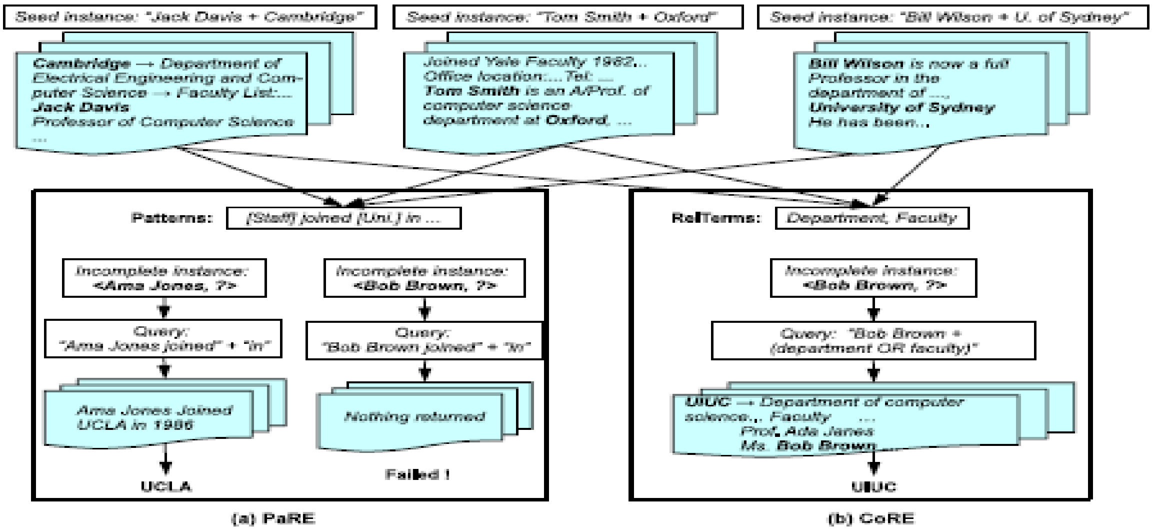
ABSTRACT Identify relation completion (RC) as one recurring problem that is central to the success of novel big data applications such as Entity Reconstruction and Data Enrichment. Given a semantic relation R, RC attempts at linking entity pairs between two entity lists under the relation R. To accomplish the RC goals, propose to formulate search queries for each query entity a based on some auxiliary information, so that to detect its target entity b from the set of retrieved documents. For instance, a pattern-based method (PaRE) uses extracted patterns as the auxiliary information in formulating search queries. However, high-quality patterns may decrease the probability of finding suitable target entities. As an alternative, propose CoRE method that uses context terms learned surrounding the expression of a relation as the auxiliary information in formulating queries. The experimental results based on several real-world web data collections demonstrate that CoRE reaches a much higher accuracy than PaRE for the purpose of RC.

KEYWORDS: Core, Extraction, Semantic relation , formulation

I. INTRODUCTION

The abundance of Big Data is giving rise to a new generation of applications that attempt at linking related data from disparate sources. This data is typically unstructured and naturally lacks any binding information (i.e., foreign keys). Linking this data clearly goes beyond the capabilities of current data integration systems. This motivated novel frameworks that incorporate information extraction (IE) tasks such as named entity recognition (NER) [8], [20] and relation extraction (RE) [23], [31].General RE tasks target those frameworks have been used to enable some of the emerging data linking applications such as entity re construction. Identify relation completion (RC) as one recurring problem that is central to the success of the novel application mentioned above. In particular, an underlying task that is common across those applications can be simply modeled as follows: for each query entity a from a Query List La, find its target entity b from a Target List Lb where ða; bÞ is an instance of some relation.
This is precisely the relation completion task, which is the focus of the work presented in this paper. To further illustrate that task, consider the following scenarios:
Scenario 1: A research institution needs to evaluate the many researchers, however, may not provide the exact venue names within their publications record as per the ranking list. In this case, an RC task is performed between the list of publication titles and the list of venues. This is clearly an example of an entity reconstruction problem, in which each paper entity is reconstructed from different data sources.
Scenario 2: Two online book stores in different languages, such as English and Japanese, want merge their databases to provide bilingual information for each book. Literal translation is not acceptable, especially when some books already have popular and quite different names in different languages. This problem is naturally defined as an RC task between the two book lists in English and Japanese, which is an example of a data integration problem in the absence of foreign key information.

