Basis Technology at TAC 2013 Entity Linking

Yuval Merhav
Joel Barry
James Clarke
David Murgatroyd
Basis Technology Corp.
One Alewife Center
Cambridge, MA 02140, USA
{yuval|joelb|jclarke|dmurga}@basistech.com

Abstract

Basis Technology participated in the TAC Entity-Link task of the Knowledge Base Population track at TAC 2013. This paper describes the system we developed and runs submitted for English, Chinese, and Spanish evaluation. The system is an extended and improved version of the system used in TAC 2012. We focus on the novel components and error analysis.

1 Introduction

The TAC entity linking task is to link name mentions of entities in a document collection to entities in English Wikipedia (the Knowledge Base (KB)), or to new named entities discovered in the collection. This year the document collection is a combination of newswire articles, blog posts, newsgroups, and discussion fora. The KB provided by TAC was derived from a 2008 English Wikipedia dump and contains over 800K entities. The three tasks we participated in are: (1) English document linking; (2) Chinese and English document linking; and (3) Spanish and English document linking. In all tasks the KB to link to is the same English Wikipedia KB.

The general architecture of the system used in TAC 2012 has been preserved. The main changes to our system this year are:

- a new KB derived from newer English and Chinese Wikipedia dumps. The KB contains significantly more entities than the TAC KB. It also contains additional data not available in the TAC KB (redirects, important links, etc.).
- improved data pre-processing (e.g., better in-document coreference algorithm, Chinese text normalization, etc.).
- improved Chinese candidate selection.
- more training data. All our submissions are based on models trained on all TAC 2009 - 2012 datasets.

2 System Description

A full and detailed description of our system is given in (Clarke et al., 2012). In this section we give a short summary of the core system and discuss the new components.

Our entity linking system is an extension of our incremental cross-document coreference system, which we approach as a clustering problem. We wish to identify sets of in-document coreference chains that refer to the same entity. Given an in-document coreference chain x and a set of clusters Y, the goal is to determine which cluster $y \in Y$ to place x or to create a new cluster y' with the singleton x. Initially Y is empty, but grows as the system processes new documents. For the entity-linking task we seed the initial set of clusters. One cluster is created per knowledge base entry. Each seeded cluster contains a single in-document chain automatically extracted from the knowledge base for each Wikipedia language. The system operates in three stages. First it builds a representation of the in-document chain, then generates a set of candidate clusters and finally generates features for the chain and candidate clusters and performs inference. For the latter we use a structural support vector machine algorithm with the following loss function:

$$l(y, y^*) = \begin{cases} 1,\ y \neq y^* \\ 0,\ y = y^* \end{cases} \qquad (1)$$

2.1 Query Extension

TAC specifies queries as sub-strings within a document. For many queries there is usually a co-referent and less ambiguous mention in the document. This makes entity extraction and in-document coreference crucial components. We employ several techniques to extract the in-document chain used for linking given only the query and document. If the query exactly matches a named entity identified by our in-house Rosette Entity Extractor (REX) then we use the in-document chain and type from REX (unless the type is not person, organization or location). When an exact match cannot be found we override REX forcing the query to be annotated as a named entity. The entity type is determined using a large corpus automatically annotated by REX. We then run our in-house in-document coreference resolution algorithm and represent the in-document chain by the longest mention in the chain. Figure 1 shows an example of the TAC query “杰克逊” (Jackson) that we tag as a Person and chain to “阿方索 • 杰克逊” (Alfonso Jackson). The latter is less ambiguous for entity linking.

2.2 KB Construction

We start by building a KB from Wikipedia dumps. We use the Sweble Wikitext parser (Dohrn and Riehle, 2011), an open source software tool to parse the Wikitext markup language used by MediaWiki, the software behind Wikipedia. The concatenation of the dump language and Wikipedia internal page ID is used as a globally Unique Identifier (e.g., en_10178154 is the GUID for “Basis Technology Corp.”). Every page is populated with various fields such as title, name, text, links, inter-language links, categories, infobox class and facts, and others. We only store pages from the *Main* namespace, including redirect and disambiguation pages. To reduce the size of the final KB we skip pages that do not include an infobox and pages classified as Miscellaneous by our Wikipedia type classifier (TAC train and evaluation datasets only contain Person, Location, and Organization entities). In our submissions we used an English Wikipedia dump from November 2012, and a Chinese Wikipedia dump from December 2012.

