LIA-iSmart at TREC 2010 : An Unsupervised Web-based Approach for Filtering Answers

Ludovic Bonnefoy *,** , Patrice Bellot *, Michel Benoit **

Abstract—**Searching for named entities has been the subject of many researches in information retrieval. Our goal in participating in TREC 2010 Entity Ranking track is to look for reconizing any named entity in arbitrary categories and use this to rank candidate named entities. We propose to address the issue by means of a web oriented language modeling approach.**

Index Terms—**siblings entities, characteristics, TREC Entity Ranking, QA**

I. Introduction

Named entities recognition and extraction are central to many works related to information retrieval and to natural language processing. In this work, we addressed two tasks : the aim of the first one is to determine to what extent a candidate answer to a natural language question (usually a named entity) may be associated to a given type of entity. We guess that it could be used to determine whether two named entities are similar or to what extent they are. Many peoples are interested in this issue as shown by the creation, in 2009, of the *Entity* track in TREC [1]. The second issue, and probably the most important one, is to deal with any type of entities. Our goal was to deal with types as broad as "person" or as specific as "scotch whisky distilleries".

The Related Entity Finding task proposed at TREC is defined as finding an answer list of named entities, associated with their homepage, answering to a topic composed of an input named entity, its type (person, location, organization or product) and of a narrative field describing the relation to the input entity. Homepages have to be found in the ClueWeb09 corpus[1] which contains 500 million web pages written in English.

This paper is broken down as follows : in a first part, we present an unsupervised way to measure the degree of membership of any entity to any type. In a second part, we describe the approaches we employed for TREC Entity 2010. Then, we analyze and comment the official results we obtained.

II. An unsupervised measure of membership of a named entity to a given type

We aimed here to determine, in an effective way, in which degree an entity is of a given type without using any other resources than the Web.

We started our work with a study of the web pages returned by commercial web search engines for different types. We noticed that the web pages associated to each type have a specific vocabulary. For example, for the web pages responding to the query "portable mp3 players", some words like "mp3", "music", "capacity", "headphones",. . . are really frequent, compared to their frequency in general web pages.

When we have analyzed the web pages related to specific entities, we noticed that the pages related to them, have a specific vocabulary too (for example for "Winnie the pooh" one can find some high frequency words like "fictionnal", "character", "bear", "friends", "disney",...).

Our last observation was that the words distribution for web pages related to an entity is close to the one of the type which charaterizes it the most (for "iPod" we obtain some specific words like "apple", "mp3", "music", headphones", "media", . . . that look close to the ones for "portable mp3 players").

We deduced from these observations that if we compare the words distribution in web pages related to an entity to the one in web pages related to a given type, we could determine in what extent this entity belongs to this type. The steps are :

- Obtain a first set of web pages related to the type, by querying a web search engine with the type (ex : "science-fiction writers"). This set is called "reference set". Obtain a second set, related to the entity, by querying the web search engine with it (ex : "Isaac Asimov").
- Compute, for each set, its words distribution (smoothed with Dirichlet) :

$$p'(w|s) = \begin{cases} p_s(w|s) \text{ if w is in the set} \\ \alpha_d p(w|C) \text{ otherwise} \end{cases} \tag{1}$$

  where $p'(w|s)$ is the probability of the word w in the set S, $p_s(w|s)$ is the smoothed probability of w, $p(w|C)$ the Laplace smoothed probability of w in a collection C (which consist here in ten percent of the TREC 2010 Entity Track corpus) and α_d is a multiplier. $p_s(w|s)$ and α_d are estimated as :

$$p_s(w|s) = \frac{tf(w,s) + \mu.p(w|C)}{\sum_{w' \in V} tf(w',s) + \mu} \tag{2}$$

$$\alpha_d = \frac{\mu}{\sum_{w \in V} tf(w,s) + \mu} \tag{3}$$

  where $tf(w,s)$ is the term frequency of w in the set s, V is the set of all words w' in s and μ is a multiplier with a value set to 2000 (optimal value according [2] for newspapers and largest collection).
- Compare the words' probability p'_{CNE}, in documents associated to the entity, to the reference one p'_{NE},