To accomplish the RC task, a straightforward approach can be described as follows: 1) formulate a web search query for each query entity a, 2) process the retrieved documents to detect if it contains one of the entities in the target list Lb, and 3) if more than one candidate target entities is found, a ranking method is used to break the ties (e.g., frequency based[14]). Clearly, however, this approach suffers from the following drawbacks: First, the number of retrieved documents is expected to be prohibitively large and in turn, processing them incurs a large overhead. Second, those documents would include significant amount of noise, which might eventually lead to a wrong b.In contrary to the basic approach above, our goal is to formulate effective and efficient search queries based on RE methods. In general, given some semantic relation R (e.g., (Lecturer, University)), general RE tasks target at obtaining relation instances of the relation R from free text. Clearly, our approach is motivated by the observation that RC can be perceived as a more specializedand constrained version of the more general RE task. Specifically, while RE attempts to find arbitrary entity pairs that satisfy a semantic relation R, RC attempts to match sets of given entities a and b under a semantic relation R.


Seed instance: "Jack Davis + Cambridge"
Seed instance: "Tom Smith + Oxford"
Seed instance: "Bill Wilson + U. of Sydney"
Cambridge → Department of Electrical Engineering and Computer Science → Faculty List:... Jack Davis Professor of Computer Science ...
Joined Yale Faculty 1982... Office location:... Tel: ... Tom Smith is an A/Prof. of computer science department at Oxford, ...
Bill Wilson is now a full Professor in the department of ..., University of Sydney He has been...
Patterns:
[Staff] joined [Uni.] in ...
Incomplete instance: <Ama Jones, ?>
Query: "Ama Jones joined" + "in"
Ama Jones Joined UCLA in 1986
UCLA
Incomplete instance: <Bob Brown, ?>
Query: "Bob Brown joined" + "in"
Nothing returned
Failed !
RelTerms:
Department, Faculty
Incomplete instance: <Bob Brown, ?>
Query: "Bob Brown + (department OR faculty)"
UIUC → Department of computer science.,. Faculty ... Prof. Ada Janes Ms. Bob Brown
UIUC

(a) PaRE (b) CoRE

In that respect, existing general RE methods can potentially solve the more specialized RC problem.

For instance, consider employing the state-of-the-art Pattern-based semi-supervised Relation Extraction method (PaRE) [1], [4] for the purpose of RC. In general, given a small number of seed instance pairs, PaRE is able to extract patterns of the relation R from the web documents that contain those instances. Hence, a web search query can be formulated as a conjunction of a PaRE extracted pattern together with an entity query a and the target entity b is extracted from the returned documents. For example in Fig. 1a, given seed instances of the relation (Lecturer, The PaRE method, however, relies on high-quality patterns which may decrease the probability of finding suitable target entities. That probability is further reduced when an entity query a is used in conjunction with a high-quality pattern. In other words, while an entity query a provides more context for finding a target entity b, the PaRE method falls short in leveraging that context and instead it formulates a very strict search query, which could possibly return very few and irrelevant documents. For example, Fig. 1a shows that no documents have been retrieved for the query ("Bob Brown joined" + "in") and hence, an incomplete instance (Bob Brown) In fact, our experimental evaluation on real data sets shows that no more than 60 percent of query entities can be successfully linked to their target entities under the PaRE method. The remaining 40 percent query entities were mainly entities appeared in very few webpages (i.e., long tail). Though some of those pages contained the correct target entities, PaRE fell short in finding those pages since they failed to satisfy the strict patterns used in formulating the PaRE-based search queries.

Given such limitations of directly adopting PaRE, propose a novel Context-Aware Relation Extraction method(CoRE), which is particularly designed for the RC task. CoRE recognizes and exploits the particular context of an RC task. Towards this, instead of representing a relation in the form of strict high-quality patterns, CoRE uses context terms, which we call relation-context terms (RelTerms).