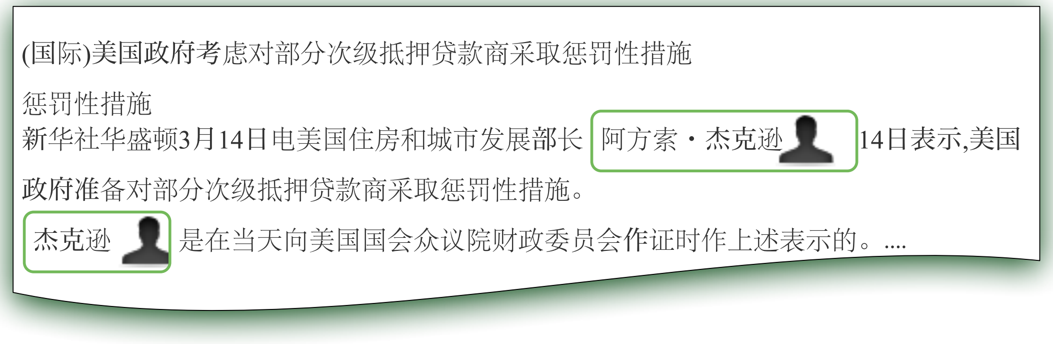
2.3 Wikipedia Redirects and Disambiguation Filters

Candidate selection has been a popular method in state-of-the-art systems in order to reduce the size of candidates per query (McNamee et al., 2011; Dredze et al., 2010). We perform candidate selection using a variety of filtering techniques that are tuned for high recall while still dramatically reducing the set of clusters to consider. As in last year, we use a Name Similarity Filter using Basis Technology’s Rosette Name Indexer (RNI) and an Anchor Text Filter with data obtained from Google Cross-wiki (Spitkovsky and Chang, 2012).

The Cross-wiki resource is useful but noisy and only maps strings of text to English Wikipedia. Also, Wikipedia is a dynamic KB that keeps growing. Consequently, we built a similar (from less data) resource from two Wikipedia sources: redirect pages and disambiguation pages.

A redirect is a page which has no content itself, but sends the reader to another article. For example, a user who searches “Man United” in Wikipedia will be taken to the article “Manchester United F.C.”. By definition a redirect is a unique mapping, i.e., a redirect can only map to a single Wikipedia page. As a result, redirects are high precision alternative names, or aliases. We augment the Anchor Text Filter with aliases from redirect pages.

In many cases a name can be ambiguous enough not to qualify as a redirect. For example, unlike “Man United”, the name “United”



(国际)美国政府考虑对部分次级抵押贷款商采取惩罚性措施

惩罚性措施
新华社华盛顿3月14日电美国住房和城市发展部长 阿方索・杰克逊 14日表示,美国政府准备对部分次级抵押贷款商采取惩罚性措施。

杰克逊 是在当天向美国国会众议院财政委员会作证时作上述表示的。....

Figure 1: Example from TAC 2013 zho-eng-eval: The TAC query "杰克逊" (Jackson) is chained to the less ambiguous mention "阿方索 • 杰克逊" (Alfonso Jackson).

is too ambiguous to be a redirect to "Manchester United F.C.". Instead, Wikipedia contains a disambiguation page titled "United", listing pages that "United" may refer to. ("Manchester United F.C.", "United Airlines", etc.). We associate every disambiguation page title with all the pages listed on the page[1], which are also added to the Anchor Text Filter. We assume a probability of 1.0 to redirects and disambiguation aliases. In the future we plan to estimate a probability based on the number of links. We use these aliases as filters in our candidate selection phase and also as indicator feature functions.

For non-English Wikipedia, we also map the redirects and disambiguation aliases to English Wikipedia if there is an inter-language link. For example, "曼聯" is a Chinese redirect to "曼徹斯特聯足球俱樂部", which is the Chinese page for "Manchester United F.C.". Since the Chinese page links to the English page, we also associate "Manchester United F.C." with the alias "曼聯".

2.4 Candidate Entity Context Feature

Every candidate retrieved by one of the filters, including a dummy new-cluster candidate, is scored based on feature values, and the top candidate is picked as the answer. Our 2012 system employs many context features. For every indoc chain we build various context vectors (the surrounding terms, entities, etc.) which we then compare against the candidate cluster contexts. This year we added a new context feature that improved the results on previous TAC datasets up to three points. Similarly to (Cucerzan, 2007), for each KB entity we create a context vector containing those entities mentioned in the first paragraph of its Wikipedia page, and those for which the corresponding pages refer back to the targeted entity. We refer to these entities as important entities. For example, "Stanford University" is an important entity for "Google" and vice versa, since their Wikipedia pages link to each other. This step is done right after the construction of the KB.

Then, for a given query in a source document, we build a context vector which we compare against the important entities vectors of KB candidates. The context vector contains KB entities retrieved by our alias filters for each of the surrounding entity mentions for the query. In other words, this context vector contains all the possible[2] disambiguations of the entity mentions in the document besides the query itself.

