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. 2, Special Issue 3, July 2014

Entity Disambiguation for Comparable Entity Mining Using MLN

S.Deepa [1], S.Vasumathi Kannagi [2], K.P Porkodi 3.

[1]ME student, Department of Computer Science and engineering, Info Institute of Engineering, Kovilpalayam, Coimbatore, Tamilnadu, India,

[2]Assistant Professor, Department of Computer Science and engineering, Info Institute of Engineering, Kovilpalayam, Coimbatore, Tamilnadu, India,

[3]Assistant Professor, Department of Computer Science and engineering, Al-Ameen Engineering College, Nanjai Uthukuli, Erode, Tamilnadu, India,

ABSTRACT*:* Comparing alternative options is one essential step in decision-makings that we carry out every step of the existing method. itdoes not mine rare patterns, the entity disambiguity problem is occurs. to over come these use the Markov Logic Network (MLN) a joint model which combines first order logic (FOL) and Markov networks.The model achieves the contextual information of the recognized entities for entity disambiguation as well as the constraints when linking an entity. Proposed MLN which is the combination of first order logic (FOL) and Markov networks with combination of NIL-filtering and entity disambiguation stages. For entity disambiguation problem the representation capture the entity information from background knowledge with familiar entities as well as the constraints while connecting an entity.

KEY WORDS: Markov Logic Network (MLN), Comparative Questions, Information Extraction, Bootstrapping, Sequential Pattern Mining and Comparable Entity Mining.

I. INTRODUCTION

Many of the existing question answering systems methods uses exterior in formation and tools for respond analytical. For reference it uses entity taggers, Word Net, specific parsers and ontology list. Though, at the latest TREC-10 QA assessment, the attractive scheme second-hand immediately single resource. The obvious authority of such patterns stunned numerous. To handle this accordingly determined to examine their possible by acquiring patterns routinely and to determine their accurateness**.**

In the World Wide Web the comparison of search results with similar concept or similar information via search the relevant pages regarding the targeted products, discovercontending products, understand writingreview, and recognize pros. In this paper focal point of discovering set of comparable pair of entities .Generally it becomes complex to choose if together entities are equivalent for a variety of reasons. To overcome this problem entity linking helps to study the possible information from background knowledge many disambiguationmove toward have been planned to deal with the entity ambiguity difficulty. For instance, Dredze et al [1] proposed the disambiguation mission as a ranking difficulty and developed features to link Wikipedia entries. Zhang et. al. [3] second-handbe automatically generate the quantity to instruct a dual classifier to reduceambiguity. Dai et al. [2] composedexteriorinformation for every entity and intended likelihoods stating the correspondence of the presenttextbookby means of the information to get better the disambiguation presentation.

In adding together to the entity ambiguity difficulty, the EL task in Text Analysis Conference (TAC) 2009 establish the nonappearanceconcern McNamee et.al [4] for entities that include no equivalent entry in the KB a NIL be supposed towardexist returned.

In this paper current an approach move toward for automatically learning such mining comparators beginning comparative questions and additionally, make available and grade comparable entities intended for a user's input entity suitablybeginning the web. It is very useful method for help to users to choose alternative choices by suggestive ofsimilar entities based on additional users' previousdesires. To mine comparators pairs result first need to detect whether the question is present in comparator or not Richardson and Domingos et .al [7] developed markov logic network based joint model which combine first order logic (FOL) and Markov networks. The model captures the contextual information of the recognized entities for entity disambiguation as well as the constraints when linking an entity mention to a KB entry. Our method uses the machine learning based weakly supervised method for bootstrapping to formulate a huge tagged corpus preliminarythroughsimply a small number of examples of QA pairs. Comparablemethods have been investigated expansively in the field of information extraction. These methods are significantly aided by the information that there is no necessitatein the direction of corpus, whereas the profusion of data on the web makes it easier to concludedependable statistical estimates.

II. RELATED WORK

In conditions of discovering associatedsubstance for an entity, our employment is comparable to the investigate on recommender systems, which suggestsubstance to a consumer.

Recommender systems mostly rely on similarities among items and their arithmeticalcorrelation in consumer log data [8]. While consideration of Amazon, the principle of commendation is to attract their customers to appendadditionalsubstanceitems. Still thesetypes of questions posted by web users are complex to be predictingbasically based on item similarity among them. They are comparable but also dissimilarsorequestassessment with every other. It is obvious that comparator mining and item recommendation are related other than not the similar.