For example in Fig. 1b, CoRE searches the web for documents that contain each of the seed instance pairs and from those documents it learns some RelTerms such as "department" and "faculty". Based on those RelTerms, CoRE can formulate a query such as "Bob Brown + (department OR faculty)"In comparison to PaRE, given the large number of possible RelTerms, and in turn the large number of possible query formulations, realizing an effective and efficient CoRE involves further challenges:

1) Learning high-quality Rel Terms: as for PaRE, it is quite straightforward to learn patterns which are exactly the same sequences of words surrounding some pairs of linked entities across different webpages. RelTerms, however, can be any terms that are mentioned frequently with some entity pairs,
2) Query formulation: as for PaRE, each pattern can be used to formulate one search query for each query entity.

Terms, however, can be used in different combinations, and each combination corresponds to a potential search query.Meanwhile, not every combination can be used to formulate an effective search query for a given query entity. Given those challenges, we proposed different techniques that are employed by CoRE so as to maximize both efficiency and effectiveness. Our main contributions in this work are summarized as following:

Propose CoRE, a novel Context-Aware Relation Extraction method, which is particularly designed for the RC task. propose an integrated model to learn high-quality relation-context terms for CoRE. This model incorporates and expands methods that are based on terms' frequency, positional proximity and discrimination information. propose a tree-based query formulation method, which selects a small subset of search queries to be issued as well as schedules the order of issuing queries. propose a confidence-aware method that estimates the confidence that a candidate target entity is the correct one. This enables CoRE to reduce the number of issued search queries by terminating the search whenever it extracts a high-confidence target entity.as demonstrated by our experimental evaluation, CoRE provides more flexibility in extracting relation instances while maintaining high accuracy, which are desirable features for fulfilling the RC task. Also demonstrate the effectiveness and efficiency of our proposed techniques inlearning relation terms and formulating search queries.

Roadmap:1)give an overview of CoRE in Section 2) The Rel Terms learning algorithm is introduced in Section 3)while the Query Formulation algorithm is presented in Section
4) The experimental setup is described in Section 5) and the experimental results are in Section 6. We cover related work in Section 7, and then conclude in Section.

II. BACKGROUND AND CORE OVERVIEW

Relation completion is rapidly becoming one of the fundamental tasks underlying many of the emerging applications that capitalize on the opportunities provided by the abundance of big data (e.g., entity reconstruction [9], [13], data enrichment [5], [16], etc.). We formally define the relation completion task as follows. Definition 1 (Relation completion). Given two entity lists La and Lb and a semantic binary relation R, the goal of relation completion is to identify for each entity a 2 La an entity b 2 Lb which satisfies ða; bÞ 2 R. Accordingly, La is a query list, Lb is a target list, a is a query entity and b is a's target entity.
Similar to classical semi-supervised relation extraction [1], [7], the semantic binary relation R is expressed in terms of a few seed linked entity pairs between La and Lb. Differently, however, the goal of relation extraction is to detect semantic relationship mentions in natural language. Formally given a binary relationship R between two types of entities, then an entity pair ða; bÞ is linked under R, i.e. ða; bÞ 2 R, if a and b satisfy the semantic relation R. For instance, givenR ¼(Company, Headquarter), we have (Microsoft, Redmond)2 R. Hence, relation completion is a more specialized and constrained version of the more general RE task. In particular, RC is a targeted task, which is driven by a set of predefined entities (i.e., query list La as shown in Fig. 2). RC attempts to match sets of given entities a and b under a relation R.For instance, consider employing the state-of-the-art Pattern-based semi-supervised Relation Extraction method [1], [4], [7] for the purpose of RC. Hence, a web search query can be formulated as a conjunction of a PaRE extracted pattern together with an entity query a, and the target entity b is extracted from the returned documents. The PaRE method, however, relies on high-quality patterns which may decrease the probability of finding suitable target entities. That probability is further reduced when a is used in conjunction with a high-quality pattern. To overcome the limitations of PaRE, we propose a novel Context-Aware Relation Extraction method, which can recognize and exploit the particular context of an RC task. CoRE represents a semantic relation R in the form of context terms, which we call relation-context terms. In our work, RelTerms provide the basis for formulating web search queries that are especially composed for the purpose of RC. Specifically, CoRE employs what we call a Relation Query (RelQuery), which is basically a web search query that is specially formulated for the purpose of relation completion. Such Rel Query is formally defined as follows:Definition 2 (Relation Query). A Relation Query is a web.

