# CUNY-UIUC-SRI TAC-KBP2011 Entity Linking System Description

**Taylor Cassidy**[1], **Zheng Chen**[1], **Javier Artiles**[1], **Heng Ji**[1],
**Hongbo Deng**[2], **Lev-Arie Ratinov**[2], **Jing Zheng**[3], **Jiawei Han**[2], **Dan Roth**[2]

[1]Computer Science Department and Linguistics Department
Queens College and Graduate Center, City University of New York, New York, NY, USA

[2] Computer Science Department
University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA

[3] SRI International, Menlo Park, CA, USA

hengji@cs.qc.cuny.edu

## Abstract

In this paper we describe a joint effort by the City University of New York (CUNY), University of Illinois at Urbana-Champaign (UIUC) and SRI International at participating in the mono-lingual entity linking (MLEL) and cross-lingual entity linking (CLEL) tasks for the NIST Text Analysis Conference (TAC) Knowledge Base Population (KBP2011) track. The MLEL system is based on a simple combination of two published systems by CUNY (Chen and Ji, 2011) and UIUC (Ratinov et al., 2011). Therefore, we mainly focus on describing our new CLEL system. In addition to a baseline system based on name translation, machine translation and MLEL, we propose two novel approaches. One is based on a cross-lingual name similarity matrix, iteratively updated based on monolingual co-occurrence, and the other uses topic modeling to enhance performance. Our best systems placed 4th in mono-lingual track and 2nd in cross-lingual track.

## 1 Introduction

Entity Linking in Knowledge Base Population (KBP) has been designed in a "top-down" fashion, namely, a system is provided with a name mention in a context, and is evaluated based on whether it can determine which of a set of target entities that mention denotes. However, a "bottom-up" task in which a system must link all name mentions in a set of source documents with knowledge base (KB) entities would be more desirable, albeit more difficult to perform and evaluate. The introduction of the NIL clustering component this year brings the entity linking task a step closer to this goal. Many useful applications will require this capability, such as assisting scientific paper reading by providing links to Wikipedia pages for all technical terms. In fact, our existing UIUC "Wikification" system (Ratinov et al., 2011) aims to link all possible concepts to their corresponding Wikipedia pages. We can extract all entities in the context of a given query, and disambiguate all entities at the same time. Our existing CUNY system (Chen and Ji, 2011) further extends this idea to the cross-document level by constructing "collaborators" for each query and exploiting the global context from the entire collaborator cluster for each query. Both systems exploit information redundancy for entity linking and achieved similar performance. However, as they have used algorithms and resources that are quite different, the next logical step is to combine them in an attempt to enhance entity linking performance.

There are two conventional ways to extend mono-lingual entity linking systems to a cross-lingual setting: (1) Apply a Source Language (SL) MLEL system to link SL name mentions to SL KB entries, and then link the SL KB entry to the corresponding Target Language (TL) KB entry; (2) Apply machine translation to translate

the SL document into the TL, and then apply a TL MLEL system to link the translated document to a TL KB entry. However, these approaches may be problematic: approach (1) relies heavily on the existence of an SL KB whose size is comparable to the TL KB, and thus is not easily adaptable to other low-density languages; approach (2) relies on machine translation (MT) output, and as such it may suffer from translation errors, particularly those involving named entities.

In order to enhance both the portability and reduce the cost of cross-lingual entity linking, we have developed a novel approach that does not need MT nor source language KB. Our research hypothesis is that the query entities can be disambiguated based on their "collaborators" or "supporters", namely other entities which co-occur or are related to the queries. For example, three different entities with the same name spelling "阿尔伯特/Albert" can be disambiguated by their respective context entities (affiliations): "比利时/Belgium", "国际奥委会/International Olympic Committee" and "美国科学院/National Academy of Sciences". We construct a large entity supporting matrix to jointly translate and disambiguate entities.