Our effort on comparator mining is associated to the investigate on entity and relative extraction in information extraction [9]. Jindal and Liu [10], [11] also proposed a comparator mining methods for mining relative sentences and relationships. Both class and sequential rules learned to annotate the result of news and review domain to mine relative sentences as well as relationship. The similar methods followed by author [10] also applied to comparative question identification. Though, their methods characteristically can accomplishelevated precision but endure from low recall [11].

Solving Entity Linking problem for MineralIndustry Research Laboratoryproposed an MLN. With a joint conclusionprocedure can carry out together tasks concurrently to let alone this kind of inaccuracyproliferation by Poon and Domingo's et.al [12]. Joint inferences havedeveloped intowell-likedlately, since they make it probable for features and constraints to be communalamongst tasks. For instance, word sense disambiguation (WSD) solved by using representation of joint model by Che and Liu [13] and integrated parsing as well as entity recognition in a joint representation by Finkel and Manning et.al [14].

III. WEAKLY SUPERVISED AND MARKOV-LOGIC NETWORK COMPARABLE ENTITY MINING

Markov logic network (MLN) to representation of interweaved constraints. It is one of the major types of entity linking method with genetic material state relating. Proposed MLN which is the combination of first order logic (FOL) and Markov networks with combination of NIL-filtering and entity disambiguation stages. The representation captures the background information of the familiar entities for entity disambiguation as well as consideration of entity linking in the Knowledge Base (KB) .For instance, an individualdeclarepreservesimply be linked to a KB entry when the state has not been familiar as an NIL.The KB bases the formula are demonstrated with four keywords: constants, variables, functions, and predicates. Whereas constants are referred to as objects in the database entries, that related variables are denoted as x and y for selected objects. Relationship among the data objects are represented as predicates. A world is an obligation of reality values to everyoneprobableview atoms is also referred to as predicates. Knowledge Base (KB) is anincompleterequirement of a world; everyparticle in it is accurate, false or unidentified.

A Markov Logic Network (MLN) characterizes the joint distribution of a set of variables $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots\ldots \mathrm{X}_n) \in \mathrm{x}$ as a result of factors:

$$\mathrm{P}(\mathrm{X} = \mathrm{x}) = \frac{1}{\mathrm{Z}} \prod\nolimits_{\mathrm{Z}} \mathrm{f}_{\mathrm{k}}(\mathrm{x}_{\mathrm{k}})$$

As extended as intended for every one $\mathrm{P}(\mathrm{X} = \mathrm{x}) > 0$, for everyone x the distribution can be consistently represent as a log-linear representation:

In our disambiguation move toward, rely on background knowledge k, such as an entity's populated locationid. Describes a variety of aspect of the entity's ambiguous background knowledge $entry, id$. Every time the entity is discussed, a number of this aspect determination be state as well. Using $k$can write formula similar to the subsequent for disambiguation.
hasCandidate(i, id)
hasquestionInfo(i, id, sd)
hasWord(w): the abstract contain a word w.
QIKeyword(w), isQIPartner(id1, id2)
hasQIPartnerRank (i, id, r), hasGOTermRank(i, id, r),
hasTissueTermRank(i, id , r)
hasPrecedingWord(i, w, l), hasFollowingWord(i, w, l)
hasUnigramBetween(i, j, w)

Variable Type

i: an integer, which refers to the ith question mention in the given article (similarly j refers to the jth mention
id: an EntrezQuestion ID, which refers to a linked KB entry.
sd: an integer, which refers to the sentence distance.
w: a word.
r: an integer, which refers to the rank of the
matching.
l: an integer, which refers to a context window
length

$P(\mathrm{X} = \mathrm{x}) = \frac{1}{\mathrm{Z}} \exp(\sum_{\mathrm{i}} \mathrm{w}_{\mathrm{i}}\, \mathrm{g}_{\mathrm{i}}(\mathrm{x}))$, Where $\mathrm{g}_{\mathrm{i}}(\mathrm{x})$ is the features are subjective functions of the variables situation. An MLN L is a set of pairs $(\mathrm{F}_{\mathrm{i}}, \mathrm{w}_{\mathrm{i}})$, where $\mathrm{F}_{\mathrm{i}}$is a principle in FOL and w_{i} is a real numeral represent a weight. Mutually with a predetermined position of constants, it describe a Markov network,$\mathrm{M}_{\mathrm{L,C}}$where contains single node for every probable preparation of every predicate appear in L. The assessment of the node is 1 if the ground predicate is true, and 0 or else. The probability distribution in excess of probable worlds is known by $\mathrm{P}(\mathrm{X} = \mathrm{x}) = \frac{1}{\mathrm{Z}} \exp(\sum_{\mathrm{i}} \sum_{\mathrm{j}} \mathrm{w}_{\mathrm{i}}\, \mathrm{g}_{\mathrm{j}}(\mathrm{x}))$ where Z is the separation function, F is the set of every one first order formula in the MLN, is the set of groundings of the i^{th} first-order formula, and $\mathrm{g}_{\mathrm{j}}(\mathrm{x}) = 1$ if the j^{th}ground formula is true and $\mathrm{g}_{\mathrm{j}}(\mathrm{x}) = 0$ or else.