III. LEARNING RELATION EXPANSION TERMS

CoRE utilizes the existing set of linked pairs towards learning the relation expansion terms (i.e., RelTerms) for any given relation R. This task involves two main steps: 1) learning a set of candidate RelTerms for each existing linked pair, and 2) selecting a global set of RelTerms from those individual candidate sets.

3.1 Learning Candidate RelTerms

Several factors such as frequency, position, and discrimination, are typically considered in selecting good expansion terms in the conventional Query Expansion (QE) models [14], [24], [30]. In learning the candidate RelTerms for a given linked pair, we also take those factors into account and they are summarized as follows:

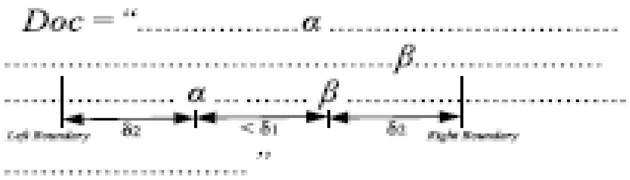
1. Frequency: The RelTerm is mentioned frequently across a number of different RelDocs that are relevant to the given linked pair

2. Position: The RelTerm is mentioned closely to the two entities in the given linked pair, such that it could help bridging the query entity to its target entity.

3. Discrimination: The RelTerm is mentioned much less in irrelevant documents (or non-RelDocs) than in Rel Docs. These factors naturally lead to three formal selection models as described below. Meanwhile, for the remainder of this section, use Qþ to denote a web search query, which takes as an argument a linked pair a+b and returns only the set of relevant documents Fþ containing both a and b. Similarly, Q_ denotes a web search query, which takes as an argument a-b and returns only the set of non-relevant documents F_ containing a but not b.

3.1.1 Frequency-Based Model

The frequency-based model we propose is an adaptation of the classical relevance model [14]. Specifically, the work in [14] assumes different levels of document relevance based on some criteria (e.g., search engine ranking), whereas in our work all retrieved documents are considered equally relevant as long as they contain a þ b. This adaptation enables CoRE to enrich the set of RelTerms with useful terms that might as well appear beyond the top-ranked documents. Accordingly, in our model Fþ is simply the set of all retrieved documents and the probability that a term e is a RelTerm for a given linked pair is estimated as follows: the Bayesian smoothing using Dirichlet priors proposed where tfðe;DÞ is the frequency of term e in document D, jDj is the length of document D. Additionally, PMLðejCÞ is the maximum likelihood estimation of the probability of e in the collection of web documents C indexed by the employed web search engine, which can be approximately estimated with the term frequencies from the Web1T corpus.1 Finally, m is the Dirichlet prior parameter of Dirichlet smoothing. In our experiments, we set m ¼ 1; 500, which is the average length of the documents in collection C.



3.2 Selecting General RelTerms

The Relation after learning all the possible candidate RelTerms from each of the existing individual linked pair, CoRE selects a set of general RelTerms from those candidates. The goal is to select a set of high-quality RelTerms for effective query formulation, and in turn accurate relation completion (i.e., finding target entities). In CoRE, this task takes place in two steps: in the first step, CoRE uses a local pruning strategy to eliminate the least effective RelTerms, and in the second step, CoRE uses a global selection strategy to choose the most effective RelTerms. During the local pruning step, CoRE verifies the effectiveness of each RelTerm in extracting the target entity for the linked pair from which it was learned. In particular, in the verification of a linked pair such as ðai; biÞ, ai is considered as a seed

RelQuery without auxiliary information and each learned RelTerm ei;j is used as a candidate to such seed query with auxiliary information. That is, to formulate a keyword-based query ai þ ei;j. Accordingly, we measure the accuracy achieved for the top-ranked documents that returned and ranked by the employed web search engine, P@N, i.e., the ratio of documents containing the actual target b (top-100 is enough to indicate the performance). To set up a baseline for comparison, we also measure the accuracy of the top-ranked documents which are retrieved with the unexpanded seed query. If the improvement of P@N is evident i.e., the improvement of P@100 is significantly (for example, more than 30 percent), then the verified Rel Terms survive the elimination step and is promoted to the second step (i.e., global selection), which is described next. During the global selection step, CoRE creates a set of a general RelTerms that are best fit for completing the relation under consideration. Intuitively, the RelTerms belonging to more linked pairs with higher probability should have a better coverage rate. Hence, one possibility is to employ a selection model based on the number of covered linked context terms, respectively.

3.2.2 Cluster-Based RelTerms Selection
The cluster-based RelTerm selection model is formalized as follows:

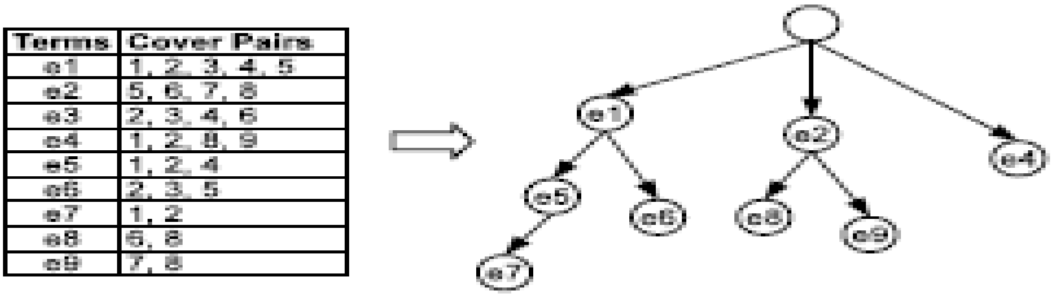
$$P(e|\mathcal{R}) = \sum_{C \in Clusters} P(e|C), \qquad (12)$$

where $P(e|C)$ measures the utility of RelTerm e in determining the target entities within cluster C, which is defined as:

$$P(e|C) = \frac{1}{|C|} \sum_{(\alpha,\beta) \in C} P\left(e|Q_{+}^{(\alpha,\beta)}\right), \qquad (13)$$

IV.RELQUERY FORMULATION

In Section 3, have addressed the challenge of learning high-quality RelTerms for some semantic relation R. In this section, we address the second major challenge towards realizing CoRE. That is, the formulation of efficient and effective RelQueries. In order to put that challenge in perspective, recall that for each query entity a, there are many possible formulations of a RelQuery, each of which is based on a and a conjunction of RelTerms. In particular, assume that n RelTerms are learned, then there are ðn2 _ 2Þ different combinations of RelTerms, leading to ðn2 _ 2Þ different formulation of RelQueries for each a. Obviously, formulating and issuing all those queries will incur a large overhead, which is impractical. Hence, our goal is to minimize the number of issued RelQueries while at the same time maintaining high-accuracy for the RC task. Towards achieving that goal, we propose the following two orthogonal techniques: 1) a confidence-aware termination condition, which estimates the confidence that a candidate target entity bc is the correct target entity (Section 4.1), and 2) a tree-based query formulation method,
which selects a small subset of RelQueries to be issued as well as schedules the order of issuing those RelQueries (Section 4.2). Our termination condition can be used independently or in synergy with our tree-based query formulation method.
When the termination condition is used independently, all the possible RelQueries for a query entity a are ordered arbitrarily and the termination condition is checked after each of those queries is issued. That is, calculate the confidence that one of the candidate target entities bc extracted from the retrieved documents is the right target entity b. If the confidence is higher than a threshold, that is the case, CoRE stops issuing more queries and the search for a target entity is terminated successfully. While the termination condition is expected to eliminate. The need for issuing many of the possible RelQueries, further improvements are attainable by tuning the issuing order of such queries. Ideally, the most effective RelQuery for each a in the query list should be issued first. In reality, however, it is impossible to determine which is the most effective RelQuery for each a. But since the different combinations of RelTerms form a hierarchical structure in which some combinations subsume others, it is often possible to predict the effectiveness of one RelQuery based on the perceived estimated effectiveness of another RelQuery that has already been issued. As such, CoRE builds a tree that captures the relationship between the different combinations of Rel Terms. Further, it employs a tree-based query formulation method which ranks the promising combinations of RelTerms while pruning those combinations that are predicted to be ineffective.



Terms	Cover Pairs
e1	1, 2, 3, 4, 5
e2	5, 6, 7, 8
e3	2, 3, 4, 6
e4	1, 2, 8, 9
e5	1, 2, 4
e6	2, 3, 5
e7	1, 2
e8	6, 8
e9	7, 8

(a) Cover Table

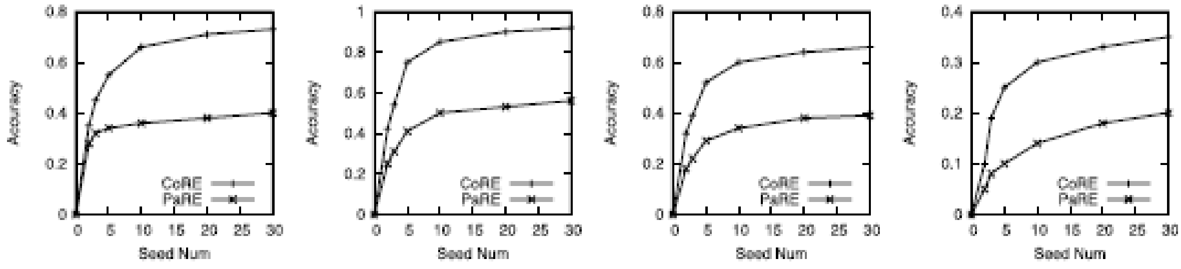
(b) Sorted RelTerm Tree

V. EXPERIMENTAL SETUP

5.1 Data Sets

Perform RC on four real-world data sets below: Academic Staff & University (Staff): About 25k academic staff's full names (from 20 different universities) and their universities have been collected. We also collected 500 university names from the SHJT world university ranking.2Book & Author (Book): This data set contains more than 43k book titles collected from Google Books.3 These books are of more than 20 different categories including education, history, etc. We have also collected about 20k book writers' names (including the chief authors of the 43k books) from Google Books. Invention & Inventor (Invention): This data set contains 512 inventions' names with their chief inventors' full names (311 different people) from an inventor list4 in Wikipedia. Drug & Disease (Drug): This data set contains 200 drug names and the names of 183 different diseases they can cure. It was extracted from a drugs list.5

All four data sets exhibit 1-1 semantic relations. That is,each query entity has only one target entity in target list.



(a) Acd. Staff & Univ.

(b) Book & Author

(c) Invention & Inventor

(d) Drug & Disease

5.2 Metrics

Three metrics are used to estimate the effectiveness or efficiency of our proposed techniques and models. (1) P@N: The precision of top N documents, that is, the percentage of RelDocs in the top N retrieved results. (2) RC Accuracy: To estimate the effectiveness of CoRE and PaRE, we apply them in the relation completion task. The accuracy of RC is the percentage of initially unlinked pairs that could be correctly linked. (3) AvgQueryNum: The average number of processed queries for each RelQuery; The first two metrics are designed for measure effectiveness, the third one measures efficiency.

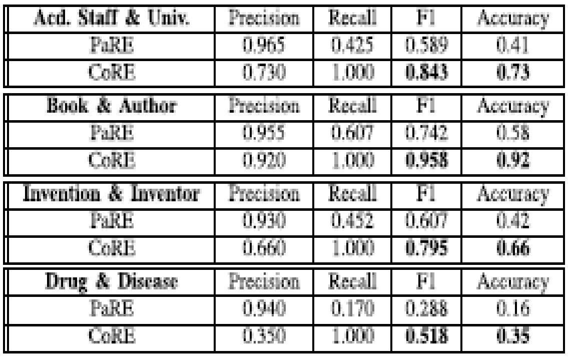
VI .EXPERIMENTAL RESULTS

Present all the experimental results in this section.6.1 CoRE versus PaRE now compare the RC accuracy of applying either CoRE or PaRE in the context of RC with different number of seed instances. As can be observed in Fig. 5, CoRE always reaches a higher RC accuracy than PaRE on all the four data sets with different number of seeds. To further illustrate our accuracy results, Table 1 provides a more comprehensive comparison based on the precision, recall and F1 metrics, in which the seed size is set to 10. Here recall is the percentage of linked pairs, precision is the percentage of pairs that are correctly linked, while F1 ¼ 2 precision recall Precision þrecall. As shown in Table 1, the precision of PaRE is usually very high(90 percent), but its recall is typically low. To the contrast, the recall provided by CoRE is always high(¼ 1:0). This is because in the absence of a confidence threshold, CoRE can always find some target entity for each query entity a even if it is a false positive. But as expected, this comes as the expense of a lower precision when compared to PaRE. Overall, however, the F1 score achieved by CoRE is always greater than PaRE, which emphasizes the advantage provided by CoRE over PaRE.

Through observations to the linked results, we found that all the pairs that can be successfully linked by PaRE were also linked by CoRE, but CoRE could link about 20-30 percent more pairs than PaRE. These 20-30 percent entity pairs are deemed as “long tail” entity pairs that appeared in very few webpages. Though some pages may mention the correct target entities, PaRE fell short in finding these pages due to the strictness of PaRE-based search queries. The experimental results also demonstrate that as the number of the seeds increases from 2 to 10, the performance of both CoRE and PaRE improves dramatically whereas, only slight improvements are observed after the seed number becomes larger than 10. Conclude that a small number of seeds are enough to launch CoRE or PaRE in the context of RC. While the number of seed instances is small (’10), the total number of training sentences is sufficiently large (’1;000), which is the reason for the accuracy of both PaRE and CoRE tends to stabilize after the seed number is larger than 10. Such behaviour is expected since our method is a web-based semi-supervised method instead of a supervised one.

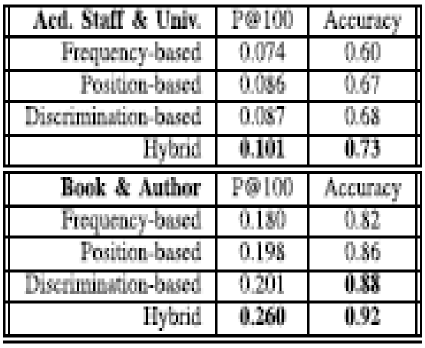
From the experiments, also observe that both PaRE and CoRE reach a relatively higher RC accuracy on the Book data set than that on the other three data sets, probably because book and author are more commonly to be mentioned in some formal formats. As a result, patterns or RelTerms are better shared amongst different instances.

TABLE 1
Comparing CoRE and PaRE Comprehensively



Acd. Staff & Univ.	Precision	Recall	F1	Accuracy
PaRE	0.965	0.425	0.589	0.41
CoRE	0.730	1.000	**0.843**	**0.73**
Book & Author	Precision	Recall	F1	Accuracy
PaRE	0.955	0.607	0.742	0.58
CoRE	0.920	1.000	**0.958**	**0.92**
Invention & Inventor	Precision	Recall	F1	Accuracy
PaRE	0.930	0.452	0.607	0.42
CoRE	0.660	1.000	**0.795**	**0.66**
Drug & Disease	Precision	Recall	F1	Accuracy
PaRE	0.940	0.170	0.288	0.16
CoRE	0.350	1.000	**0.518**	**0.35**

TABLE 2
Comparing Models for Learning Candidate RelTerms



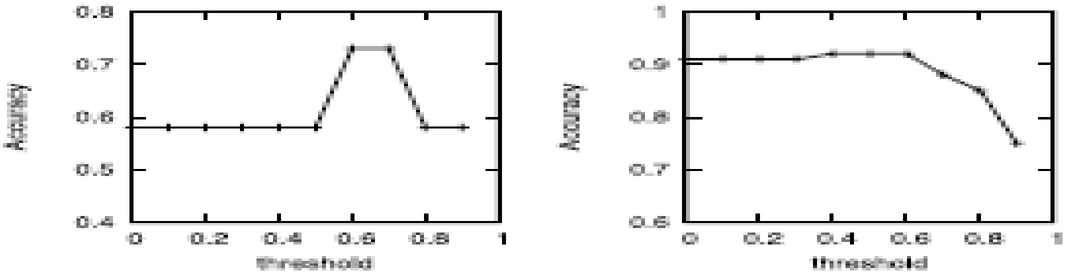
Acd. Staff & Univ.	P@100	Accuracy
Frequency-based	0.074	0.60
Position-based	0.086	0.67
Discrimination-based	0.087	0.68
Hybrid	**0.101**	**0.73**
Book & Author	P@100	Accuracy
Frequency-based	0.180	0.82
Position-based	0.198	0.86
Discrimination-based	0.201	**0.88**
Hybrid	**0.260**	**0.92**

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015



(a) Acd. Staff & Univ.

(b) Book & Author

VII.RELATED WORK

In this work, identify relation completion as one recurring problem that is central to the success of some emerging applications. The RC problem, although novel, is still related to some well-studied problems in the areas of data management and information extraction. For instance, the conventional record linkage (RL) problem whose goal is to find similar entities across two data sets (e.g., [4], [7]) can be considered a special case of the RC problem, in which the semantic relation between those two data sets is always “same as”. In RC, however, that semantic relation can take any arbitrary form such as “published in”, “study at”, “employed by” or “married to”, etc. RC is also very strongly related to the problems arise in question answering systems [15], [28]. In those systems, answers are provided to questions such as “Which country is the city ‘Amsterdam’ located in?”, or “Who is the author of the book ‘The world is flat’?”. Currently, question answering systems rely on relation extraction methods to build an offline knowledge base for providing answers to specific questions. RE methods particularly fit the purpose of question answering systems since its goal is to find arbitrary entity pairs that satisfy a semantic relation R. Meanwhile, RC can be perceived as a more specialized and constrained version of the RE task with the objective of matchings two sets of given entities under a relation R.

While general-purpose RE methods, such as PaRE can be adopted to fulfill the RC task, our experimental evaluation shows that our special-purpose CoRE method provides significant improvements in the RC accuracy. This is primarily because PaRE falls short in incorporating the particular context of the RC task in which query and target entities are given. Like PaRE, other general-purpose RE methods also suffer from the same shortcoming. In the following, we describe those methods and discuss their usage in the context of the RC task. Generally, most RE methods can be divided into three categories: supervised, unsupervised and semi-supervised. Supervised RE methods [12], [31] formulate RE as a classification task, and decide whether an extracted entity pair belongs to a given semantic relation type by exploiting its linguistic, syntactic and semantic features. Supervised method mostly built the model based on tree and sequence kernels that can also exploit structural information and interdependencies among labels [22]. However, they are expensive to be applied to new relation types for requirement of labeled data [2]. To solve this problem, recent work [21], [22] used large semantic databases such as WordNet or Freebase, to provide “distance supervision”.

In particular, they can automatically generate a proper training set with sentences containing pairs of entities in the semantic databases. However, the supervised RE methods are still not appropriate to be used in the context of RC, since we need to generate the feature vectors for each entity pair between the query list and the target list. In order to do that, we need to collect a number of webpages for each entity pair by searching for the web documents containing each term in the query list first and then identify sentences containing each entity pairs. However, this way still requires us to do pair-wised search on all retrieved documents, which will lead to a large overhead. Unsupervised RE methods [3], [25] produce relationstrings for a given relation through clustering the words between linked entities of the relation in large amounts oftext. In some sense, the relation-strings are very similar to the RelTerms learned in the CoRE method, but they only limited in learning these strings, instead of using them to formulate effective RelQueries.

VIII. CONCLUSION AND FUTURE WORK

In this work, identify relation completion as one recurring problem that is central to the success of novel big data applications. We then propose a Context-Aware Relation Extraction method, which is particularly designed for the RC task. The experimental results based on several real-world web data collections demonstrate that CoRE could reach more than 50 percent higher accuracy than a Pattern-based method in the context of RC. As future work, we will further study the RC problem under the many-to-many mapping, and investigate techniques for maintaining the high precision and recall achieved under the many-to-one case.

ACKNOWLEDGMENT

This research was partially supported by National 863 High-tech Program (Grant No. 2012AA011001) and the Australian Research Council (Grant No. DP120102829). Part of this work has appeared as a short paper in CIKM ’11 [18].

REFERENCES

[1] E. Agichtein and L. Gravano, “Snowball: Extracting Relations from Large Plain-Text Collections,” Proc. Fifth ACM Conf. Digital Libraries (ACMDL), pp. 85-94, 2000.
[2] N. Bach and S. Badaskar, “A Survey on Relation Extraction,” Language Technologies Inst., Carnegie Mellon Univ., 2007.
[3] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead, and O.Etzioni, “Open Information Extraction for the Web,” Proc. Int’l