Bhilai Institute of Technology Durg at TAC 2010: Knowledge Base Population Task Challenge

Ani Thomas, Arpana Rawal, M K Kowar, Sanjay Sharma, Sarang Pitale, Neeraj Kharya

ARPANI NLP group, Research and Development Cell, Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, Chhattisgarh, India

{aniarp@rediffmail.com, www.bitdurg.org}

ABSTRACT

The present communication aims to report the TAC forum about the system-incorporated towards Entity-Linking task. The noun-phrases relevant to the search term were used to correlate the entity-relevant happenings mentioned in the source documents to that of the entity-relevant information in the knowledge-base. Care was taken in identifying the abbreviation and full form variants of search names if present in the same set of queries or in the provided Knowledge base as Wikipedia structured nodes. The subsequent knowledge-base node hits led to the ranking of their respective content for retrieving entity-linking node-id responses. Three different runs were submitted the Task Challenge following the different ways to search for the most meaningfully relevant information about the entities mentioned in Entity-Linking Target list from the Knowledge-base of the track. The team's spirits feels elevated at the thought of using the free text from the Wikipedia pages associated with the knowledge base nodes, while building the evaluation model as it adds to the robustness and reliability of the system even at the time of inaccessibility of the Web.

1. INTRODUCTION

This piece of work is performed as a part of Entity Linking task in KBP track that posed up the problem to design some kind of automated Information Relevance ranking system that could meaningfully augment a reference knowledge-base component to the entity-names discovered from the inputted newswire articles. So, the infrastructure had to be designed to initially fetch the input corpus consisting of those newswires, whose names are given in the collection of 2250 Entity-linking queries from a total collection of 2255 no. of articles corresponding to newswire groups as apw, xin, eng, nyt, etc. For instance, document files in .sgm formats as AFP_ENG_20070414.0278.LDC2009T13 had to be fetched from TAC 2010 KBP Source Data (Corpora LDC2010E12). This also required at least 150 GB of storage for both the collections: newspaper articles and a huge set of Wikipedia Info boxes acting as Reference knowledge bases consisting of approximately 8 lakhs of the knowledge-base entity-relevant components. The targeted entities put to contextual IR task belonged to three entity types: PERSON (PER), ORGANIZATION (ORG) and GEO-POLITICAL ENTITY (GPE), although the group aimed to build a generalized prototype design without using any of the explicit web-ontology or domain-specific controlled dictionaries.

2. THE AVAILABLE DATA SETS

TAC-KBP participants were initially provided with a set of 818,741 entity-descriptions in the knowledge-base corpus of the Wikipedia. The team could browse the uniform structure of these entity descriptions i.e. a name string appearing in wiki title tagged format, a pre-assigned entity type as PER / ORG / GPE /

UKN, a unique node-id to each of the knowledge-base components and some text relevant to that knowledge-base component.

2.1 STRUCTURED REFINEMENTS UPON KNOWLEDGE BASE

Looking at the knowledge-base at a glance, the only innovative thought that could initiate the entity-oriented search, was finding all possible string-hits from the <wiki-title> string patterns. This was possible by mapping the whole set of knowledge bases into an indexed form comprising of only <wiki-title> sentential fragments. However, if an acronym form of the entity-mention-name is needed to be searched across the knowledge-base index, the knowledge-base nodes with <wiki-title> full-form of that abbreviated name string shall also not be skipped from being selected as candidate-nodes for relevance ranking procedures and is included in some of the runs generated.

2.2 FETCHING THE NEWSWIRE CONTENTS

In later stages of Task Allotment period, the group was provided with an evaluation window period of seven days to prove the team's potential along with a download of Entity-Linking test-data release from the end of Linguistic Data Consortium including 2250 queries in an .xml file. These queries consisted of a name string and a document-id in the test collection. The purpose of the associated document is to provide context that might be useful for disambiguating the name-string's reference in the knowledge-base. Meanwhile, the group already had an access to electronically shipped TAC 2010 KBP Source Data that contained a series of newspaper articles from various newsgroup sources. Now, as a preliminary task, approximately 2255 newswire articles were fetched for resolving 2500 queries, taking into consideration that the same document was referenced for two or more different queries at the input end.

2.3 FORMATTING THE ENTITY-LINKING QUERIES

The entity-linking queries were further pre-processed to extract only the entity-mentioned names as the list of name-strings had two major role plays in the KBP-Entity-linking task.

2.3.1 Selection of Candidate Nodes:

First, the names could fetch the selective knowledge-base components by providing hits to the string patterns included in the <wiki-title> portions of the knowledge-base index, as introduced in section 2.1. Hence, streams of knowledge-base sub-indexes were prepared for each fired-up entity query that provide a baseline for exploring one of the candidate-content bearing maximum relevance with the entity-name. The figure 1 illustrates the query-wise generation of knowledge-base index, by attempting either total or partial string-pattern hits, taken as a criteria for exploring the performance of TAC-KBP runs, elaborated in the section 4. The query considered here belongs to query-id **EL0011110** and query name "ICD" of the input data set.