Describe four predicates to confine the acceptedquestions environment information, together withquestion location, Question Interaction (QI), Tissue Type and Question ontology.

The “+ ” details in the beyondmethod indicates that necessitystudy a split weight for every grounded variable (sd).

Correlation information from knowledge base (KB) approach interacts with entity one to entity two to solve adisambiguating an entity problem.The QI information stored in the backend database with correlation measure. Based on this result and candidate KB entry distribution result , the id toassociated with the majority unambiguous entries is the mainlyprobableid to be linked to i.Additionaldescribe the subsequent formula to confine the dependence that an entity be supposed to be linked to id_2 if one more entity havebe linked toid_1 structure a correlation with id_2 .Filtering the subsequent mention typepersonsbelong to classes with the intention of are not in the database curation objective; called NILs. In linking question with gene are stored to KB Database and NIL filter apply the QI interaction to solve the entity disambiguation problem.The subsequent formula to make sureto, every time the entity is linked to a KB entry id , it be supposed to be an entity appropriate for linking,$\mathrm{islinkedTo(i, id)} \Rightarrow \mathrm{issuitableForlinking(i)}$

$\exists \mathrm{w.\ hasWord(w)} \Lambda \mathrm{QIKeyword(w)}$
$\Lambda \mathrm{islinkedTo(i, id_1)}$
$\Lambda \mathrm{hascandidate(j, id_2)}$
$\Lambda \mathrm{isQIPair(id_1, id_2)} \Longrightarrow \mathrm{islinkedTo(j, id_2)}$**formula(1)**

The steps involved in this Markov Logic Network are defined below:

Input : A Markov network represents the joint distribution of a set of variables $X = (X_1, X_2, \ldots\ldots X_n) \in x$, L is set of pairs (F_i, w_i)
Output: Find disambiguation result (F_i, w_i)
Step 1: Define or found the set of disambiguation pairs from using Markov Logic Network (MLN).
Step 2: Find the set of disambiguation result $(F_i, w_i)$where F_ia formula in FOL is andw_i is a real number represented a weight.
$\exists w.\, hasWord(w) \Lambda QIKeyword(w)$
$\Lambda islinkedTo(i, id_1)$
$\Lambda hascandidate(j, id_2)$
$\Lambda isQICPartner(id_1, id_2) \Longrightarrow islinkedTo(j, id_2)$**formula(1)**
Step 3: If it is $if(F_i, w_i) > C$then defines a Markov network ,$M_{L,C}$where contains one node for each possible grounding of each predicate appearing in L **Step 4:** The value of the node is 1 if the ground predicate is true, else 0 otherwise
Step 5: Find the probability distribution over possible worlds is given by ,

$$P(X = x) = \frac{1}{Z} exp(\sum_i \sum_j w_i\, g_j(x))$$

Step 6: In the step $g_j(x) = 1$ if the jth ground is true and $g_j(x) = 0$ otherwise.
Step 7: Return the best probability result for each pairs (F_i, w_i)
Step 8: Then now apply bootstrapping procedure
collection of sequence patterns is specified as S an indicative extraction pattern (IEP) ,condition it be able to be used to identify comparative questions and extract comparators in them through elevated consistency. Primary will properly describe the consistency attain of a sample. The sequence patterns is specified as S as a sequence S where s_i can be a wordor a representationof symbol denotemoreover a comparator ($\$c$), or the beginning ($\#start$) or the end of a question($\#end$).

```
Input: CP, G
$InitializesolutionQ \leftarrow \{\}, P \leftarrow \{\}\ P_{new} \leftarrow \{\} CP_{new} \leftarrow \{\}$
$Repeat$
    $P \leftarrow P + P_{new}$
    $Q_{new} \leftarrow compartiveQuestionidentify(CP_{new})$
    $Q \leftarrow Q + Q_{new}$
    $For q_i \in G do$
        $If ismatchexistingpatterns(p, q_i)\ then$
            $Q \leftarrow Q - q_i$
        $Endif$
    $Endfor$
    $p_{new} \leftarrow mineGoodpatterns(Q)$
    $cp_{new} \leftarrow \{\quad\}$
    $For q_i \in G do$
        $cp \leftarrow extractcomparablerpatterns(p, q_i)$
        $If cp \neq NULL and cp \notin CP then$
            $CP_{new} \leftarrow CP_{new} + \{CP\}$
        $Endif$
    $Endfor$
$Until P_{new} = \{\}$
$Return P$
```

