TAC Entity Linking by Performing Full-document Entity Extraction and Disambiguation

Silviu Cucerzan
Microsoft Research
1 Microsoft Way
Redmond, WA 98052
silviu@microsoft.com

Abstract

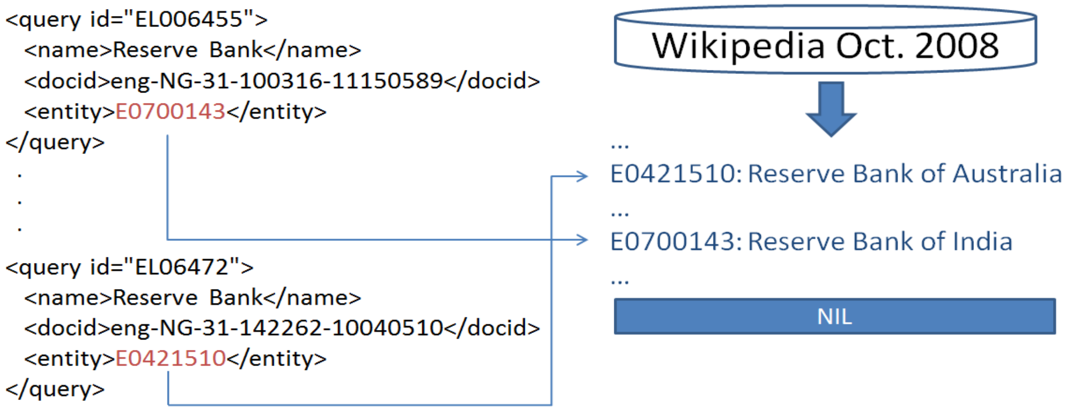
The paper describes the system submitted to TAC 2011 for the English entity linking task of the Knowledge Base Population track. Instead of focusing only on the provided target strings, this system extracts and disambiguates globally all entities from each target document and then maps the target string to one of the entities extracted from the document. The main features employed by the system are topics associated with the entities in the knowledge base, which are derived from Wikipedia categories, list pages, and lexico-syntactic link patterns. The submitted run achieved high scores on the official test data (accuracy of 86.8% and B-cubed+ F-measure of 0.841).

1 Introduction

The TAC entity linking task, which was first introduced in 2009 (McNamee and Dang, 2009) consists of mapping name strings from given text documents to entities from a given knowledge base. It promotes large-scale entity extraction and disambiguation, both in terms of size of the reference entity collection (over 800,000 entities, which were extracted from the Wikipedia dump from October 2008) and size of the target document collection (over 1.3 million Web documents, including blogs and news stories). Not all target strings refer to entities in the given collection, and participating systems must be able to return NIL for such data points. Therefore, overall linking accuracy (A), as well as known-entity linking accuracy (A_{Wiki}) and unknown-entity accuracy (A_{NIL}) are used to measure system performance. Additionally, the 2011 evaluation requires a clustering component of the entities resolved to NIL, in which all instances of each *unknown entity* (i.e., not belonging to the knowledge base derived from the 2008 Wikipedia reference) must be associated with a uniquely numbered NIL id.

Figure 1 shows two query examples from the TAC 2010 data set, in which the target name string “Reserve Bank” must be disambiguated to two distinct Wikipedia entities based on the context of the documents in which it appears. The TAC reference entity set extracted from the October 2008 dump of Wikipedia has 8 entities that contain the string “reserve bank”, which are likely to cover most popular uses of this name string in news and blog data published prior to the date of the employed Wikipedia collection.

The number of entities that contain a target name from the TAC 2010 data set varies between 0 (e.g., “Manhattan Institute” and “Michael Petrelis”) and 3680 (“Ohio”), with an average number of 78.1 and standard deviation of 285.3. While the TAC data do not contain enough labeled instances of each name string to allow us estimate the usage distributions of the corresponding disambiguations in this collection of documents, we expect them to be skewed. The highest number of non-NIL disambiguations for a name string in this data set is 3.



```
<query id="EL006455">
  <name>Reserve Bank</name>
  <docid>eng-NG-31-100316-11150589</docid>
  <entity>E0700143</entity>
</query>
 .
 .
 .
<query id="EL06472">
  <name>Reserve Bank</name>
  <docid>eng-NG-31-142262-10040510</docid>
  <entity>E0421510</entity>
</query>
```

Wikipedia Oct. 2008
…
E0421510: Reserve Bank of Australia
…
E0700143: Reserve Bank of India
…
NIL

Figure 1. Example queries from the TAC 2010 set and corresponding entity entries in the knowledge base as extracted from the October 2008 dump of Wikipedia.