2.3.2 Exploring the Contextual Reference in Newswire:

Secondly, these entity-mention-names could initiate the contextual search within the associated document file fetched in section 2.2 so that some meaningful information can be revealed as to what kind of event or happening took place in the published article in relation to that entity. This context serves the purpose of disambiguating the embedded texts in the candidate-nodes of the generated knowledge-base sub-index owing to the underlying fact that same entity appeared in multiple queries using different document-ids.

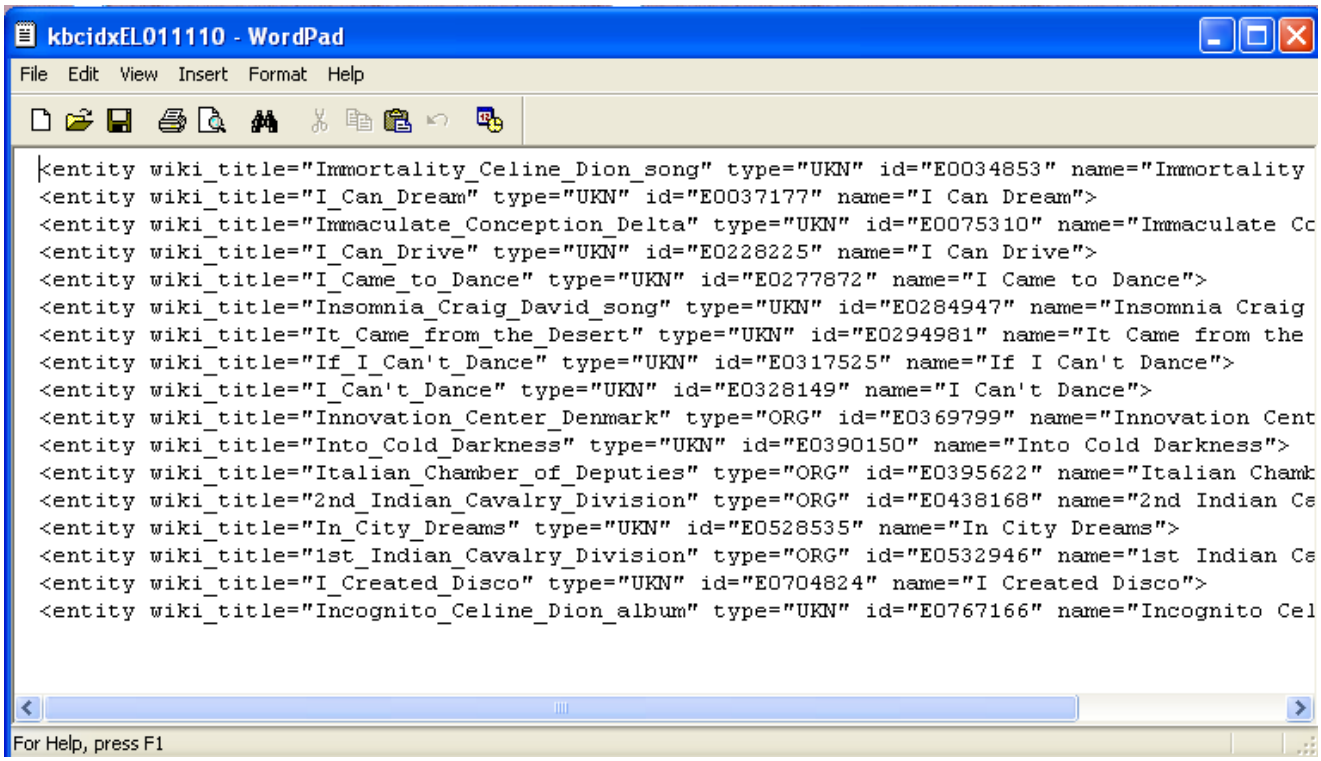


Figure 1: Query-wise generation of knowledge-base index for query-id "EL0011110" and name "ICD"

3. THE PROPOSED APPROACH

The experiments carried out by the research group outline the entity-linking task by funneling the relevant knowledge-base nodes at different conceptual depths as illustrated in figure 2.

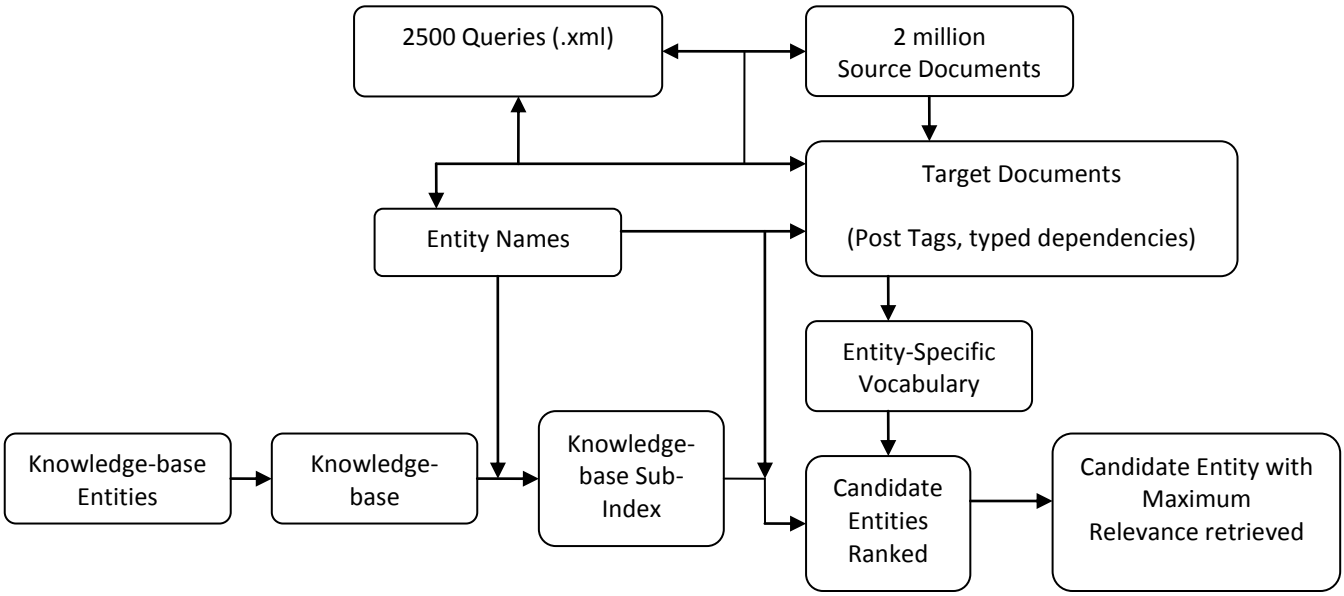


Figure 2: Process Flow Diagram for TAC-KBP 2010 task challenge

3.1 USE OF PART-OF-SPEECH TAGS

As for any concept space to understand at machine's end, the governing parameters are contributed by the extraction of Noun / verb phrases parsed from natural language syllabus text-strings. Stanford NLP group offers an efficient Part-Of-Speech Tagger among the competent ones put as open-source tools by the computational linguistic communities. The POS Tagger reads English text and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally applications use more fine-grained POS tags like 'noun-plural'. Similarly, in the current scenario, all possible Noun variants in the pool of Part-Of-Speech tags shall be used to extract the entity-specific noun phrases of the happenings described in newswire.

3.2 EXPLORING THE TEXT SEMANTICS

The whole task plays around the only objective of finding the degree of semantic similarities between the mentioned news article and the referenced knowledge-base node. The team found after a vast in-depth literature survey, that it is the Stanford typed-dependencies that are designed to provide a simple description of the grammatical relationships in a sentence that can easily be understood and effectively used by computational linguist-expertise who wishes to extract meaningful textual relations. The most appealing feature, with which it excites the community, is that it represents all sentence relationships uniformly as typed dependency relations between pairs of words - simple, uniform representation that can be tailored to any application realm in Natural Language Processing tasks. However, the team realized that news-article needed some degree of pre-processing to be enabled to fed to this selected NL Parser.

3.3 EXTRACTION OF ENTITY-SPECIFIC VOCABULARY

The Stanford typed-dependencies of the newswire articles are now suitably manipulated so that only noun-phrases, may be existing in form of single-word nouns or n-word compound noun phrases in shape of n-gram patterns, are generated in series, revealing semantically related chains of nouns that dominate the major event vocabulary happening in that published article. For accomplishing the above, at one end, Part-Of-Speech Tagged form of the news-text is fetched and at the other end, the seed-words that exist as query-names along with the alias-names (if any) trigger the generation of these noun-phrase chains. For instance, the noun-phrases confining to the major news-domain vocabulary can be viewed in figure 3 for the news-article, ENG-NG-31-142900-10121571.sgm document associated for one of the inputted queries, for the query-id: EL000848.

4. THE KBP EVALUATION SETUP

For finding the results of the runs, the relevant KB components corresponding to any query had to be fetched from approximately 8 lakh components otherwise the job would never finish within the speculated time period.

4.1 FETCHING THE KNOWLEDGE BASE COMPONENTS

The initial work done was to fetch the entity wiki title list of the entity names and its full form or short form appended to the queries as already explained in the previous section. The kb components linked to these entity wiki titles were evaluated one at a time with the corresponding newswire article to provide

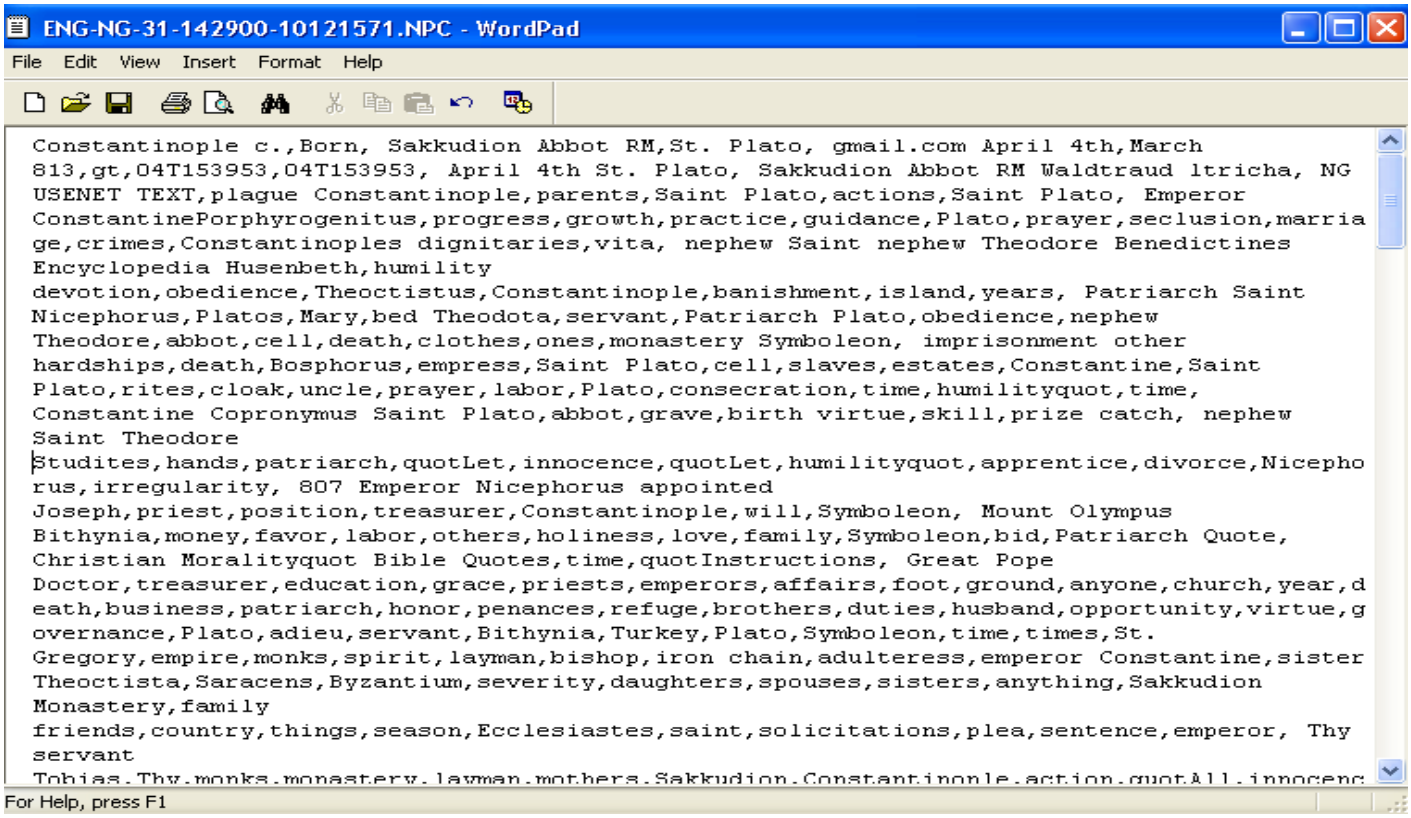


Figure 3: Entity-specific vocabulary generated for ENG-NG-31-142900-10121571.sgm source document

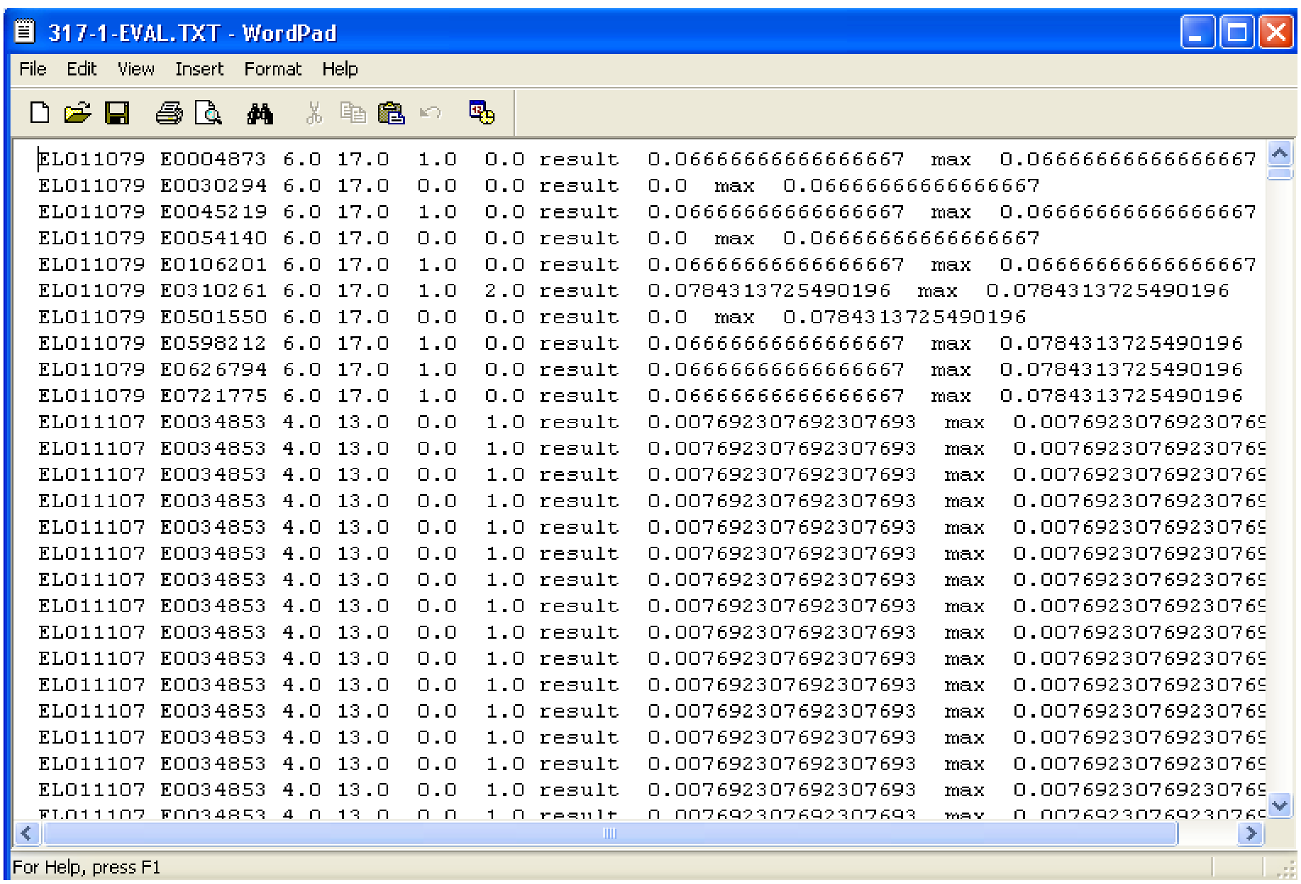
the level of relevance of that knowledge-base components with respect to the newswire document provided in the query. The extended version of this module was to include the name as well as the last name for each and every name retrieved as full form / short form in the grown query list. The knowledge-base components which contained only the last names were also searched and if found, it was appended to the entity wiki title list.

4.2 KBP EVALUATION METRICS

The noun-phrase chains formed from newswire articles are matched with each knowledge-base component to find how closely related they are. The NIL Entity-linking results were contributed before performing any evaluation if no knowledge-base component was retrieved corresponding to the target entity name in a query. The noun phrases tagged by _NNP extracted from Stanford P-O-S Tagger was given 80% thematic weightage as they contribute to the Proper Nouns in the subject of domain being discussed in the news event whereas other words like common-nouns (single word or compound words) if matched to the newswire was processed by giving only 20% contextual weightage. Hence, all the Proper Noun phrases (written in Uppercase style, *upcnt*) and Common Nouns (written in Lowercase forms, *lowcnt*) were counted for from the newswire articles and then matched with those of corresponding knowledge-base components (identified as *mupcnt* and *mlowcnt* measures) to realize 80% or 20% contributive portions for the following evaluation metric described as the algorithm below in figure 4. The matching of each of the knowledge-base components retrieved with respect to the chain of noun phrases extracted from the given newswire article for queries EL011079 and EL011107 is shown in figure 5 and the automatic result generated by finding the maximum score of each individual query is shown in figure 6.

```
if(upcnt==0)
        result=mlowcnt/lowcnt;
    else if(lowcnt==0)
            result=mupcnt/upcnt;
    else
    result=(0.8*(mupcnt/upcnt)+0.2*(mlowcnt/lowcnt))/2;
if(maxresult<result) { maxresult=result;
    maxkbcid=kbcid;
    }
```

Figure 4: Proposed Evaluation Metric for TAC-KBP task



317-1-EVAL.TXT - WordPad

```
EL011079 E0004873 6.0 17.0  1.0  0.0 result  0.0666666666666667  max  0.0666666666666667
EL011079 E0030294 6.0 17.0  0.0  0.0 result  0.0  max  0.0666666666666667
EL011079 E0045219 6.0 17.0  1.0  0.0 result  0.0666666666666667  max  0.0666666666666667
EL011079 E0054140 6.0 17.0  0.0  0.0 result  0.0  max  0.0666666666666667
EL011079 E0106201 6.0 17.0  1.0  0.0 result  0.0666666666666667  max  0.0666666666666667
EL011079 E0310261 6.0 17.0  1.0  2.0 result  0.0784313725490196  max  0.0784313725490196
EL011079 E0501550 6.0 17.0  0.0  0.0 result  0.0  max  0.0784313725490196
EL011079 E0598212 6.0 17.0  1.0  0.0 result  0.0666666666666667  max  0.0784313725490196
EL011079 E0626794 6.0 17.0  1.0  0.0 result  0.0666666666666667  max  0.0784313725490196
EL011079 E0721775 6.0 17.0  1.0  0.0 result  0.0666666666666667  max  0.0784313725490196
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
EL011107 E0034853 4.0 13.0  0.0  1.0 result  0.00769230769230769 3  max  0.0076923076923076
```

For Help, press F1

Figure 5: Knowledge-base Node Ranking Evaluation Metric for one batch of queries.

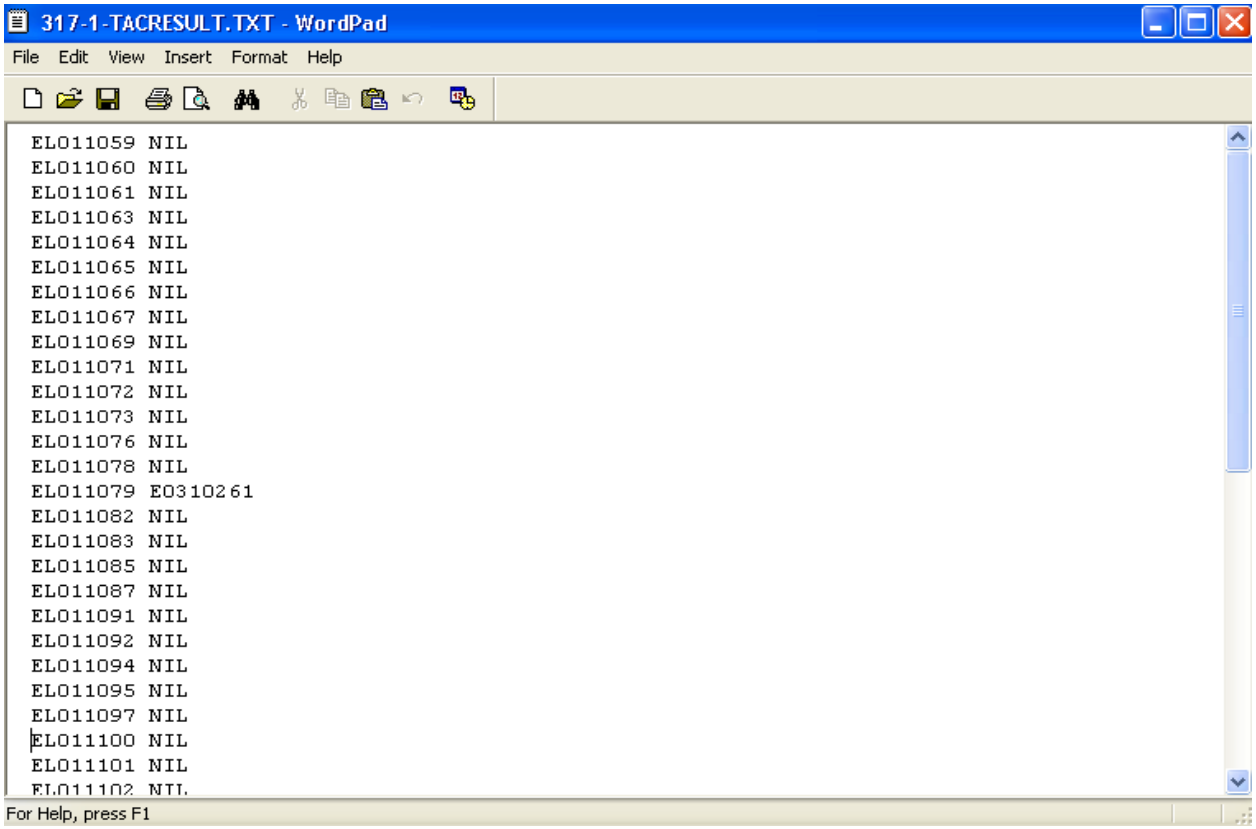


Figure 6: TAC-KBP Entity-Linking Results for the above batch of queries.

5. THE TAC RUN DESCRIPTIONS

The different TAC runs are based on the following criteria:

1. The entity names in the input query were modified by including their name aliases that included both, expanded forms for the abbreviated forms of the strings and vice-versa. However, the search for name-aliases were performed with in the mentioned queries, this was done as a recall to the point given in one of the declared statement of Preliminary Task-definition document, (published version 1/15/2010) that entity queries may be provided using different name variants. In this way, all possible semantically relevant knowledge-base nodes were expected to get fetched as candidates for ranking. The team called this modified version of expanded queries as modified version 1.
2. Another variation in the approach was thought, as to why not to search of abbreviation forms and / or expanded forms of the query name strings from the <wiki title> name-string portions of the knowledge-base components and then proceed for the selection process of candidate knowledge-base nodes. The team used this logic to frame another form (modified version 2) of the inputted query with still more aliases appended in the list of entity-names to be searched for relevance among knowledge-base as illustrated in figure 7.
3. Further, it was assumed that IR performance measure may seem to increase if the last portion of n-gram name strings is also used as a variant in string-matching procedures to identify knowledge-base candidates in wider dimensions. Hence, the team sought for generating two versions of generating knowledge-base candidates, version 1 without incorporating the mentioned processing and version 2, incorporating the feature.

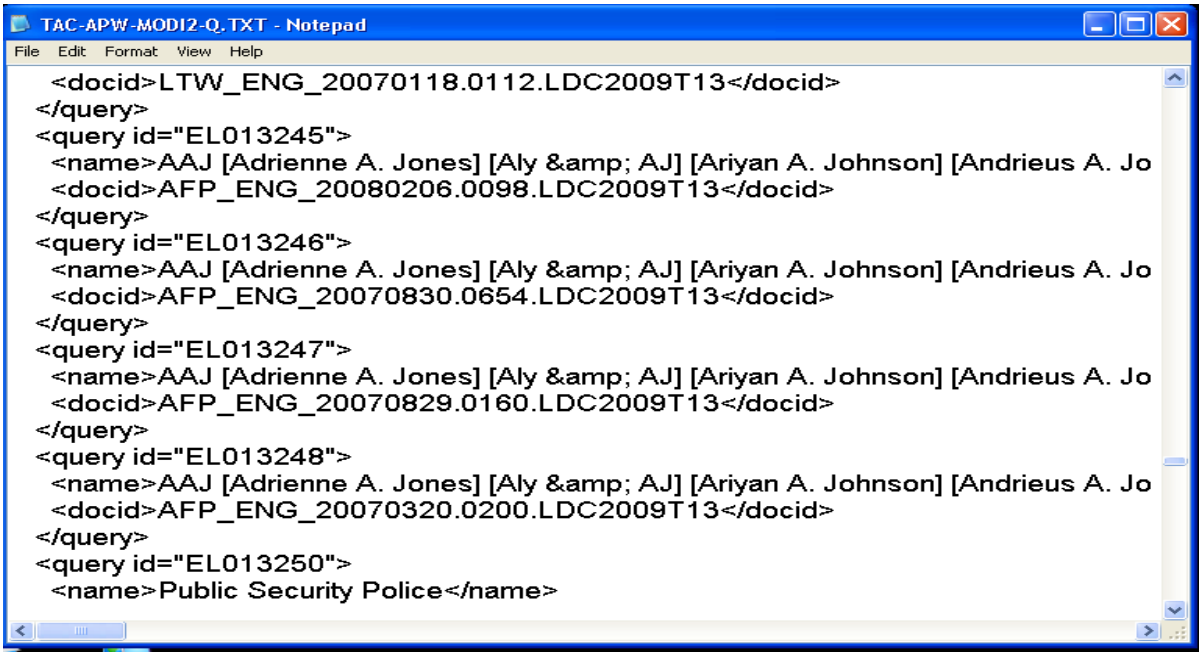


Figure 7: TAC-KBP Modified Version 2 of Entity-Linking queries.

With the above laid presumptions, the first TAC run was initiated with an expanded query consisting of entity names along with their aliases, i.e. modified version 1 of the query. At the other end, knowledge-base components were retrieved using version 2 style of generating the knowledge-base candidates. While, the second TAC run contained the modified version 2 of the queries hitting the knowledge-base in such a way that the knowledge-base components are again extracted using version 2 style of generating the knowledge-base candidates from approximately 8 lakhs of the total knowledge-base entity references. As an alternative evaluation methodology, the third TAC run was initiated with modified version 2 of the queries hitting the knowledge-base but this time, retrieving the knowledge-base components using version 1 style of generating the knowledge-base candidates. The tables below show the obtained micro-averages for the above mentioned three approaches upon the final test KBP corpus. It may be noted that TAC run 3 instance performs well, as version 1 generation style of knowledge-base candidates are found to give less number of misleading referential nodes particularly in linking PER entity nodes and hence highest performance accuracy relative to other entity types. It may also be noted that KBP evaluation of ORG entity types fairly perform well with the micro-averaging scores found in the TAC-KBP entity-linking Gold Standard result list.

TAC 2010 KBP Runs	2250 Queries	1020 Non-NIL	1230 NIL
RUN 1	0.6311	0.4961	0.7431
RUN 2	0.6320	0.4971	0.7439
RUN 3	0.6347	0.4814	0.7618

Table 1: Entity Linking Micro-Averaging Scores for TAC 2010 KBP Evaluation Corpus.

TAC 2010 KBP Runs	750 ORG Queries	304 ORG Non-NIL	446 ORG NIL
RUN 1	0.5787	0.3125	0.7601
RUN 2	0.5813	0.3158	0.7623
RUN 3	0.5880	0.3224	0.7691

Table 2: Micro-Averaging Entity Linking Scores for 'ORG' Entities in TAC 2010 KBP Evaluation Corpus.

TAC 2010 KBP Runs	749 GPE Queries	503 GPE Non-NIL	246 GPE NIL
RUN 1	0.5127	0.6044	0.3252
RUN 2	0.5127	0.6044	0.3252
RUN 3	0.5100	0.5686	0.3902

Table 3: Micro-Averaging Entity Linking Scores for 'GPE' Entities in TAC 2010 KBP Evaluation Corpus.

TAC 2010 KBP Runs	751 PER Queries	213 PER Non-NIL	538 PER NIL
RUN 1	0.8016	0.5023	0.9201
RUN 2	0.8016	0.5023	0.9201
RUN 3	0.8056	0.5023	0.9257

Table 4: Micro-Averaging Entity Linking Scores for 'PER' Entities in TAC 2010 KBP Evaluation Corpus.

6. CONCLUSION

The developed KBP evaluation system being a beginning of its prototype design series, delivered a high-performance system without having the need to explore the explicit-ontology help-in-aid, the World-Wide-Web. This encourages the research group to work upon increasing the robustness of module design in generating the knowledge-base candidates by customizing the retrieved modified versions of query formats for each of the entity-types so that promising results above 0.9000 can be expected further in all the entity-type (NIL and NON-NIL) categories as compared to those found in tables 1,2, 3 and 4. The group also could not deal with the NIL result generation where one or more knowledge bases were found to be matched against any given query.

Acknowledgements

This work was carried out in Research and Development Cell of Natural Language Processing Laboratory, as a part of ongoing research in the unfolding fields of Subjective Question-Answering and Text-Document Relevance Ranking Domains through text-mining techniques. The research is currently registered in Chhattisgarh Swami Vivekanand Technical University. The authors sincerely thank the aspiring team members of final year and pre-final year students, Harsh Bafna, Prafulla Malviya, Alok Kumar Singh and Akash Singhal for their strenuous efforts in module executions to extract the TAC-KBP runs within scheduled time deadlines.

References

1. *D. Pinto, M. Tovar, D. Vilariño, B. Beltrán, J. Somodevilla, (B. Autonomous University of Puebla),* **BUAP_1: A Naïve Approach to the Entity Linking Task**, Proceedings of Text Analysis Conference, 2009, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.
2. *M. Honnibal, R. Dale (Macquarie University),* **DAMSEL: The DSTO / Mcquarie System for Entity-Linking**, Proceedings of Text Analysis Conference, 2009, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.
3. *D. Bikel, V. Castelli, R. Florian, D. Han (IBM TJ Watson Research Center),* **IBM: Entity-Linking and Slot-Filling through Statistical Processing and Inference Rules**, Proceedings of Text Analysis Conference, 2009, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.

4. *H. Srinivasan, J. Chen, R. Srihari (Janya Inc),* **Janya**: Cross document person name disambiguation using entity profiles, Proceedings of Text Analysis Conference, 2009, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.
5. *X. Han, J. Zhao (Chinese Academy of Sciences,* **NLPR_KBP in TAC 2009 KBP Track: A Two-stage Method to Entity-linking**, Proceedings of Text Analysis Conference, 2009, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.
6. *S. Li, S. Gao, Z. Zhang, X. Li, J. Guan, W. Xu, J. Guo (Beijing University of Posts and Telecommunications),* **PRIS at TAC 2009: Experiments in KBP Track**, Proceedings of Text Analysis Conference, 2009, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.
7. *V. Varma, V. Bharat, S. Kovelamudi, P. Bysani, S. GSK, K. Kumar N, K. Reddy, K. Kumar, N. Maganti (IIIT Hyderabad),* **Siel_09: IIIT Hyderabad at TAC 2009**, Proceedings of Text Analysis Conference, 2009, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.
8. *E. Agirre (University of the Basque Country)*, *A.X. Chang, D.S. Jurafsky, C.D. Manning, V.I. Spitkovsky, E. Yeh (Stanford University),* **Stanford_UBC at TAC-KBP**, Proceedings of Text Analysis Conference, 2009, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.
9. *P. Schone, A. Goldschen, C. Langley, S. Lewis, B. Onyshkevych (U.S. Department of Defense)*, *R. Cutts (Henggeler Computer Consultants)*, *B. Dawson, C. Pfeifer, M. Ursiak (MITRE)*, *J. MacBride (BBN)*, *G. Matrangola (SRA International)*, *C. McDonough (Northrop-Grumman)*, **TCAR_r6a: TCAR at TAC-KBP 2009**, Proceedings of Text Analysis Conference, 2009, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.
10. *C. de Pablo-Sánchez, J. Perea, I. Segura-Bedmar, P. Martínez (Charles III University of Madrid),* **uc3m: The uc3m team at KBP task**, Proceedings of Text Analysis Conference, 2009, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.