In addition, most context-based ranking features follow the distributional hypothesis (Harris, 1954), namely that queries sharing topically-related contexts tend to link to the same KB entry. If we consider the KB entry denoted by a query to be its *sense*, we can also follow the "One Sense Per Discourse" hypothesis proposed by Gale et al. (1992). Topic modeling provides a natural and effective way to model the context profile of each query (Kozareva and Ravi, 2011). Similar entities in a single coherent latent topic tend to express the same sense, and thus should be linked to the same KB entry. For example, the query, "*Li Na*" is associated with a sports topic cluster represented by, "{*tennis, player, Russia, final, single, gain, half, male, ...*}", and another query, "*Li Na*" is associated with a politics topic cluster represented by, "{*Pakistan, relation, express, vice president, country, Prime minister, ...*}", so they are likely referring to two different entities. In addition, each KB article can be considered as an entity-level semantic topic. Therefore we also applied topic modeling with a biased propagation method to the Chinese source documents, implicitly assuming consistency of results among entities in each topic cluster based on our second hypothesis: "one entity per topic cluster". Compared to the best baseline system that used English entity linking, this new approach achieved 11.2% improvement in B-Cubed+ F-measure.

## 2 Related Work

Although this year's evaluation is the first to include the CLEL task, similar efforts have been published in recent papers (McNamee et al., 2011), but with evaluation settings and query selection criteria that are quite different. Almost all CLEL systems participating in the KBP2011 track (e.g. (McNamee et al., 2011), (Monahan et al., 2011), (Fahrni and Strube, 2011)) followed the approaches outlined above (MLEL using a source language KB or MLEL on MT output).

Some previous work applied similarity metrics to or used links between each multilingual pair of names to summarize multi-lingual Wikipedias (Filatova, 2009), find similar sentences (Adafre and Risjke, 2006; Bharadwaj and Varma, 2011) or extract bi-lingual terminology (Erdmann and Nakayama, 2009). Some recent name mining work has been based on aligning Multi-lingual Wikipedia Pages (Hassan et al., 2007; Richman and Schone, 2008), Infoboxes (Adar et al., 2009; Bouma et al., 2009; de Melo and Weikum, 2010; Navigli and Ponzetto, 2010; Lin et al., 2011), information networks (Ji, 2009) or social networks (You et. al., 2010). To the best of our knowledge, our joint modeling of name mining and disambiguation is the first work to apply unsupervised name mining to enhance entity linking.

(Kozareva and Ravi, 2011) applied topic modeling for the Web People Search task (Artiles et al., 2010). We extended this idea from the mono-lingual to the cross-lingual setting. Our topic modeling method treats multi-typed entities differently along with their inherent textual information and the rich semantics of the relationships, therefore it can provide more gains to entity linking.

## 3 Task Definition

The cross-lingual entity linking (CLEL) task we are addressing is that of the NIST TAC Knowledge Base Population (KBP2011) evaluations (Ji et al., 2011). In the CLEL task, given a Chinese or English query that consists of a name string - which can be a person (PER), organization (ORG) or geo-political entity (GPE, a location with a government) - and a source document ID, the system is required to provide the ID of an English Knowledge Base (KB) entry to which the name refers; or NIL if there is no such KB entry. In addition, a CLEL system is required to cluster together queries referring to the same entity not present in the KB and provide a unique ID for each cluster.

KBP2011 used a modified B-Cubed (Bagga and Baldwin, 1998) metric (called B-Cubed+) to evaluate entity clusters. Given an entity mention $e$, we use the following notation: $L(e)$ denotes the category of the entity mention, $C(e)$ denotes its cluster, $SI(e)$ denotes its KB identifier as provided by the system in question, and $GI(e)$ denotes the gold-standard KB identifier for $e$. The correctness of the relation between two entity mentions $e$ and $e'$ in the distribution is defined as:

$$\mathrm{G}(e,e') = \begin{cases} 1 \text{ iff } L(e) = L(e') \wedge C(e) = C(e') \wedge \\ GI(e) = SI(e) = GI(e') = SI(e') \\ \\ 0 \text{ otherwise} \end{cases}$$

