The MSR System for Entity Linking at TAC 2012

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

Abstract

The paper describes the system submitted to TAC 2012 for the English entity linking task of the Knowledge Base Population track. It focuses on the components that are novel with reference to the MSR system that participated in TAC 2011. Three runs were submitted for evaluation, and the best of them achieved the highest B-cubed+ F score of all systems participating in this evaluation (0.730) and the highest disambiguation accuracy (0.766).

1 Introduction

The TAC entity linking task, which was first introduced in 2009 (McNamee and Dang, 2009), consists of mapping name strings from text documents harvested from newswire, blogs, and other Web sources to entities from a knowledge base with over 800,000 entries, which was derived from the Wikipedia dump from October 2008. A large percentage of the target name strings in the training and test data account for entities that are not present in the given reference collection and should be mapped to NIL. Because of this, the current task definition includes the requirement of grouping all references of each entity across the documents in the test collection, whether or not the entity is in the reference collection, which amounts to performing inter-document coreference resolution.

The runs submitted by Microsoft Research (MSR) to the 2012 evaluation were based on variants of the MSR system that participated in the previous year's TAC evaluation (Cucerzan, 2011). Two of these runs used as reference the Wikipedia dump from February 11, 2012, while the third employed the Wikipedia dump from June 1, 2012. One of the runs that employed the February dump used a code base almost identical to that from TAC 2011. In all cases, the entities from the 2012 Wikipedia dumps were linked to the TAC reference collection (derived from a 2008 dump) by mapping to the TAC reference entities those entities in the newer collections with identical page titles (seen as canonical names) and those with pages referred by Wikipedia redirect pages with identical titles.

The general architecture of the system participating in TAC 2011 has been preserved. The most notable changes refer to entity boundary detection, the introduction of geo-spatial features, and improvements in the processing of Wikipedia dumps.

While the organizers have provided offsets for the target names in the TAC 2012 data (and the TAC 2011 training data) to avoid problems with multiple senses per document, the MSR system that generated the submitted runs did not employ this information, in part because the provided offsets correspond to the position of target names rather than the actual *surface forms*, which represent the textual mentions of the entities targeted for disambiguation (an example is shown in Figure 1).

> This so-called "Tom **Bradley** Effect" had not yet come into play in Iowa because Obama was not yet taken as a serious threat to win the nomination and because the caucus process was so intimate and open. His dramatic victory in that state means that no one can any longer doubt that he has a real chance to win.

Figure 1. Snippet of text from a target document in the TAC 2011 test set for the target name `Bradley`. Note that the target name is part of two different entities, "Tom Bradley" and "Bradley Effect", both of which are present in the Wikipedia collection. Thus, coming up with a disambiguation that matches the gold standard employed requires the knowledge of exact boundaries in text of the surface form targeted for disambiguation.

Figure 2. Screenshot of a TAC 2011 target document analyzed by the system, with superimposed disambiguations for several of the extracted surface forms (as displayed by the MSR system on hovering over those surface forms).

2 System Description

The submitted system is an extended version of the system described in Cucerzan (2011), which employs both entity representations in context/topic spaces and statistical mappings of *surface forms* (strings used for mentioning entities in text) to entities, as extracted from the Wikipedia collection.

The system takes as input a text document and 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 that are mapped by the system to each of those entities. To perform the TAC entity linking task, we match the target name string from a TAC query against the output surface forms extracted from the target document. In the matching process, the target name can be identical to one or more of the surface forms extracted from text, can be a substring of one or more of the extracted surface forms, or a superstring of one or more of those forms. The entities corresponding to all matched surface forms are ranked based on the type of match and frequency of the surface form and the top-ranked entity is returned as the answer. 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. When no such match is found, the target document is processed a second time while enforcing that the target name string is a candidate surface form to be disambiguated.

When his father died, the caliph made of him his principal counselor, his Grand Vizier. Thus it was through Saint John Damascene that the advanced sciences made their apparition among the Arab Moslems, who had burnt the library of **Alexandria** in Egypt; it was not the Moslems who instructed the Christians, as was believed for some time in Europe.

After the attacks, sales of **Bordeaux** wine to the United States fell by 29 percent in volume during the final quarter of 2001 -- the key Thanksgiving, Christmas and New Year period, which accounts for half of annual sales.

Figure 3. Snippets of text from two target documents in the TAC 2011 test set for the target names `Alexandria` and `Bordeaux`. Note that the names can be disambiguated either in the given form or as part of longer surface forms in text (`library of Alexandria` and `Bordeaux wine`, respectively), and that the disambiguations depend on the boundaries selected.