2 System description

The submitted system is an extended version of that described by Cucerzan (2007), which employs both entity information (such as entity type, contexts, and topics) and statistics on *surface forms* (i.e., strings that can be used to refer to entities). Rather than focusing only on the disambiguation of the name string from each input query (like most systems participating in the TAC evaluations from previous years), the proposed system performs a full analysis of each target document, through which it attempts to extract and disambiguate all entities from the document. The output of this analysis process is a list of entities (identified by their canonical Wikipedia name) together with lists of the surface forms extracted from the document and mapped to each of those entities. The system then matches the target name string against the output surface forms as an exact, substring, or superstring match. Further, the entities corresponding to the matched surface forms are ranked based on the type of match and frequency of the surface form. In case no match is found, the document is reprocessed by enforcing this time that the target name string be a candidate surface form, which is achieved by separating the name string from its surrounding context in the target document with a pair of commas and by explicitly adding the name string to the candidate surface form list. This strategy allows the system use its own boundary detection method first to decide the best segmentation of the text into surface forms, including the identification of entities mentioned by substrings and superstrings of the target name string. For the TAC 2010 training set, the target name string does not match exactly any surface form extracted from the text by the system for 7% of the data. For example, the target name string “USC” of one TAC 2010 query gets mapped to the surface form “USC baseball”, which is disambiguated as “USC Trojans baseball”. In another instance, the target name string “Koran Tempo newspaper” gets mapped to the surface string “Koran Tempo” disambiguated as the Wikipedia entity “Koran Tempo”.

The reference collection of entities employed by this system was generated by processing the Wikipedia dump from June 20, 2011 (the latest available at the time the system was trained), which contains approximately 3.7 million entity pages. The mapping of these entities to the official TAC entity id collection was done automatically by using exact matching of Wikipedia page titles (which are employed as the canonical forms of the entities) as well as the redirects extracted from the 2011 collection. The latter heuristic is based on the assumption that in the editorial process of renaming pages, Wikipedia contributors employ the previous name of a page to define a redirect page to the newly renamed page. While this automatic process is prone to mapping errors from which the system cannot recover, employing the more recent 2011 Wikipedia collection presents the important advantage of clustering implicitly those mentions of entities not included in the reference Wikipedia 2008 dump but included in the almost five times larger 2011 version.

As described by Cucerzan (2007), the system employs two components derived from the Wikipedia collection: a set of Wikipedia entities,