* University of Avignon - CERI/LIA,
Agroparc – B.P. 1228, F-84911 Avignon cedex 9, France.
e-mail : patrice.bellot@univ-avignon.fr
** iSmart,
Le Mercure A, F-13851 Aix-en-Provence Cedex 3, France.
e-mails : ludovic.bonnefoy@ismart.fr, michel.benoit@ismart.fr

[1] http://boston.lti.cs.cmu.edu/Data/clueweb09/

associated to the type. For this, we compute the Kullback-Leibel divergence (KLD) between them :

$$KLD(E, type) = \sum_i p'_E(i).log\frac{p'_E(i)}{p'_{type}(i)} \qquad (4)$$

where $KLD(E, type)$ is the Kullback-Leibler divergence for the given entity E and the type, $p'_E(i)$ (resp. $p'_{type}(i)$) is the probability of the i^{th} word in documents associated to the entity E (resp. to the type).

This is a fully automatic method which allows to compute in what extent an entity is of a given type (for any entity/type couple). Then, to answer to the topics of the Entity task, we propose to combine this membership score with a classical relevance score obtained from our Question-Answering (QA) system.

III. Related Entity Finding at TREC 2010

A. System overview

For our first participation to TREC Entity, we adopt (like many other participants[2]) a QA-like approach :

- First, the topic is analyzed in order to extract significant words,
- In a second time, these elements are used to retrieve a set of related web pages,
- Next, some candidate named entities are extracted from this set of web pages. Many teams in TREC Entity used a named entity recognition tool like the Stanford-NER[3] or the LBJ-based NER[4] [4]. Some other teams have tried different ways like using Wikipedia categories (as a complement to a NER tool [5] or not [6]), using ontologies such as DBPedia, Yago or Wordnet [7].
- The next step deals with candidate named entities ranking. Many ways have been explored like the estimation of the probability to get an entity from a topic, by using the word overlap between support documents and topic [9] or by computing the cosine similarity between the homepage and the topic [10], co-occurences between the candidate entity and the source entity [11]. Many other criterions have been used by some other people like entity frequency [12] or the use of hypertext links [13].
- The aim of our final step is to find the named entitie's homepage. Some participants used Wikipedia or other knowledges bases like Freebase [8] or DBPedia [14]. Other works employed a machine learning approach [3]. [14] shows that the use of a confidence score in the homepage finding for re-ranking candidate named entities could bring a significative improvement of the results.

B. Detailed implementation

Our system follows the steps as described above (see figure 1) :

The aim of the first step is to get a set of web pages related to the topic. For this purpose we queried the web search engine Yahoo! with the source entity and all the common and proper nouns extracted from the narrative filed (they were found by means of the TreeTagger[5]). The 100 top ranked web pages were downloaded, cleaned of HTML tags and parsed in sentences. The sentences are then indexed with the search engine Indri[6]. Finally, we queried Indri with the same query that the one used to query Yahoo! and we keeped the 500 top ranked passages.

The Stanford-NER is used for identifying candidate named entities of types ”person”, ”location” and ”organization”. For the type ”product”, we had to design a specific heuristic because the Stanford-NER is not able to deal with it (because it is trained on CoNLL and MUC corpus). First, we get, as candidate products, all proper nouns and word sequences in capital letters (ex: ”Epson Stylus”) which were not recognized by the Stanford-NER. If one of those sequences is followed by a number, we concatenated it to the word sequence (ex : ”Playstation 3”).

For the type ”person”, we maped the different spellings of a candidate entity (*Barichello*, *R. Barichello*) to their guessed canonic form(*Rubens Barichello*) by keeping the most frequent form (in the snippets found by Yahoo!, by querying it with the different spellings). In this way, we could get the entire name when we had the last name only (ex : *Rubens* ⇒ *Rubens Barichello*) and reduce the candidate named entities list by two or three by removing a lot of redundancy.