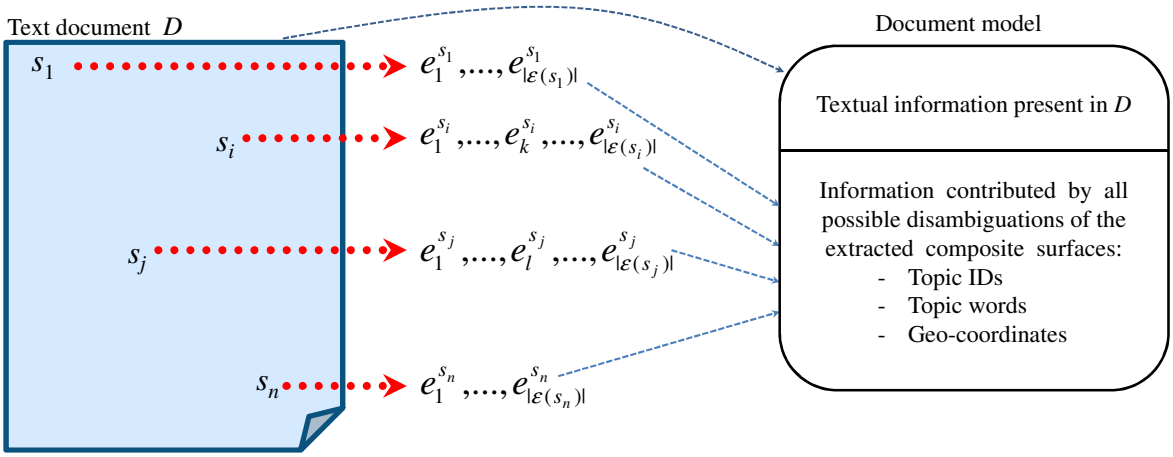

Figure 4. Given a target document, the system first extracts surface forms (either simple or composite). The system then builds a document model that includes the original text of the document, as well as vectorial models that aggregate various types of information associated with all possible entity disambiguations for all extracted surface forms.

While this strategy works well overall, a number of the errors observed in the TAC 2011 evaluation are due to the fact that the system selects a wrong surface form to disambiguate when multiple entity boundaries are possible. The heuristic employed, of using the longer surface forms when multiple boundaries are possible, gives better results than heuristically choosing the shorter forms. However, the gold standard does not follow such a rule, and includes both target names with entity resolutions based on the longer surface forms in text and target names with entity resolutions based on the exact target string or shorter surface forms containing it. Two examples of the latter are shown in Figure 3. Figure 1 shows also an example in which the disambiguation of the target name depends on which longer surface form is chosen for disambiguation.

2.1 General Architecture

The analysis of a text is done in three stages, which are similar to those described by Cucerzan (2007):

- text normalization and sentence breaking;
- surface form boundary detection;
- document model construction and entity disambiguation.

However, by contrast to Cucerzan (2007), the system employed in the TAC evaluation uses only resources derived from Wikipedia in all three stages, and does not perform queries against Web data. For example, the boundary detection does not employ Web statistics for solving structural ambiguities, such as conjunctive constructions and prepositional attachments. To address these ambiguities as well as to attempt selecting the surface forms that correspond to the most appropriate entities in the context of a given document (see Figure 3), the MSR system for TAC employs a strategy in which multiple possible boundaries are used to create composite surface forms; the possible disambiguations of all participating surface forms are merged, and the decision about the entity mention boundaries is postponed until the third stage. (More details on this process are provided in Section 2.2.1.) Additionally, the second stage performs only surface form detection and does not label the detected mentions with entity classes.

The document model construction and entity disambiguation stage follow the ideas presented in Cucerzan (2011) for the TAC 2011 evaluation: Multiple features are computed for each possible disambiguation associated with a surface form. Some of these are based on similarities between the information extracted from Wikipedia for the candidate entities (such as contexts) and the target document, while others measure the similarity of the information associated with the candidate entities for each surface form and the document representation obtained by aggregating the information from all possible disambiguations (candidate entities) for all surface forms in the given document. Figure 4 sketches the construction of the document model used to compute these features.

Finally, as with the MSR system that participated in the TAC 2011 evaluation, the entity assignment is calculated for each surface form as the

argmax of a linear combination the features computed for all candidate entity disambiguations.

2.2 Novel Components

This section provides details on the main differences between the system that participated in the TAC 2011 evaluation and the current system. Most of these components were implemented without specifically targeting the TAC evaluation and their individual contribution to the reported performance on the TAC data sets is not very large.

2.2.1 Composite Surface Forms

In many situations, it is difficult to hypothesize accurately the boundaries of surface forms in text without employing precise syntactic chunks and additional semantic representations. Therefore, it can be beneficial to postpone making a decision on the boundaries of some surface forms until the disambiguation stage, in which information about the other entity candidates in the document becomes available through the document model.

In those cases in which the boundary identification decision is postponed, the information stored in the knowledge base for the surface forms corresponding to all possible boundaries at a given location in text is merged to obtain a composite surface form record, which is employed further in the document model building and in the disambiguation component in the same manner as data for regular surface forms are employed. This strategy ensures that each entity to be extracted from the document contributes the same amount of information to the document model and to the overall disambiguation process regardless of how many different possibilities exist for identifying its surface form boundaries at the particular location of that entity in text. As with the examples shown in Figure 3, note that the resulted entities (extracted by the system from the document) often depend of the selected surface form boundaries and vice versa.

2.2.2 Geo-coordinates

In addition to the features described in Cucerzan (2011), the MSR system participating in TAC 2012 employs two novel features, which are based on geo-coordinates associated with entities in Wikipedia, as mined from the Wikipedia dumps. Similarly to the way topics associated with all candidate disambiguations for all surface forms get aggregated in the document model, geo-coordinates are contributed to the document model by all candidate disambiguations that have associated such information. We compute as features for each candidate disambiguation of a surface form that has geo-coordinates the minimum geo-distance to all other locations/geo-coordinates in the document model, as well as the average with respect to all other surface forms in the document of the minimum distances to any of those surface forms' possible disambiguations.

2.2.3 Concepts versus Entities

On one hand, Wikipedia contains numerous pages for works of art with titles that are typically used as common noun phrases or for other common meanings. For example, "yesterday", "every morning", "this is the day", and "let's go" can all be song titles, for which dedicated Wikipedia pages exist. However, there are no pages for the common meanings of these terms. Attempting to disambiguate an occurrence of such a term in a document could result in most cases in the addition spurious information to the document model. On the other hand, Wikipedia contains also numerous pages for common concepts/noun phrases, such as "economics", "board of directors", and "shareholder", which could contribute to a document model information potentially valuable for the disambiguation stage. To distinguish between the two, binary labels are hypothesized for all Wikipedia pages to indicate whether they describe an entity (which is typically a proper noun phrase) or a common concept. This is done by employing a logistic regression classifier trained on 1,000 manually labeled Wikipedia pages. The classifier achieved 99% accuracy in cross-validation experiments on this labeled set. The binary labels are stored in the knowledge base together with the other entity information (topics, contexts, and geo-coordinates).

When a document is analyzed, surface forms with lowercase spellings are extracted and retained for the disambiguation stage only if they have at least one possible disambiguation that is a common concept (i.e., Wikipedia page labeled as describing a common concept), and that disambiguation has associated topics that also belong to other entities that are candidate disambiguations for the other surface forms present in text. Implemented as a two-step process, this heuristic rule results in the elimination of most lowercase surface forms with

spurious disambiguations, while it still allows for the use of common concepts that are important for the meaning/topic of the target document.

2.3 Other Changes

For one of the runs submitted (denoted R3), we also employed a new code in the off-line process that extracts the knowledge base of the MSR system from a Wikipedia dump. This new code performs an improved analysis of Wikipedia and handles various changes that have been made over time in the format of the dumps. The new Wikipedia analyzer based on this code also makes use of additional information available more consistently in the more recent dumps, such as "See also" and "Main" templates, from which new types of topics are built. Additionally, this new Wikipedia analyzer generates extra surface forms to handle more robustly spacing and punctuation variations, short forms of person names, etc. A software bug that was identified in the code of the new analyzer after the TAC evaluation resulted in the erroneous extraction (and further usage) of geo-coordinate values employed for generating the run R3.

3 Clustering of Unknown Entities

Similarly to TAC 2011, the systems participating in the 2012 evaluation are required to cluster the NIL values across target documents, so that all instances of each *unknown entity* (i.e., that is not present in the given knowledge base) get assigned the same unique identifier.