3 Evaluation

We learn the weight vector for the model using a structural support vector machine. This is a supervised learning setting which requires train-

[1] For simplicity we extract all links from every disambiguation page.

[2] All the possible disambiguations considering only our alias filters.

ing examples. Our training examples consist of the filtered set of candidate clusters (the truth cluster is not added if it was not included in the filters). Recall that the feature vector is generated between the current in-document chain (query) and candidate cluster.

3.1 Experimental Results

Table 1 shows a summary of our 2012 results from last year. Our **current** system performs significantly higher on the 2012 evaluation datasets: 0.6773 on English, 0.7219 on Chinese, and 0.6905 on Spanish. These scores are based on a model trained on TAC 2009 - 2012 datasets, not including the 2012 evaluation datasets.

For our submission this year we trained a model on the TAC 2009, 2010, 2011, and 2012, training and evaluation data. Table 2 shows a summary of our 2013 results. The measurements against the 2013 evaluation data are in the low-to-mid 60s of F1. In comparison to 2012, we can see a significant improvement in ORG and LOC types.

We also submitted other runs with slight modifications:

- Not using Cross-wiki in the candidate selection phase to reduce the number of noisy candidates

- Using an English only KB to reduce the number of entities in the KB

- Sorting input documents by source: News first (NYT), discussion forums last. Our system processes document incrementally so it might be better for it to see "good" queries first

- Using Google Translate to translate Chinese queries

The last modification is the only one which made a significant difference. It improved Chinese F1 scores from 0.63 to 0.66, which is roughly equivalent to the highest TAC score achieved by any team on this dataset, including systems that accessed the web at run-time. Among runs that have not accessed the web at run-time, our system was the top performer on Chinese.

Google Translate. As mentioned previously our system employs multiple candidate selection filters. One of these filters is based on our Rosette Name Indexer (RNI) that provides a similarity score between two name strings. RNI has multilingual functionality which is capable of comparing names from different languages and scripts. However, its support for Chinese ORG and LOC entities is still under development. As a result, we added a filter that queries the KB with the translation of Chinese queries using Google Translate. In addition an indicator feature function that fires when there is a match between the translation and KB candidate.

4 TAC 2013 Error Analysis

Table 3 lists F1 scores for KB and non-KB queries, and also a breakdown by document source. Not surprisingly our system performs worst on the discussion forums (DF) data. Many of the DF documents this year contain very long discussion threads. We trained our system on previous years data which consist mainly of news stories.

We discovered that our candidate selection filters retrieved the correct KB candidate for only 86%, 79%, and 80% of the queries, for English, Chinese, and Spanish, respectively. Further analysis showed that many of these queries for which we failed to filter the right KB candidate, are slang/insulting nicknames found in discussion forums and are highly uncommon in newswire. Here are several examples from the gold data:

- Barack Obama: "Obomber", "Bamster", "Bammy", "Owebama", "Obambi" and "Obamadinejad". (among others)

- George W. Bush: "Dubya", "Bushitler" and "Shrub". (among others)

- Mitt Romney: "Romnuts" and "Mittens".

- Toronto: "hogtown", "T Dot" and "T.O.".

- Sarah Palin: "Caribou Barbie".

Run ID	Dataset	PER	ORG	LOC	ALL
`basistech1`	English-English	0.784	0.387	0.440	0.566
`basistech1`	English-Chinese	0.561	0.495	0.634	0.565
`basistech1`	English-Spanish	0.826	0.521	0.454	0.595

Table 1: TAC F1 measurements for 2012 evaluation data (broken down by query type).

Run ID	Dataset	PER	ORG	LOC	ALL
`basistech1`	English-English	0.726	0.624	0.557	0.633
`basistech2`	English-Chinese	0.558	0.624	0.703	0.631
`basistech2`	English-Spanish	0.646	0.687	0.611	0.651

Table 2: TAC F1 measurements for 2013 evaluation data (broken down by query type).

- London: “big smoke” and “Londres”.

Such queries are of less importance to our target audience.

4.1 Linking Errors

Also in Table 3 we can see that our system makes more errors on KB queries (linking) than NIL ones. One reason for that is the relatively low recall of the candidate selection filters. To shed more light we quantified the different errors our system makes on KB queries: When the system makes an error it means our ranking algorithm placed a wrong candidate at the top; this candidate can be either a KB cluster, a non-KB cluster (existing NIL cluster), or the dummy new cluster candidate that is always added. Table 4 shows that for all datasets, in the majority of linking errors, it is the new cluster that is placed incorrectly at the top. One reason for this is when the correct candidate is not picked by the filters, new cluster is actually the correct candidate to place at the top. Another reason is that our system does not perform well on very short documents, which are more common this year than in previous years. In such documents there is little useful context and the main evidence the system can rely on is the “prior” probability of a candidate, sometimes refereed to as Commonness. We are planning on investigating why our system often prefers new clusters over candidates with high “prior” probability. One possibility is that our training data characteristics are significantly different than the 2013 datasets.