Patterns Generation and evaluation

To produce sequential patterns, become accustomed the exterior text pattern mining techniqueintroduced. For somespecified comparative question and its pairs, questions of each comparator are replaced with representation$Cs. Together symbols, #start and #end, are emotionally involved to the start and the end of every sentence in the question. To decreasevariety of seriesinformation and extractpossible patterns, expression chunking is practical. After that, the next three kinds of sequential patterns are generated beginningseries of questions:

Lexical patterns point toward sequential patterns containing only the representation of symbols and of only words. Generalized patterns are able to be as wellprecisesimplify lexical patterns by replacing one or additional words/phrases by means of their POS tags. 2n - 1 generalized patterns can be fashionedbeginning a lexical pattern containing N words exclusive of$Cs.

Specialized patterns a pattern be able touniversaleven though a question is relative, According to our primarysupposition, a reliability score $R^k(p_i)$ for a contestant pattern p_i at iteration k might bedefinite as follows

$$R^k(p_i) = \frac{\sum_{\forall cp_j \in cp^{k-1}} N_Q(p_1 \rightarrow cp_j)}{N_Q(p_1 \rightarrow cp_j)}$$

Comparator Extraction

Comparator extraction used a random based strategy to perform comparator, it randomlychoose a pattern amongst patterns which be able to be useful to the question. Another type of strategy is Maximum length strategy. These strategies select a maximum pattern for given a questionwhich is able to be applied to the question comparator extraction. From the discussion above comparator extraction in this work uses a maximum length method is able to exist exactly enclosed which means that the model is additionalappropriateintended for the query.

Comparable ranking methods

The major importance of comparable based ranking methods is to compare the extra attractiveentity for an entity if it is compared with the entity furtherregularly. Based on this insight, describe a straightforward ranking function $R_{freq}(c, e)$which ranks the comparator results corresponding to the amount of time when the comparator$c$ is comparetoward the user's key$e$in relativequestions collection$Q$:

$$R_{freq}(c, e) = N(Q_{c,e})$$

$$R_{rel}(c, e) = \sum_{q \in Q_{c,e}} R(p_{q,c,e})$$

Graph-Based Ranking

Althoughregularity is well-organized for comparator ranking, the frequency-based technique can experiencewhilst an effortoccurinfrequently in question collection; for instance, understand the case that all probable comparators to the effort are compared simplyon one occasion in questions. In this case, the Frequency-based method mightbe unsuccessful to create a significant ranking end result. Then, Representability is supposed tomoreover be considered. For instance, when individualrequirements to buy a smart phone and allowing for"iphone-89","iphone 87" is the primarylone he/she needs to evaluate. It uses a graph-based PageRanking method to compare questions. If a comparator is compared to numerousadditionalsignificant comparators which are able to be moreover compared to the input entity, it would be considered as a precious comparator in ranking. Based on this scheme, examine PageRank algorithm to rank comparators for a known input entity which mergeregularity and representability.

IV. EXPERIMENTAL RESULTS

All experimentationwas conducted on concerning questionsthat are mined beginning Yahoo! Answers' question name field. The motivationto facilitate used simply a name field is that they obviouslyconvey a majorpurpose of an asker by means of a structure of straightforward questions in all-purpose. Physically constructed keyword set which contains upto 53 words such as “otherwise” and “rather,” which are superior indicators of comparative questions. Categorizes of each and every questions set into SET-A and SET-B one or more keywords from each set ,it randomly selected other than earlier selected questions beginningevery Yahoo! Answers category with atleast one keyword present as mentioned above. It contains 765 comparative questions and 1,456 noncomparative questions. For comparative question identification experiments were conducted for each set category separately.Whereas comparator extraction is applied only for SET-B. All the left behind unlabeled questions that is SET-R used for weakly supervised method.

Table 1 shows experimental result in the category of Identification, extraction and all results. Identification says that the comparative questions are identified correctly, Extraction only says that the in which the comparator extracts the question correctly extracted are used as input, and Allindicate the back-to-back performances whilst question detectionoutcome were second-hand in comparator extraction. Reminder that the outcome of WSN-MLN technique on our collections are extremely comparable to what is reported in their manuscript and the figure 1,2,3 values are tabulated in 1.

Table 1: Performance Comparison between Weakly supervised model (WSM) and Weakly supervised model with Markov logic network(WSM-MLN)

Results	Identification only		Extraction only		All	
	Weakly supervised model(WSM) and Weakly supervised model with Markov logic network(WSM-MLN)					
Recall	0.817	0.915	0.760	0.854	0.760	0.870
Precision	0.833	0.925	0.716	0.925	0.776	0.916
F-score	0.825	0.935	0.833	0.889	0.768	0.936

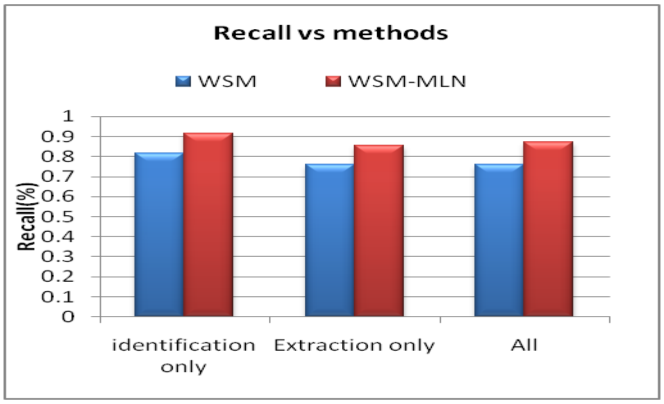

Figure 1:Recall vs. types

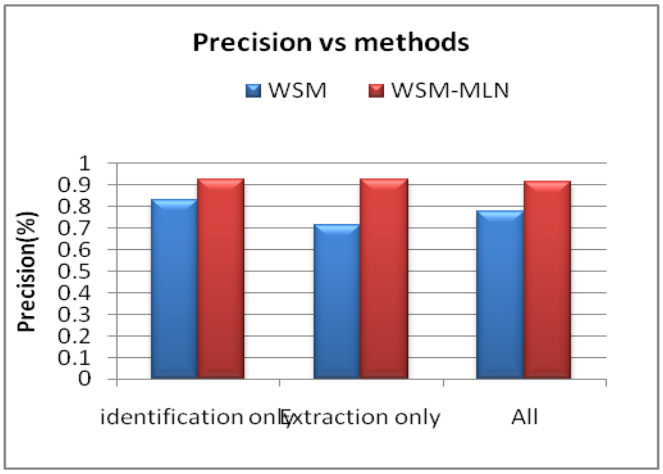

Figure 2: Precision vs. types

TABLE 2: Effect of Pattern Specialization and Generalization in the End-to-End Experiments

Methods	Recall		Precision		F-Score	
	Weakly supervised model(WSM) Weakly supervised model with Markov logic network(WSM-MLN)					
Original patterns	0.689	0.815	0.449	0.760	0.544	0.750
Specialized	0.731	0.850	0.602	0.810	0.665	0.851
Generalized	0.760	0.860	0.776	0.854	0.768	0.825

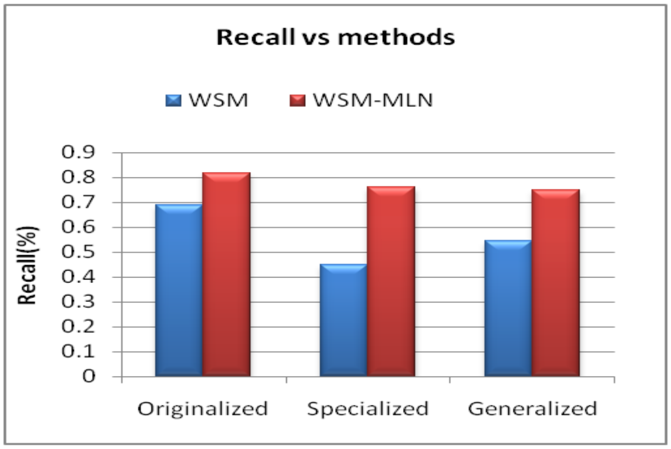

Figure 3: Effect of Pattern vs. recall

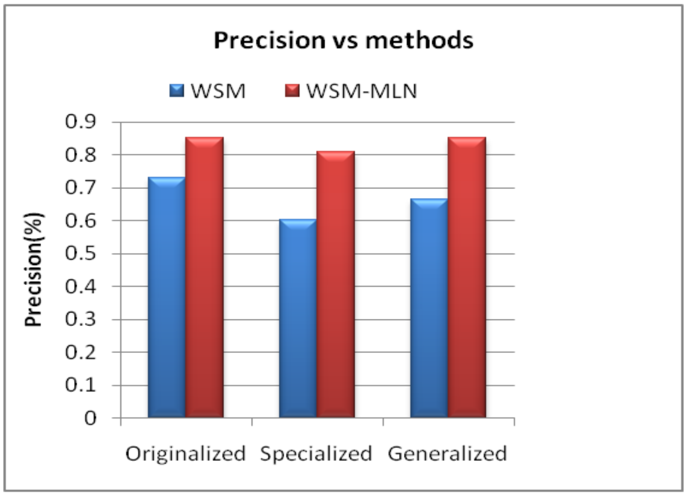

Figure 4: Effect of Pattern vs. precision

V. CONCLUSION AND FUTURE WORK

In this paper current an original entity disambiguation by means of weakly supervised process to recognize comparative questions and extract comparator pairs concurrently. It depends on insight of key patterns that are generated by high-quality comparative question detection pattern be supposed to extort good comparators, and a good quality comparator pair be supposed to suggest itself in good comparative questions to bootstrap the extraction process. By leveraging hugequantity of unlabeled data and the bootstrapping procedure with in significan tmanagement .The investigation alout come demonstrate that our method is effectual in together comparative question detection and comparator extraction. It considerably improve recall in together tasks whils tmaintainelevated precision. Our examples demonstrate that these comparator pairs replicate interested in comparing which is actually wanted by user. Our comparator mining outcome can be second-hand for a commerce exploration or product recommendation organization. For instance, automatic proposition of comparable entities can help out users in their assessment activities earlier than building their acquiredecision. In addition, our outcome can make available helpful information to companies which would like to recognize their competitors.In future work also map to extend technique to summarize answers pooled by a specified comparator pair.

REFERENCES

1. Dredze, Mark, Paul McNamee, Delip Rao, Adam Gerber and Tim Finin. 2010," Entity Disambiguation for Knowledge Base Population", In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing.
2. Dai, Hong-Jie, Po-Ting Lai and Richard Tzong-Han Tsai," Multistage Gene Normalization and SVM-Based Ranking for Protein Interactor Extraction in Full-Text Articles", IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 7(3): 412- 420,2010.
3. Zhang, Wei, Jian Su, Chew Lim Tan and Wen Ting Wang," Entity Linking Leveraging Automatically Generated Annotation", In: Proceedings of the 23rd International Conference on Computational Linguistics Beijing,2010.
4. McNamee, Paul and Hoa Trang Dang," Overview of the TAC 2009 Knowledge Base Population Track", In: Proceedings of the Second Text Analysis Conference (TAC 2009), Gaithersburg, Maryland,2009.
5. Bunescu, R and M Pasca," Using encyclopedic knowledge for named entity disambiguation" ,In: European Chapter of the Association for Computational Linguistics,2006.

6. Li, Fangtao, Zhicheng Zheng, Fan Bu, Yang Tang, Xiaoyan Zhu and Minlie Huang," THU QUANTA at TAC 2009 KBP and RTE Track", In: Proceedings of Test Analysis Conference 2009 (TAC 09), Gaithersburg, Maryland USA,2009.
7. Richardson, Matthew and Pedro Domingos," Markov logic networks. Machine Learning, 62(Special Issue: Multi-Relational Data Mining and Statistical Relational Learning): 107-136,2006.
8. S. Li, C.-Y. Lin, Y.-I. Song, and Z. Li, "Comparable Entity Mining from Comparative Questions," Proc. 48th Ann. Meeting of the Assoc. for Computational Linguistics (ACL '10), 2010.
9. E. Riloff and R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping," Proc. 16th Nat'l Conf. Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conf. (AAAI '99/IAAI '99), pp. 474-479, 1999.
10. N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.