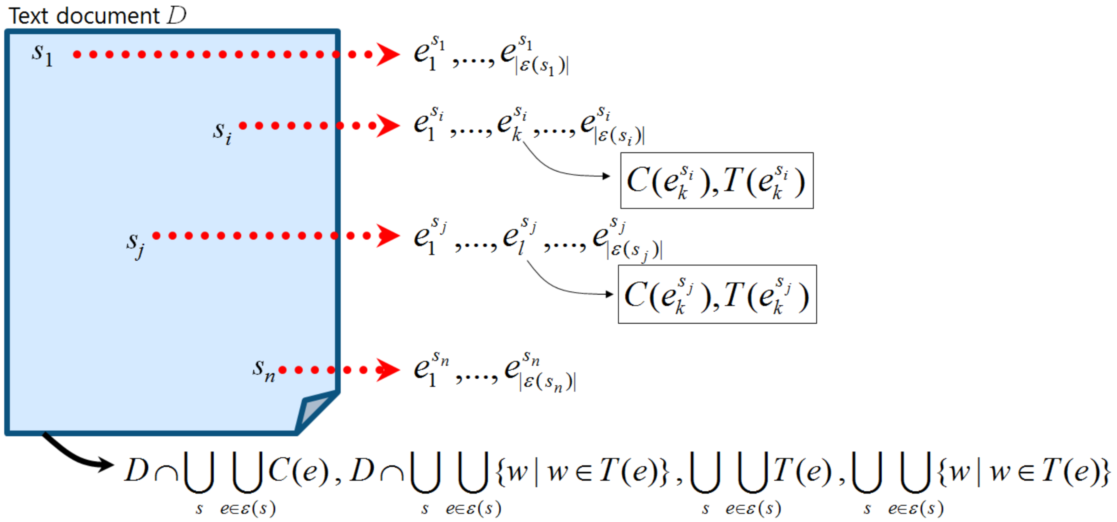

Text document D
s_1
$e_1^{s_1},\dots,e_{|\varepsilon(s_1)|}^{s_1}$
s_i
$e_1^{s_i},\dots,e_k^{s_i},\dots,e_{|\varepsilon(s_i)|}^{s_i}$
$C(e_k^{s_i}),T(e_k^{s_i})$
s_j
$e_1^{s_j},\dots,e_l^{s_j},\dots,e_{|\varepsilon(s_j)|}^{s_j}$
$C(e_k^{s_j}),T(e_k^{s_j})$
s_n
$e_1^{s_n},\dots,e_{|\varepsilon(s_n)|}^{s_n}$
$D\cap\bigcup_s\bigcup_{e\in\varepsilon(s)}C(e),\ D\cap\bigcup_s\bigcup_{e\in\varepsilon(s)}\{w\,|\,w\in T(e)\},\ \bigcup_s\bigcup_{e\in\varepsilon(s)}T(e),\ \bigcup_s\bigcup_{e\in\varepsilon(s)}\{w\,|\,w\in T(e)\}$

Figure 2. Document analysis: the system attempts to find an assignment of entities to the surface forms extracted from a target document *D* that maximizes the similarity between the document representations in the context and topic spaces and the known contexts and topics for each candidate entity from that assignment.

together with contexts and topics that they belong to, and a set of surface form to entity mappings. However, there are several important differences in how these components are populated and used. First, the topics are extracted not only from Wikipedia categories (approximately 456k topics) and list pages (80k topics) but also based on lexico-syntactic pattern matches (more than 852k topics), which are named by using the title of the Wikipedia page in which they were identified in concatenation with a numeric counter. The topics of this new type are by far the most numerous in the proposed system. Figure 3 shows a histogram for the number of entities that have assigned to them various numbers of topics. The average number of topics associated with an entity is 4.5 and the average number of entities that belong to a topic is 12.0.

Additionally, the new system makes use not only of vectors of topic ids but also of the vocabulary of the topic names. The latter is useful for computing both an additional measure of contextual similarity with the target document and an additional measure of lexical similarity between topics, which overcomes the problems of topic sparsity and topic redundancy (which also leads to topic id mismatching). These problems can be noticed in Figure 5, which shows the Wikipedia categories for the correct disambiguations of the target name string (surface form) "AZ" and three other surface forms from a target document in the TAC 2010 data set (eng-WL-11-174574-12934438, shown in Figure 4). Note that there is only one category shared by all four entities ("Living people") despite the assignment of very similar categories to these entities, such as the pair "African American actors" and "African American film actors" or the pair "Hispanic and Latino American rappers" and "African American Rappers".

The extensive use of topics associated with entities made possible a more conservative approach for extracting contexts, basically eliminating the need of using the bidirectional Wikipedia linkage employed by the previously published system.

The disambiguation process is similar to a large extent to that presented by Cucerzan (2007). Figure 2 sketches this disambiguation paradigm, in which the system attempts to find an assignment of entities to the set of surface forms extracted from the target document *D* that maximizes the similarity between the contexts and topics of each entity in the assignment and a particular representation of document in the context and topic spaces. These document representations employ both the text of

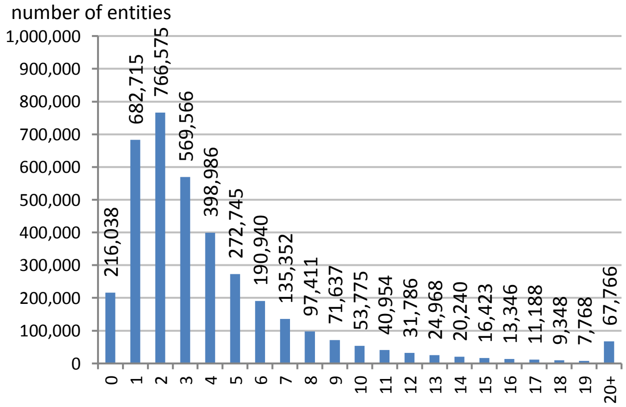
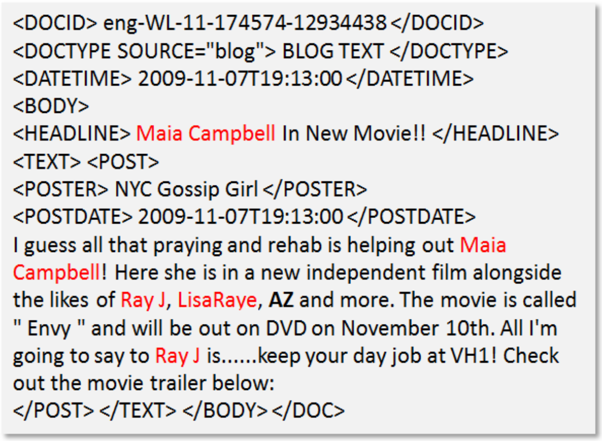


Figure 3. Number of Wikipedia entities in the system that are assigned a certain number of topics (0 to 20+).



```
<DOCID> eng-WL-11-174574-12934438 </DOCID>
<DOCTYPE SOURCE="blog"> BLOG TEXT </DOCTYPE>
<DATETIME> 2009-11-07T19:13:00 </DATETIME>
<BODY>
<HEADLINE> Maia Campbell In New Movie!! </HEADLINE>
<TEXT> <POST>
<POSTER> NYC Gossip Girl </POSTER>
<POSTDATE> 2009-11-07T19:13:00 </POSTDATE>
I guess all that praying and rehab is helping out Maia
Campbell! Here she is in a new independent film alongside
the likes of Ray J, LisaRaye, AZ and more. The movie is called
" Envy " and will be out on DVD on November 10th. All I'm
going to say to Ray J is......keep your day job at VH1! Check
out the movie trailer below:
</POST> </TEXT> </BODY> </DOC>
```

Figure 4. Example target document for the name string "AZ" from the TAC 2010 data set.

the document and the contexts and topics of all possible entity disambiguations of the set of surface forms extracted from the document. In total, four document representations are used, of which one is in the entity-context space (the projection of the document text onto the space of all known contexts for all possible entity disambiguations for the surface form set) and three in the topic space (one that uses the topic ids and two that use the topic vocabulary of all possible disambiguations).

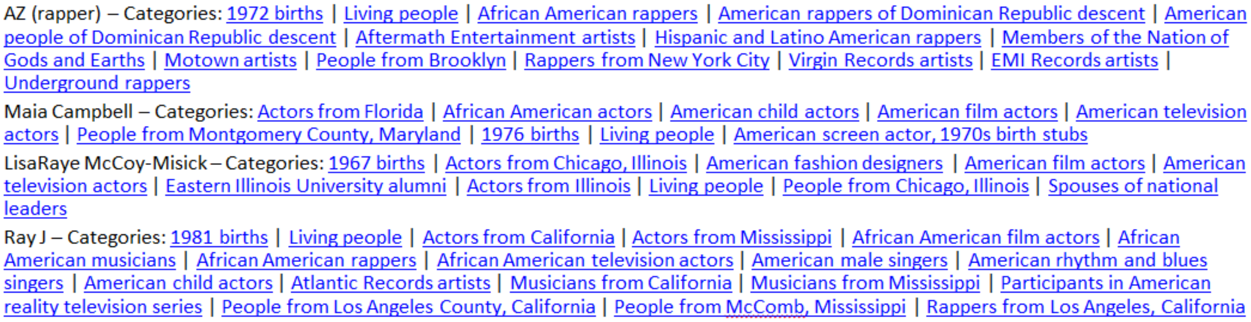
Contextual similarities between each candidate disambiguation and the target document are then computed as scalar products between the known context and topic vocabulary vectors for the candidate and the two corresponding document projections. Similarly, topic similarities are computed between individual topic and topic vocabulary vectors of the candidate disambiguations and the two corresponding document representation in the topic and respectively, topic vocabulary spaces.

The entity assignment is finally calculated as the argmax of a linear combination of these similarities and several additional features:

$$\underset{e_i \in \varepsilon(s_i)}{\arg\max} \sum_{j=1..|F|} \beta_j \cdot f_j(e_i, \overline{D}), i \in 1..n$$

where s_i denotes one of the surface forms from document D to be disambiguated, $\varepsilon(s_i)$ denotes the set of all known disambiguations for that surface form, and F comprises the following ten features:

- Wikipedia-based prior (computed based on the number of times surface forms are used as anchor texts for Wikipedia interlinks);
- Context similarity of the target document with the candidate entity's contextual vector;
- Lexical similarity between the target document and the candidate's topic vocabulary;
- Topic-identifier similarity between the aggregated topic id model for the document and the candidate entity's topic id vector;
- Topic vocabulary similarity between the document representation in the topic vocabulary space and each topic vocabulary vector;
- Number of different surface forms in the target document that lead to the same candidate entity;
- A binary flag encoding whether a required context is found in the document (such as the context "India" for the surface form "Ministry of Education" and the candidate disambiguation "Ministry of Education (India)";
- String similarity between the surface form and the candidate entity's canonical form;
- Acronym matching flag, which indicates whether the surface form is an acronym of the



AZ (rapper) – Categories: 1972 births | Living people | African American rappers | American rappers of Dominican Republic descent | American people of Dominican Republic descent | Aftermath Entertainment artists | Hispanic and Latino American rappers | Members of the Nation of Gods and Earths | Motown artists | People from Brooklyn | Rappers from New York City | Virgin Records artists | EMI Records artists | Underground rappers

Maia Campbell – Categories: Actors from Florida | African American actors | American child actors | American film actors | American television actors | People from Montgomery County, Maryland | 1976 births | Living people | American screen actor, 1970s birth stubs

LisaRaye McCoy-Misick – Categories: 1967 births | Actors from Chicago, Illinois | American fashion designers | American film actors | American television actors | Eastern Illinois University alumni | Actors from Illinois | Living people | People from Chicago, Illinois | Spouses of national leaders

Ray J – Categories: 1981 births | Living people | Actors from California | Actors from Mississippi | African American film actors | African American musicians | African American rappers | African American television actors | American male singers | American rhythm and blues singers | American child actors | Atlantic Records artists | Musicians from California | Musicians from Mississippi | Participants in American reality television series | People from Los Angeles County, California | People from McComb, Mississippi | Rappers from Los Angeles, California

Figure 5. The categories associated with the entities "AZ (rapper)", "Maia Campbell", "LisaRaye McCoy-Misick", and "Ray J" in the June 2011 Wikipedia dump. While the vocabularies of the category sets associated with each entity overlaps with each other to a high degree, there is only one category shared by all four entities ("Living people"), and no other categories are shared by the categories associated with the entity "AZ (rapper)" and the other three entities.

candidate disambiguation and whether the latter is present in the text of the document;

- Binary flag for unlikely types of entities (such as works of art and artifacts).

The final weights of the linear combination (β_j, $j = 1..|F|$) were trained on the TAC 2010 data, by using as objective function the system's accuracy.

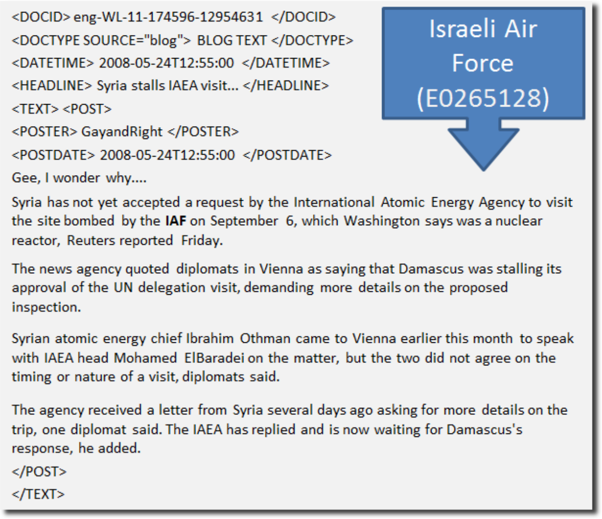
3 Subtasks

This section describes several system components particularly useful for the TAC evaluation.

3.1 TAC Name Strings versus Surface Forms

Target name strings employed in TAC evaluations do not necessarily match the exact surface forms extracted from text by an entity recognition system (in particular, the employed system). For example, the string "`Dick`" is used as a target name in the TAC 2010 for references to "Andy Dick", "Dick Cheney", "Kirby Dick", and "Dick Ebersol". In some instances (e.g., the target document `eng-WL-11-174595-12967356`), a full known surface form for one of these entities is present in the target document, but in some other cases (e.g., `eng-WL-11-174643-13000483`), such a surface form is not present. The former case can be seen as requiring an additional coreference task. The later suggests that the number of candidate disambiguations to be considered for a target name string is much larger than the number of entities for which that name string was extracted as an exact surface form from the Wikipedia collection. To account for such cases, all Wikipedia surface forms that contain a target name as a substring must become disambiguation candidates for that name. For the example "`Reserve Bank`" in Figure 1, while there are only 9 entities for which "reserve bank" is a known surface form in the June 2011 Wikipedia dump, there are no less than 105 surface forms that contain it as a substring in this data collection. Overall, these surface forms are associated with 68 distinct entities, and thus, the apparent 10-way classification problem (9 entities and NIL) becomes instead a 69-way classification (68 entities and NIL).

Instead of focusing on all possible disambiguations for a target string, the proposed system addresses these issues by performing full-document entity extraction and disambiguation, which includes coreference and partial name resolution, followed by a stage in which the target string is mapped to one of the extracted entities. The coreference and partial name resolution employ relatively simple positional and string matching heuristics to map shorter surface forms to longer surface forms labeled with the same entity type, as well as a set of about 4,000 name map-


Israeli Air Force (E0265128)

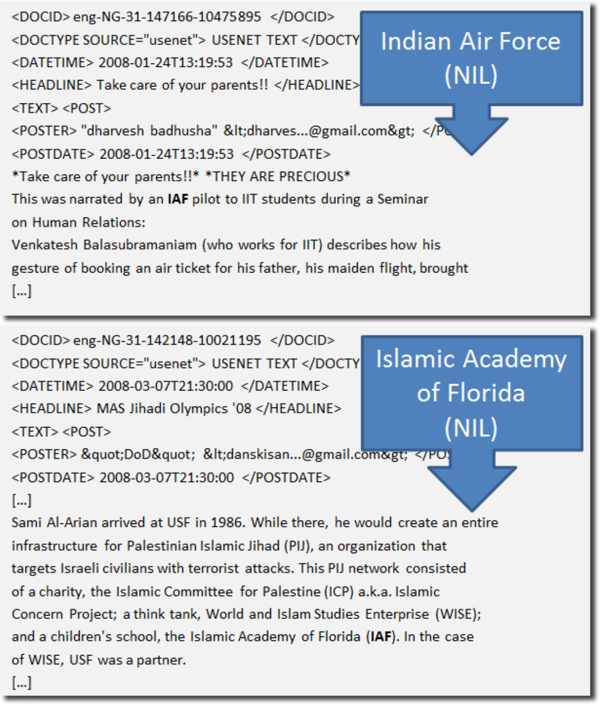
```
<DOCID> eng-WL-11-174596-12954631 </DOCID>
<DOCTYPE SOURCE="blog"> BLOG TEXT </DOCTYPE>
<DATETIME> 2008-05-24T12:55:00 </DATETIME>
<HEADLINE> Syria stalls IAEA visit... </HEADLINE>
<TEXT> <POST>
<POSTER> GayandRight </POSTER>
<POSTDATE> 2008-05-24T12:55:00 </POSTDATE>
Gee, I wonder why....
Syria has not yet accepted a request by the International Atomic Energy Agency to visit
the site bombed by the IAF on September 6, which Washington says was a nuclear
reactor, Reuters reported Friday.

The news agency quoted diplomats in Vienna as saying that Damascus was stalling its
approval of the UN delegation visit, demanding more details on the proposed
inspection.

Syrian atomic energy chief Ibrahim Othman came to Vienna earlier this month to speak
with IAEA head Mohamed ElBaradei on the matter, but the two did not agree on the
timing or nature of a visit, diplomats said.

The agency received a letter from Syria several days ago asking for more details on the
trip, one diplomat said. The IAEA has replied and is now waiting for Damascus's
response, he added.
</POST>
</TEXT>
```

Figure 6. Example of acronym ("`IAF`") used as a target name in the TAC 2010 data set.


Indian Air Force (NIL)
Islamic Academy of Florida (NIL)

```
<DOCID> eng-NG-31-147166-10475895 </DOCID>
<DOCTYPE SOURCE="usenet"> USENET TEXT </DOCTY
<DATETIME> 2008-01-24T13:19:53 </DATETIME>
<HEADLINE> Take care of your parents!! </HEADLINE>
<TEXT> <POST>
<POSTER> "dharvesh badhusha" &lt;dharves...@gmail.com&gt; </P
<POSTDATE> 2008-01-24T13:19:53 </POSTDATE>
*Take care of your parents!!* *THEY ARE PRECIOUS*
This was narrated by an IAF pilot to IIT students during a Seminar
on Human Relations:
Venkatesh Balasubramaniam (who works for IIT) describes how his
gesture of booking an air ticket for his father, his maiden flight, brought
[...]
```

```
<DOCID> eng-NG-31-142148-10021195 </DOCID>
<DOCTYPE SOURCE="usenet"> USENET TEXT </DOCTY
<DATETIME> 2008-03-07T21:30:00 </DATETIME>
<HEADLINE> MAS Jihadi Olympics '08 </HEADLINE>
<TEXT> <POST>
<POSTER> &quot;DoD&quot; &lt;danskisan...@gmail.com&gt; </PO
<POSTDATE> 2008-03-07T21:30:00 </POSTDATE>
[...]
Sami Al-Arian arrived at USF in 1986. While there, he would create an entire
infrastructure for Palestinian Islamic Jihad (PIJ), an organization that
targets Israeli civilians with terrorist attacks. This PIJ network consisted
of a charity, the Islamic Committee for Palestine (ICP) a.k.a. Islamic
Concern Project; a think tank, World and Islam Studies Enterprise (WISE);
and a children's school, the Islamic Academy of Florida (IAF). In the case
of WISE, USF was a partner.
[...]
```

Figure 7. Two more examples of usage of the acronym "`IAF`" as a target name in the TAC 2010 data set.

pings (such as Bill → William, Alex → Alessandro, and Ahmed → Ahmad) automatically derived from the Wikipedia collection. For partial personal names that cannot be resolved to a known superstring surface form in text, the system uses the complementary parts of the entities of the type person identified in the document to create disambiguation candidates.

3.2 Acronyms

More than 10% of the target names in the TAC 2010 data are acronyms. Figures 6 and 7 show three examples of target texts for the target name "IAF". In one case, the correct disambiguation is an entity in the 2008-based knowledge base, in a second case, the correct disambiguation does not appear in the 2008 collection but is included in the 2011 Wikipedia collection, and in the third case, the correct disambiguation does not appear in any of the Wikipedia versions employed.

For the first two examples, the correct expansion does not appear in the text of the target document and the system must rely only on the surface form to entity mappings extracted from Wikipedia for "IAF" (as shown in Table 1) and the general disambiguation paradigm described in Section 2. Employing the much larger 2011 collection is beneficial for the second example because the correct entity becomes a candidate disambiguation despite the fact that its expansion does not appear as a surface form in the target document.

The system also employs an acronym detector and matcher similar to that described by Jain et al. (2007), which is able to identify expansions of a target acronym in the text irrespective of their presence as surface forms in the Wikipedia-based knowledge base (such as the third example shown for "IAF", with the disambiguation "Islamic Academy of Florida"). Whenever at least one possible expansion in the text is present in the Jain et al. acronym-expansion list (the entries for "IAF" in this list are shown in Table 2), the system restricts the candidate space to only acronym expansions from the list that occur in the target document.

3.3 Capitalization

The proposed system uses a system similar to that described by Cucerzan (2010) to trucase the beginning of the sentences and the text lines that appear to be titles (i.e., all words but function words start with uppercase letters). For this particular evaluation, the employed capitalization n-gram statistics were derived from Wikipedia and the Gigabyte corpus.

4 Clustering of Unknown Entities

To address the labeling of NIL entities, the system used to generate the official run submitted for evaluation to TAC 2011 relied only on the much larger size of the 2011 Wikipedia dump and exact string matching for entities that could not be disambiguated to a Wikipedia entity. No additional explicit clustering algorithm was employed for this official run.

IAF
IAF (disambiguation)
Israeli Air Force
Indian Air Force
Indonesian Air Force
International Accreditation Forum
International Astronautical Federation
Islamic Action Front

Table 1. Disambiguations for the surface form "IAF", as extracted from the June 2011 Wikipedia dump.

IAF
israeli air force
international association of facilitators
institute for alternative futures
industrial areas foundation
international accreditation forum
inter american foundation
israel air force
integrated architecture framework
intelligent audio file
iraqi accordance front
inspired art fair
international astronautical federation
international advertising festival
indian air force
islamic action front
infrastructure assessment framework
industrial air filtration
international apparel federation
islamic academy of florida
integration adapter framework
international of anarchist federations
inject a floor
international academy of flint

Table 2. Ranked list of normalized expansions extracted for "IAF" from Web search logs and Web data.

5 Development and Evaluation

The system submitted for evaluation to TAC 2011 was developed within one and a half person months starting from a production system based on the work of Cucerzan (2007). The processing is done in real-time, with an average speed of over 10 news/blog documents per second. When retrained on the June 2011 Wikipedia dump, the out-of-the-box accuracy of the existing system on the TAC 2010 training data was 86.3%. Once redirection information was employed to map entities from the 2011 dump to those from the TAC reference collection, the performance increased to 88.2%. The additional increase to the final numbers shown in Table 3 were due to employing a reprocessing step for cases in which the target name string could not be aligned to one of the extracted surface forms, better ranking of these aligned forms when multiple alignments were possible, mapping of Wikipedia disambiguation pages to NIL, as well as the addition of two new features for acronym matching and for types of entities unlikely to be appear in the target data. The final results obtained (A=90%), compare favorably to those reported by the best two participants in the TAC 2010 evaluation, A = 86.8% (Lehmann et al., 2010), and respectively, A= 84.4% (Radfordyz et al., 2010).

While in the TAC 2010 evaluation, the NIL class encompassed all entities that are not present in the reference collection, for the TAC 2011 evaluation, systems were required to cluster the NIL values, so that each unknown entity gets assigned one unique identifier. This requires new evaluation metrics, which measure the overlap between the gold standard clusters and those hypothesized by participating systems. The metrics employed, derived from those proposed by Bagga and Baldwin (1998), are B-cubed+ precision and B-cubed+ recall with equal element weighting:

$$P_B^3 = \mathrm{Avg}_x(\mathrm{Avg}_{x'|\mathrm{T}(x)=\mathrm{T}(x')}(\delta\,(\mathrm{T}(x),\mathrm{S}(x),\mathrm{S}(x'))),$$

$$R_B^3 = \mathrm{Avg}_x(\mathrm{Avg}_{x'|\mathrm{S}(x)=\mathrm{S}(x')}\,(\delta\,(\mathrm{S}(x),\mathrm{T}(x),\mathrm{T}(x'))),$$

and the corresponding F-measure:

$$F_{B^3} = 2 \,/ \left(\frac{1}{P_{B^3}} + \frac{1}{R_{B^3}} \right),$$

where T(x) denotes the true label of an instance x and S(x) denotes the label predicted by the evaluated system for x.

TAC 2010	A	A_{Wiki}	A_{NIL}
Training	89.9%	90.6%	88.3%
Test	90.0%	87.3%	92.2%

Table 3. Scores obtained by the proposed system on the TAC 2010 data.

TAC 2011	A	P_B^3	R_B^3	F_B^3
Official Scores	**86.8%**	**0.848**	0.834	0.841

Table 4. Official scores of the submitted system. Bold indicates the score is the best obtained in TAC 2011.

The scores obtained on the run submitted for the system described in this paper, are shown in Table 4. The system achieved the best accuracy and the second best B-cubed+ F-measure (the maximum achieved was 0.846, the median was 0.716).

6 Conclusion

The paper described an entity linking system that performs full entity analysis of a target document by modeling it in the space of Wikipedia-derived topics and contexts of all candidate entity disambiguations for all surface forms extracted from that document. The empirical results obtained on the test set suggest that this system is achieving current state-of-the-art entity linking performance.

References

A. Bagga and B. Baldwin. 1998. Algorithms for Scoring Coreference Chains. Proceedings LREC 1998 Workshop on Linguistic Coreference, pages 563–566.

S. Cucerzan. 2007. Large Scale Named Entity Disambiguation Based on Wikipedia Data. The EMNLP-CoNLL Joint Conference 2007, pages 708—716.

S. Cucerzan. 2010. A Case Study of Using Web Search Statistics: Case Restoration. Lecture Notes in Computer Science, Vol. 6008, Springer, pages 199—211.

A. Jain, S. Cucerzan, and S. Azzam. 2007. Acronym-Expansion Recognition and Ranking on the Web. IEEE-IRI 2007, pages 209—214.

J. Lehmann, S. Monahan, L. Nezda, A. Jung, and Y. Shi. 2010. LCC Approaches to Knowledge Base Population at TAC 2010. Text Analysis Conference, http://www.nist.gov/tac/publications/2010/papers.html

P. McNamee and H.T. Dang. 2009. Overview of the TAC 2009 Knowledge Base Population Track. TAC. http://www.nist.gov/tac/publications/2009/papers.html

W. Radfordyz, B. Hacheyz, J. Nothmanyz, M. Honnibalyz, and J.R. Curran. 2010. Document-level entity linking: CMCRC at TAC 2010. TAC. http://www.nist.gov/tac/publications/2010/papers.html