4.2 Chinese Linking

Linking Chinese queries to an English KB poses the same challenges as linking English and many more. One such challenge is Chinese NER (Duan and Zheng, 2011). While our tools support many languages including Chinese, accuracy is not always as good as it is in English. One important change we made this year is normalizing all Chinese text (including in the KB) to Simplified Chinese[3]. This is also consistent with our entity extraction model that was trained on Simplified Chinese. We also normalized the “middledot” character since it is used frequently to separate the given and family name of non-Chinese names (See Figure 1 for one example). Common Unicode code points used for the “middledot” in Wikipedia and source documents are “U+00B7”, “U+2027”, and “U+30FB” among others.

In the following section we provide a few linking errors the system made due to early errors in the pipeline.

4.2.1 Chinese Examples

In the following document the TAC query is “阿諾” (Arnold), which is quite ambiguous. REX missed the full mention “阿諾舒華辛力” (Arnold Schwarzenegger) (in bold) which made

[3] Normalizing to Simplified Chinese is an unambiguous direct mapping unlike converting Simplified Chinese to Traditional Chinese.

Run ID	Dataset	in KB	not in KB	News docs	Web docs	Forum
`basistech1`	English-English	0.565	0.709	0.729	0.595	0.492
`basistech2`	English-Chinese	0.621	0.645	0.623	0.641	–
`basistech2`	English-Spanish	0.612	0.705	0.652	–	–

Table 3: TAC F1 measurements for 2013 evaluation data (broken down by query and document type).

Dataset	Correct	Incorrect, KB Cluster	Incorrect, NIL Cluster	Incorrect, New Cluster
English-English	64%	7%	4%	25%
English-Chinese	68%	5%	9%	18%
English-Spanish	66%	7%	3%	24%

Table 4: Distribution of the top ranked candidates by our system on TAC 2013 evaluation data (KB queries only): Correct means the top candidate is the correct KB candidate. When the top candidate is incorrect there are three possibilities: (1) It is a KB cluster; (2) It is an existing NIL cluster; or (3) It is the dummy new cluster that is added for every query.

the linking task much harder than it could have been.

肯雅舉國準備慶祝奧巴馬當選美國總統

美國的軟力量在世界上仍然有舉足輕重的地位。

六年前，奧地利之子，大隻佬**阿諾舒華辛力加**參選美國加州州長，奧地利整國人民徹夜看加州州長大選，直到阿諾當選為止。...

Mixed language documents are also a challenge. For example, in the following document the TAC query is “CEA”, an English entity in a Chinese document. Unlike the Schwarzenegger case, in this document REX did tag the less ambiguous mention “美国消费电子协会” (“U.S. Consumer Electronics Association”). However, our in-house in-document coreference resolution did not chain the two. Consequently, our system placed the right candidate below the top candidate “Cinema Exhibitors’ Association”.

... 本届博览会由商务部、信息产业部、科技部、山东省人民政府主办, 美国消费电子协会 (CEA) 担任海外主办单位, 中国电子商会和青岛市人民政府承办。...

5 Conclusions

The paper described the system developed for entity linking at Basis Technology and evaluation results in the TAC 2013 evaluation for English, Chinese, and Spanish. We found that many errors are due to difficult document sources (discussion forums and short snippets) and early pipeline errors in the Chinese case.

Acknowledgments

This research is partially supported by the U.S. government. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of any U.S. government agency.

References

Clarke, James, Yuval Merhav, Ghalib Suleiman, Shuai Zheng, and David Murgatroyd. 2012. Basis technology at tac 2012 entity linking. In *Proceedings of the Text Analysis Conference 2012*.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP-CoNLL 2007*. pages 708–716.

Dohrn, Hannes and Dirk Riehle. 2011. Design and implementation of the sweble wikitext parser: unlocking the structured data of wikipedia. In *Int. Sym. Wikis*. ACM, pages 72–81.

Dredze, Mark, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *COLING*. pages 277–285.

Duan, Huanzhong and Yan Zheng. 2011. A study on features of the crfs-based chinese named entity recognition. *International Journal of Advanced Intelligence* 3(2):287–294.

McNamee, Paul, James Mayfield, Dawn Lawrie, Doug Oard, and David Doermann. 2011. Cross language entity linking. In *International Joint Conference on Natural Language Processing*.

Spitkovsky, Valentin I. and Angel X. Chang. 2012. A cross-lingual dictionary for English Wikipedia concepts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey.