CoRE: A Context-Aware Relation

Extraction Method for Relation Completion

**Keywords:**

Context-aware relation extraction, relation completion, relation query expansion.

**Project Description:**

In this project this project propose CoRE, a novel Context-Aware Relation Extraction method, which is particularly designed for the RC task. This project 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. This project 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. This project 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.

**Techniques:**

1. Context-Aware Relation Extraction method
2. Relation-context terms.

**Module Implementation:**

**1. Pattern Based Search**

**2. Real term Search**

**3. Context-aware relation extraction**

**4. Relation Completion**

**Pattern Based Search:**

Multiple RelQueries are posed, each of which is based on the query entity a in conjunction with one of the patterns extracted by the PaRE method (e.g., (“Bob Brown joined” þ “in”), (“Bob Brown works at”’), etc.). Using patterns as auxiliary information will generate very strict RelQueries, which will return the least number of web documents, but most of which are RelDocs. Hence, if a query entity a happened to appear in a webpage under one of the used patterns, it will be quickly matched with its correct target entity. However, such assumption is unrealistic for many query entities that appear in very few webpages (i.e., long tail). For those entities, no webpages will be returned and will remain unmatched. This formulation is orthogonal to the pattern-based one above, where Multiple RelQueries are posed, each of which is based on the query entity a and an entity bc from the target list. Hence, each of the retrieved documents is processed to detect any of the patterns extracted by the PaRE method to justify whether ða; bcÞ 2 R. Obviously, this formulation incurs a large overhead as it requires posing a large number of RelQueries for each query entity as well as processing the documents retrieved by those queries.

**Real Term Search:**

CoRE utilizes the existing set of linked pairs towards learning Relation Expansion Terms (i.e., RelTerms) for any relation R. This task involves two main challenges:

1. learning a set of high-quality candidate RelTerms from each existing linked pair
2. Consolidating and pruning those individual candidate sets into a minimal global set of RelTerms that are used in the formulation of RelQueries.

CoRE formulates and issues a set of Relation Queries (i.e., RelQueries) for each query entity a based on the set of learned RelTerms. However, there are many possible formulations, each of which is based on a and a conjunction of RelTerms. Clearly, formulating and issuing all those queries will incur a large overhead, which is impractical. Hence, one major challenge 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 two orthogonal techniques:

1) a confidence-aware termination (CA-Term) condition, which estimates the confidence that a candidate target entity is the correct one 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.

**Context aware relation extraction:**

This formulation is based on our proposed CoRE, in which multiple RelQueries are posed, each of which is based on the query entity a in conjunction with several RelTerms extracted by the CoRE method (e.g., (“Bob Brown” þ “Department”), (“Bob Brown” þ” Faculty”), etc.). By using RelTerms, a limited number of documents are retrieved, among which some are RelDocs that contain the correct target entity. Context-based formulation tries to strike a fine balance between a very strict RelQuery formulation (i.e., pattern based) and a very relaxed one (i.e., query-based). Towards this, CoRE exploits Rel Term towards a flexible query formulation in which a RelQuery is formulated based on the query entity a in conjunction with one or more RelTerms.

**Relation Completion:**

Rel Terms 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 considered as a seed RelQuery without auxiliary information and each learned RelTerm used as a candidate to such seed query with auxiliary information.

**Scope:**

In this work, we 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 inve-

stigate techniques for maintaining the high precision and recall achieved under

the many-to-one case.

**Objectives:**

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) and

relation extraction (RE) Those frameworks have been used to enable some of the emerging data linking applications such as entity reconstruction and data enrichment. In this work, we 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 appli-cations 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 semantic relation R. This is precisely the relation completion task, which is the focus of the work presented in this project.

**Example**

****

****

# System Requirements:

HARDWARE REQUIREMENTS:

Hardware - Pentium

Speed - 1.1 GHz

RAM - 1GB

Hard Disk - 20 GB

Floppy Drive - 1.44 MB

Key Board - Standard Windows Keyboard

Mouse - Two or Three Button Mouse

Monitor - SVGA

SOFTWARE REQUIREMENTS:

Operating System : Windows

Technology : Java and J2EE

Web Technologies : Html, JavaScript, CSS

IDE : My Eclipse

Web Server : Tomcat

Tool kit : Android Phone

Database : My SQL

Java Version : J2SDK1.5

**Introduction**

* **Big Data is giving rise to a new generation of applications that attempt at linking related data from disparate source.**
* **data is typically unstructured and naturally lacks any binding information.**
* **capabilities of current data integration systems .**
* **This motivated novel frameworks that incorporate information extraction (IE) tasks such as named entity recognition (NER) and relation extraction (RE).**
* **Those frameworks have been used to enable some of the emerging data linking applications such as entity reconstruction and data enrichment.**

Project Execution Plan:

What is a Project Execution Plan?

The *Project Execution Plan* is the ‘road map’ used by the Project Team to deliver the agreed project outputs. It outlines the responsibilities of the Project Team and key stakeholders.

Why would you develop a Project Execution Plan?

A *Project Execution Plan* is developed to expand on the Project Business Plan by specifying the day-to-day (operational) management procedures and control plans including:

* detailed project plans;
* resource schedules;
* quality procedures;
* reporting procedures;
* product purchasing and development plans;
* risk management planning; and project budgets.

1. CoRE versus PaRE: We mainly compare CoRE with PaRE in the context of RC. For CoRE, we use Google API6 to retrieve documents and associated snippets from the web.The number of feedback documents we use for each query

is fixed to 100, which is the maximum that the search engine returns at a time.

We learn RelTerms from single relation query with the Hybrid Positional+Semi-Negative model,and then do the tree-based QF method. For PaRE, we adapt

the state-of-the-art method proposed in NELL (Never-EndingLanguage Learner)

for PaRE. Both CoRE and PaRE are semi-supervised methods, which require a small number of seed linked entity pairs.Therefore, we perform experiments with several numbers of seed linked pairs (jTj ¼ 2; 3; 5; 10; 20; 30). For each size,cross-validation is achieved by generating five different random sets for training, while all the remaining pairs in each data set are used for testing.

2. Models for learning Candidate RelTerms: Four models were proposed for

learning RelTerms from single linked pairs (1) The Frequency-based model;

(2) The Position-based model; (3) The Discriminationbased model;

(4) The Hybrid of all the above models. To demonstrate the effectiveness of the four models, we evaluate them in two dimensions: (1) RC accuracy; and (2) P@100.

3. Models for selecting General RelTerms: Three possible models are compared in selecting general RelTerms: (1) The Query-Based Model; (2) The Cluster-Based Model; (3) The Unexpanded Model. We evaluate in two dimensions:

(1) RC

Accuracy; and (2) P@100.

4. Confidence-Aware Termination: Four different options are available to estimate the confidence of each retrieved document, including: (1) The

Uniform option; (2) The Page Rank option; (3) The Number of RelTerms option;

(4) The Conf. of RelTerms Combination option. We will work out a proper threshold for the Confidence-Aware Termination for each of the options,

and then compare the four options in one dimension: RC Accuracy. Finally, to demonstrate the effectiveness of the Confidence-Aware Termination strategy,

we compare the RC Accuracy and AvgQueryNum of our Tree-based QF method

with or without using the Confidence-Aware Termination strategy.

5. RelQuery formulation methods: We will compare proposed

efficient QF method with two baselines. Thus we

have three methods as following: (1) The Linear Coveragebased

QF uses top-K RelTerm (combinations) for expansion

one after one; (2) The Linear MinCoverSet-based QF uses

RelTerms in the MinCoverSet to do the expansion one after

one; (3) The Tree-based QF selects RelTerms and their

combinations with the guidance of the Sorted Graph. We

will evaluate them in two dimensions: (1) RC Accuracy; and

(2) AvgQueryNum.

**Application and Usefulness:**

1. A Context-Aware Relation Extraction method, which is particularly designed for the RC task.
2. CoRE could reach more than 50 percent higher accuracy than a Pattern-based method in the context of RC.
3. 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.

**Demonstration of relation completion problem**



**References**

****

****