The MSR system does not employ a sophisticated clustering component for NIL-mapped target names. Its NIL-tag labeling relies on the following:

- the much larger size of the 2012 Wikipedia dumps (e.g., the target name `Appleton` gets disambiguated by the system to "Appleton, Wisconsin" in two documents and to "Appleton, New York" in two other documents from the TAC 2011 evaluation set, depending on the context of those documents; while both entities appear in the 2012 dumps, only the former is listed in the 2008 reference entity list; however both instances of the latter get assigned the same NIL label despite the fact that the surface forms extracted from text are different from each other: `Appleton, New York` in one instance, and `Appleton, N.Y.` in the other instance);

F_B^3	Submitted MSR runs			All participants	
	R1	R2	R3	max	median
All	0.721	0.694	**0.730**	0.730	0.536
in KB	**0.687**	0.641	0.685	0.687	0.496
NIL ($\notin$ KB)	0.758	0.754	0.781	0.847	0.594
NW docs	0.775	0.742	**0.782**	0.782	0.574
WB docs	0.615	0.601	0.630	0.646	0.492
PER	0.788	0.790	0.809	0.840	0.646
ORG	0.655	0.649	0.715	0.717	0.486
GPE	**0.694**	0.601	0.627	0.694	0.447

Table 1. Official F_B^3 scores for the submitted runs. Bold indicates the score is the best obtained in TAC 2012.

Accuracy	Submitted MSR runs			All participants	
	R1	R2	R3	max	median
All	0.762	0.739	**0.766**	0.766	0.601
in KB	**0.720**	0.676	0.712	0.720	0.526

Table 2. Official accuracy results for the runs of the submitted system. Bold indicates the score is the best obtained in TAC 2012.

- acronym expansion matching in the text (e.g., `CASA` gets expanded to "Civil Aviation Safety Authority" in several different documents in the TAC 2011 evaluation set, and thus, it gets mapped to the same NIL identifier; ADF gets mapped to `Alliance Defense Fund` in one document, `American Dance Festival` in another document, and `Australian Defence Force` in a third document, and are assigned identifiers different from each other);
- the identity of surface forms extracted in different target documents by the boundary detection and in-document coreference components (e.g., the target name `Harpootlian` gets mapped to the surface form `Dick Harpootlian` in several documents in the TAC 2011 test set; while there exists no Wikipedia page for this person entity even in the newer collections/dumps, all instances in the TAC 2011 set get clustered together because of the identical surface form to which the target name is mapped).

4 Evaluation

Three runs, ordered by their performance on the TAC 2011 data, were submitted for evaluation in TAC 2012. Run R2 was generated by a version of the system that is the most similar to the MSR system that participated in the TAC 2011 evaluation.

It employs the additional components discussed in Sections 2.2.1 and 2.2.3. Run R1 additionally employs the geo-coordinate features discussed in Section 2.2.2 together with some minor variations in how surface boundaries are handled. Both systems that generated R1 and R2 use a knowledge base derived from the Wikipedia dump from February 11, 2012 by employing the older Wikipedia analyzer code. The third run (R3) was generated by a system that employs all of the novel components presented, as well as a knowledge base generated by the new Wikipedia analyzer (as discussed in Section 2.3) from the June 1, 2012 dump.

The parameters of these systems were tuned by employing annotated data from the TAC 2009, 2010, and 2011 evaluations.

Because of the requirement to cluster NIL values across target documents, the performance of the participating systems is measured using B-cubed+ clustering metrics (Bagga and Baldwin, 1998), which account for the overlap between the gold standard clusters and those hypothesized by the systems. Table 1 shows the official F_B^3 scores for runs R1, R2, and R3, as well as the maximum and median scores for all participating systems. Note that the official median scores reported were obtained by considering only the highest score of all runs submitted by each of the 25 participating teams (rather than taking the median of all 98 runs submitted for evaluation).

By collapsing all NIL labels into one class, linking accuracy can also be measured on the TAC test set. Table 2 shows the accuracy numbers obtained by the three MSR runs, as well as the maximum and median accuracy with respect to all systems submitted to TAC 2012.

5 Conclusion

The paper described an entity linking system that performs full-document entity extraction and linking to Wikipedia, which obtained very good empirical results in the TAC 2012 evaluation.

References

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

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

S. Cucerzan. 2011. TAC Entity Linking by Performing Full-document Entity Extraction and Disambiguation. Proceedings of the Text Analysis Conference 2011, http://www.nist.gov/tac/publications/2011/participant.papers/MS_MLI.proceedings.pdf

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

J. Mayfield, J. Artiles, and H. T. Dang. 2012. Overview of the TAC2012 Knowledge Base Population Track. Text Analysis Conference 2012. http://www.nist.gov/tac/publications/2012/additional.papers/KBP2012_overview.notebook.pdf

S. Monahan, J. Lehmann, T. Nyberg, J. Plymale, and A. Jung. 2011. Cross-lingual Cross-document Coreference with Entity Linking. Proceedings of Text Analysis Conference 2011, http://www.nist.gov/tac/publications/2011/participant.papers/lcc.proceedings.pdf.

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