The first criterion we used to rank the candidates named entities is the compacity score as presented in [15]. It measures the density of the query words around a given candidate entity (a correct answer to a query tends to appear in the texts near of the query words). Compacity is defined as :

$$Compacity(E_i) = \frac{1}{|QW|}\sum_{w \in QW}\frac{Z_w}{R_w + 1} \qquad (5)$$

with QW the set of query words (elements extract from the topic to get the web pages), $|QW|$ the cardinality of this set and w one of them. Let E_i be a candidate named entity, R_w the distance (in number of words) between w and the candidate named entity. Let Z_w be the number of query words between w and the E_i (both included).

In the Entity track, only high level named entity types are given, but we noticed that a fine-grained type is expressed in each narrative indicated in topics field. For each topic, we got it by extracting the first plural common name of the narrative's main clause and adjacents common names and adjectives (for finer types). For example we can determine from the narrative (analyzed by the TreeTagger) *”Scotch (NN) whisky (NN) distilleries (NNS) on (IN end of main phrase) the (DT) island (NN) of (IN) Islay (NN) . (SENT)”* (topic 20 from TREC 2009 Entity Ranking track) the fine-grained type ”distilleries” and with even more precision ”scotch whisky distilleries”.

[2] http ://ilps.science.uva.nl/trec-entity/les/trec2010/trec2010-entity-workshop.pdf

[3] http ://nlp.stanford.edu/ner/index.shtml

[4] http ://cogcomp.cs.illinois.edu/page/software_view/3

[5] http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6] http ://www.lemurproject.org/indri/

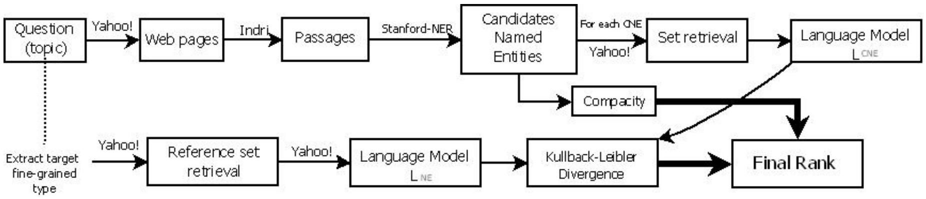

Fig. 1. System Overview

Then, for each candidate named entity, we computed its degree of belonging to this fine-grained type. We downloaded 100 web pages by querying Yahoo! with the fine-grained type name as a query to build the "reference set". For each candidate entity, we built its associated set by querying Yahoo! with the entity alone as a query and download the 10 top ranked web pages. The Kullback-Leibler divergence was then computed between the language models estimate on the sets (see section 2).

We then had to rank the candidate named entities by using all the different scores that we had. We imagined four different ways to rank the candidate named entities corresponding to each run:

- Comp : The first one, which was our baseline, was the use of the compacity only.
- Type : Our second run was to only use the degree of belonging to the target type to rank the entities.
- HM : The third way combined the two previous measures by computing the harmonic mean between the rank of the named entity according to the compacity to its rank according to the Kullback-Leibler divergence.
- ML : The last run that we adopted was a machine learning approach to estimate the weight of each score in a linear combination.

  For each candidate named entity we wanted to combine four scores : the best compacity score of the candidate entity $bCompacity(E_i)$, the degree of belonging to the type $KLD(E_i, type)$, its idf ($idf(E_i) = log\frac{N}{n_i}$ with N the total number of documents and n_i the number of documents in which entity E_i appears) in the 500 top ranked passages retrieved by Indri and the best score $bP(E_i)$ of passages where the entity was found. For learning, we extracted a set of 45 topics from TREC QA 2007&2006 list questions (with the same proportions for each coarse-grained target type of questions as Entity Ranking 2009). Then, we got all the named entities in output of our system for this 45 topics and we have assigned, for each named entity, the class "Yes" if it is a correct answer or "No" otherwise.

$$\begin{aligned} S(E_i) = \lambda_1 bCompacity(E_i) + \lambda_2 KLD(E_i, type) \\ + \lambda_3 idf(E_i) + \lambda_4 bP(E_i) \end{aligned} \quad (6)$$

  With this set of correct and incorrect answers, we trained a multilayer perceptron classifier (Weka's one[7]) with each score as a feature. Lastly, we ranked named entities according to the score of membership to the class "Yes" given by the classifier (a named entity with a confidence of 80% for "Yes" will have a higher rank than a named entity with 50%).

Then, we had to found the homepages of candidate named entities in ClueWeb09. We noticed that web search engines deal with the word "homepage" in a specific way. It is why we asked Yahoo! with queries *"candidate_named_entity" homepage"* (ex : "Lufthansa homepage") and hence retrieved the five top ranked web pages. We deleted, from this set of candidate homepages, the ones which did not have the characteristics of an homepage. For this purpose we trained a SVM classifier on 7-web genre collection[8] (the genre are : "blog", "shop", "personal homepage", "frontpage", ...). Features used are word frequencies (any of them [16]), POS frequencies, sentence, word, document average size (in words), . . . [17] and HTML tag count [18]. The trained classifier has a precision of 99.7% on a 10-folds cross validation. If a web page was categorized as an homepage and its url is present in the ClueWeb09 we keep this one (if it the case of more than one web page we kept the top ranked one).

IV. Results

In this part are presented our official results for the Related Entity Finding task at TREC Entity 2010. This evaluation is made with 48 (new) topics and four different evaluation measures are used : precision at 10 elements, map, nDCG@R and Rprec with R the number of existing correct answers. Table 1 shows the average results that we obtained for each of our methods and the best and median results (if reported).

These first global results allow us to draw preliminary conclusions. The first observation is that, on 48 runs, ours are 27,29,36 and 40 which position our best run close to the median one. In this 48 runs, 19 of them are manual or partially manual (ex : the keywords selection from the topic). If we only compare our 4 runs to the automatic runs, they are ranked 14,16,18 and 21-th (among 29 runs), which ranks our best run just above the median one.

It is difficult to explain why a method does better than an other one (given the many parameters we have to consider), but there are some clear ways to improve our system.

[7] www.cs.waikato.ac.nz/ml/weka/
[8] http://www.webgenrewiki.org/index.php5/Genre_Collection_Repository

	Our runs					
Metric	Compacity alone	Type alone	HM Comp./Type	Learned Combi. Comp. /Type	Best	Median
P@10	.0468	.0213	.0362	**.0532**		
nDCG@R	.0737	.0428	.0610	**.0766**	≈ .39	.0857
map	.0261	.0129	.0200	**.0305**		
Rprec	.0463	.0189	.0373	**.0591**		

TABLE I

Official evaluation of our 4 approaches (Comp, Type, HM and ML (see 3.2)) for the Entity track at TREC 2010 for precision at 10 elements (P@10), nDCG@R, map and Rprec compared to the best and the median official runs.

	Our runs					
Topic	Compacity alone	Type alone	HM Comp./Type	Learned Combi. Comp. /Type	Best	Median
21	.0166	**.0213**	.0177	0	.4094	.0260
22	.0954	.0852	.096	**.1009**	.2818	.1008
30	.2342	.1941	.1705	**.3155**	.3739	.0810
44	0	0	0	0	.7099	0
49	.139	.1295	.1321	**.1451**	.3081	.1233
66	0	0	0	0	.4306	0
67	**.0443**	*.0443*	.0397	.0260	.4526	.0725
Mean	.0756	.0678	.0651	**.0839**	.4238	.0577

TABLE II

Official evaluation for type "location" for our 4 approaches (Comp, Type, HM and ML (see 3.2)) for the Entity track at TREC 2010 with nDCG@R and compared to the best and median official runs.

By comparing our approaches, we can see that using the degree of membership to the target type (run Type) only does not allow to rank effectively named entities. This is obvious because in this way, their context are not take into account.

Secondly, one can observe a diminution of our baseline results (run Comp) when we combine the membership score over the type with compacity by means of straightforward harmonic mean (run HM) giving an equal weight to both parameters. If we look at the results for each topic, we can see that for some of them, the use of a language modeling approach to compute the named entities relatedness / membership to the target type is not relevant. The reason is that, for some target types as "students" or "winners", their language model are not enough specific (too close to a generic model of the world) and, moreover, they do not really characterize a named entity. Maybe could we penalize the weight of this score depending on the degree of relevance of the target type (by taking into account, for example, the distance from its language model to a language model of the world).

Lastly, we can see that the use of a machine learning approach to estimate the weight of different informations to rank named entities (run ML) brings significant improvements : +14% for P@10, +4% for nDCG@R, +17% for map and +27% for Rprec. These results confirm our intuition, the use of a degree of membership to the target type of a candidate named entity can improve results.

Tables 2,3 and 4 show results for each broad type (except for "product" because only one topic has been evaluated) for the nDCG@R measure.

The first important point that we can notice is that, for 10 of the 47 topics, more than half of the runs get zero for this metric while the best systems still have good results. Some topics were more difficult than others for the main reason that answers could not be found in Wikipedia pages[9]. Even if we don't use the Wikipedia pages specifically, we used them in a indirect way, because they tend to appear in the top five web pages retrieved by a commercial web search engine. Some participants were able to exploit other resources like the input entity's homepage or knowledge bases (Freebase[10], DbPedia[11], ...). No information are given about the best runs and we doesn't know if they're manual or automatic...

The second important point is that, for two of the three types ("person" and "organization"), we obtain results above median. We can also see that we get better results for "person" then "location" and finally "organization". This order suggest that the results are significantly impacted by the precision of the Stanford-NER for each type because they are directly correlated [19]. Moreover, our general good results for "person" (better than 60% from the median) seem to show that our way to find canonic form of candidate named entities is interesting.

[9] http ://ilps.science.uva.nl/trec-entity/files/trec2010/trec2010-entity-workshop.pdf

[10] http ://www.freebase.com/

[11] http ://dbpedia.org/About

	Our runs					
Topic	Compacity alone	Type alone	HM Comp./Type	Learned Combi. Comp. /Type	Best	Median
24	.1873	0	0	**.2743**	.4348	0
37	0	0	0	0	.6518	0
38	**.3444**	0	.2397	.3046	.6490	.2193
41	**.2979**	.0289	.2105	.1940	.4941	.1614
43	**.0232**	.0502	.0209	.1440	.6664	.1172
52	**.1586**	.0221	.1062	0	.5829	.0551
55	0	0	0	0	.7661	.0500
57	0	0	0	0	.4693	0
Mean	**.1264**	.0127	.0722	.1146	.5893	.0754

TABLE III

Official evaluation for type ”person” for our 4 approaches (Comp, Type, HM and ML (see 3.2)) for the Entity track at TREC 2010 with nDCG@R and compared to the best and median official runs.

	Our runs					
Topic	Compacity alone	Type alone	HM Comp./Type	Learned Combi. Comp. /Type	Best	Median
23	.1015	0	.055	**.1115**	.4855	.0550
25	0	0	0	**.1365**	.5718	.0476
26	.1223	**.3273**	.2821	.2106	.3998	.0732
27	0	0	0	0	.7638	0
29	**.0087**	0	.0073	.0076	.5216	.2248
31	.0383	**.0730**	.0562	.0497	.4627	.0730
32	0	0	0	0	.4522	0
33	.2119	.1798	**.2275**	.0601	.3403	.1111
34	.0349	**.045**	.0381	.0381	.6573	0
36	0	0	0	0	.3026	0
39	.0504	.0325	**.0637**	0	.5292	.1404
40	0	0	0	0	.5016	.1361
42	.0182	.0138	**.0246**	.0108	.3932	.0682
45	**.1203**	0	0	.0907	.6920	.1203
47	0	0	0	0	.7800	.0913
48	.08	.0881	**.0939**	.0698	.7628	.1870
50	**.0682**	.0484	.0595	.0674	.4626	.1299
51	.1713	.1628	.1596	**.251**	.5365	.3807
53	0	0	0	0	.5877	0
54	.0488	**.1198**	.0337	.0785	.3175	.1047
56	0	0	0	0	.4171	0
58	0	0	0	0	.3667	0
60	0	0	0	0	.6988	.0212
61	.0373	.0301	.0184	**.0502**	.7115	.0502
62	0	0	0	0	.4715	.1850
63	.1466	.1479	.1361	**.2004**	.5255	.1888
64	**.2673**	0	.255	.0891	.6814	.1621
65	0	0	0	0	.6131	0
68	.0157	.0404	.0242	**.0444**	.4779	.0978
69	.3050	.0428	.1664	**.4165**	.7326	.2880
70	0	0	0	0	.5312	0
Mean	.0596	.0436	.0549	**.0640**	.5403	.0947

TABLE IV

Official evaluation for type ”organization” for our 4 approaches (Comp, Type, HM and ML (see 3.2)) for the Entity track at TREC 2010 with nDCG@R and compared to the best and median official runs.

Before	After
nico rosberg	nico rosberg
eddie irvine	**felipe massa**
felipe massa	muhammad ali
rubens barrichello	sebastian vettel
johnny herbert	joe louis

TABLE V

5 top output elements of our system, before and after homepages finding. Bold answers are correct. Topic 1, Entity track, TREC 2009 training set.

Measure	Comp + baseline homepages	Comp + HPFindingGoogle	Improvement
P@10	.0468	.0574	+22%
nDCG@R	.0737	.0941	+28%
map	.0261	.0489	+87%
Rprec	.0463	.0724	+56%

TABLE VI

Measure of the impact of homepage finding on global results. Comparison on our run Comp of our baseline to find the homepages and the one of Potsdam university (HPFindingGoogle) for the 4 officials measures.

The third point is that the use of a divergence between language models for quantifying the degree of membership of a named entity to the target type, combined in a effective way with the compacity, (run ML) brings in most cases, significant improvements (up to 10% pour locations and to 8% for organizations on average). Topics for "person" are exceptions and the use of the compacity only (run Comp) seems to be better. If we look to the results obtained by the divergence only (run Type) for "person" we can see that they are very low. Indeed, for the two other types, although they are the worst, they don't be lower than 70% of our best result. For "person" it falls until 10%. . . This points out that this way is not adapted and it is the reason that explains that combining it with compacity do not improve our global results. Again, the reason is that most of the fine grained types extracted from query topics are not enough specific and do not really characterize a named entity (ex: "members") while we have more interesting and discriminatory types for "organization" and "location" (ex: "countries", "airlines", . . .)

The last important point is the very low capacity of our system to find named entities' homepage. This have for direct consequence that we lost a lot of good answers because we're not able to find their homepage. On table 5, we can see that this particular point, on the first topic of TREC Entity 2009 training set, decrease dramatically the precision at 5 elements from 80% to 20%. To confirm it, the university of Potsdam allows us to use their homepage finding module HPFindingGoogle [3]. To rank candidate homepages, it compute for each one a set of 17 different features and combine them with the best weight configuration found with a genetic algorithm on training datas. Table 6 presents the results obtained with it on our run with the compacity only and we can see a very significant improvement (from 22% to 87%, depending on the metric). This is clearly one of the specificities of the task that we neglected and plan to improve for future participations.

V. Conclusion and Future work

We presented here a method to estimate to what extent an entity belongs to a given type in an unsupervised way. We have shown that this information may be taken into account in a QA-like system, to answer to the Related Entity Finding task more efficiently.

In a future participation, better results could be reached by better considering some specificities of the task and by employing a better homepage classification module. Effort has be also to be done for better combining relevance scores by means of a better and larger training corpus.

Lastly, we whish to explore whether one can identify and extract some ontological characteristics of named entities by means of our approach.

References

[1] Balog K., Serdyukov P. and de Vries A.P. : *Overview of the TREC 2010 Entity Track*, NIST Special Publication : TREC 2010.

[2] Chen S. F. and Goodman J. : *An empirical study of smoothing techniques for language modeling*, 34th Annual Meeting of the ACL, Vol. 13 (1996), pp. 310-318.

[3] Hold A., Leben M., Emde B., Thiele C. and Naumann F. : *ECIR - A Lightweight Approach for Entity-Centric Information Retrieval*, NIST Special Publication : TREC 2010.

[4] Vechtomova O. : *Related Entity Finding: University of Waterloo at TREC 2010 Entity Track*, NIST Special Publication : TREC 2010.

[5] Wang Z., Tang C., Sun X. and Ouyang H. : *PRIS at TREC 2010: Related Entity Finding Task of Entity Track*, NIST Special Publication : TREC 2010.

[6] Kaptein R. and Koolen M. : *Result Diversity and Entity Ranking Experiments: Anchors, Links, Text and Wikipedia*, NIST Special Publication 500-278 : TREC 2009.

[7] Serdyukov P. and de Vries A. : *Delft University at the TREC 2009 Entity Track : Ranking Wikipedia Entities*, NIST Special Publication 500-278 : TREC 2009.

[8] Bron M., He J., Hofmann K., Meij E., Tsagkias M. and Weerkamp W. : *The University of Amsterdam at TREC 2010 Session, Entity and Relevance Feedback*, NIST Special Publication : TREC 2010.

[9] Wu Y. and Kashioka H. : *NiCT at TREC 2009 : Employing Three Models for Entity Ranking Track*, NIST Special Publication 500-278 : TREC 2009.

[10] Zhai H., Cheng X., Guo J., Xu H. and Liu Y. : *A Novel Framework for Related Entities Finding : ICTNET at TREC 2009 Entity Track*, NIST Special Publication 500-278 : TREC 2009.

[11] Bron M., Balog K. and de Rijke M. : *Related Entity Finding Based on Co-Occurance*, NIST Special Publication 500-278 : TREC 2009.

[12] Wang D., Wu Q., Chen H. and Niu J. : *A Multiple-Stage Framework for Related Entity Finding: FDWIM at TREC 2010 Entity Track*, NIST Special Publication : TREC 2010.

[13] Kaptein R., Koolen M. and Kamps J. : *Result Diversity and Entity Ranking Experiments: Anchors, Links, Text and Wikipedia, University of Amsterdam*, NIST Special Publication 500-278 : TREC 2009.

[14] Wu Y., and Kawai H. : *NiCT at TREC 2010: Related Entity Finding*, NIST Special Publication : TREC 2010.

[15] Gillard L., Sitbon L., Blaudez E., Bellot P. and El-Beze M.: *Relevance Measures for Question Answering, The*

LIA at QA@CLEF-2006, Lecture Notes in Computer Science,4730/2007, Evaluation of Multilingual and Multi-modal Information Retrieval , p. 440 449, 2007.

[16] Stamatatos E., Falotakis N. and Kokkinakis G. : *Text genre detection using common word frequencies*, Proceedings of the 18th conference on Computational linguistics - Vol. 2 (2000), pp. 808-814.

[17] Dewdney N., VanEss-Dykema C. and MacMillan R. : *The form is the substance: classification of genres in text*, Annual Meeting of the ACL, Proceedings of the workshop on Human Language Technology and Knowledge Management - Volume 2001, Article 7.

[18] Levering R., Cutler M. and Yu L. : *Visual Features for Fine-Grained Genre Classification of Web Pages*, Hawaii International Conference on System Sciences, Proceedings of the 41st Annual 2008, pp. 131 - 131.

[19] Finkel J. R., Grenager T. and Manning C. : *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling*, Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370