That is, two entity mentions are correctly related when they share a category if and only if they appear in the same cluster and share the same KB identifier in the system and the gold-standard. The B-cubed+ precision of an entity mention is the proportion of correctly related entity mentions in its cluster (including itself). The overall B-Cubed+ precision is the averaged precision of all mentions in the distribution. Since the average is calculated over mentions, it is not necessary to apply any weighting according to the size of clusters or categories. The B-Cubed+ recall is analogous, replacing "cluster" with "category". Formally:

$$\text{Precision} = \text{Avg}_e[\text{Avg}_{e'.C(e)=C(e')}[\mathrm{G}(e,e')]]$$

$$\text{Recall} = \text{Avg}_e[\text{Avg}_{e'.L(e)=L(e')}[\mathrm{G}(e,e')]]$$

$$\text{F-Measure} = 2{\times}Precision{\times}Recall/(Precision{+}Recall)$$

## 4 Mono-lingual Entity Linking

### 4.1 CUNY System

Our mono-lingual entity linking system is a combined approach of the collaborative ranking framework described in the CUNY system (Chen and Ji, 2011) and the UIUC Wikification techniques described in (Ratinov et al., 2011; Ratinov, 2011). The combination is done based on majority voting on the following three submissions from CUNY system and three submissions from UIUC system.

In CUNY system, we developed 5 baseline rankers, including 2 unsupervised rankers ($f_2, f_3$) and 3 supervised rankers($f_5, f_6, f_8$):

•Entity ($f_2$): $f_2$ is defined as weighted combination of entity similarities in three types (person, organization and geo-political). Name entities are extracted from $q.text$ and KB text respectively using Stanford NER toolkit[1]. The formulas to compute entity similarities are defined in (Yoshida et al., 2010).

•Tfidf ($f_3$): $f_3$ is defined as the cosine similarity between $q.text$ and KB text using tfidf weights.

•Maxent ($f_5$): a pointwise ranker implemented using OpenNLP Maxent toolkit[2] which is based on a maximum entropy model.

•SVM ($f_6$): a pointwise ranker implemented using $SVM^{light}$ (Joachims, 1999).

•ListNet ($f_8$): a listwise ranker presented in (Cao et al., 2007).

The four supervised rankers apply exactly the same set of features except that SVM ranking ($f_7$) needs to double expand the feature vector. The features are categorized into three levels, surface features (Dredze et al., 2010; Zheng et al., 2010), document features (Dredze et al., 2010; Zheng et al., 2010), and profiling features (entity slots that are extracted by the slot filling toolkit (Chen and Ji, 2011)).

### 4.2 UIUC System

For the UIUC systemwe used GLOW, an off-the-shelf system we have developed for the

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

[2] http://maxent.sourceforge.net/about.html

related task of Disambiguation to Wikipedia (D2W) . The GLOW system takes as input a text document $d$ and a set of mentions $M = \{m_1, m_2, \ldots, m_N\}$ in $d$, and cross-links them to Wikipedia, which acts as a knowledge base. This is done through combining local (lexical and Wikipedia title prevalence) clues with global coherence of the candidate joint cross-linking assignment which is done by analyzing Wikipedia link structure and estimating pairwise article relatedness. The key advantage of GLOW as reported by (Ratinov et al., 2011) is using different strategies for forming an approximate solution to the input problem and using it as a semantic disambiguation context for all the mentions. This allows GLOW to maintain a tractable inference by disambiguating each mention locally while capturing important global properties. In fact, GLOW stands for Global and Local Wikification.

However, GLOW cannot be applied to entity linking directly since there are subtle differences between the D2W and entity linking tasks. More specifically, in D2W the set of input mentions is tied to specific locations in the text, thus potentially the same surface form may refer to different entities. For example, a review about a movie Titanic may use the same surface form "Titanic" to refer both to the ship and to the movie. In D2W, each mention of the ship sense would be cross-linked to `http://en.wikipedia.org/wiki/RMS_Titanic`, while each mention corresponding to the movie would be cross-linked to `http://en.wikipedia.org/wiki/Titanic_(1997_film)`. This scenario does not occur in the TAC KBP entity-linking task where a "one sense per document" requirement holds. On the other hand, in the entity linking task, the following query is possible ( *Query_ID, "Ford", "The Ford Library is named after Gerald Ford"* ). In D2W, the above text would contain two mentions: "Ford Library" and "Gerald Ford", both of which are easy to disambiguate, but in the entity linking task it is necessary to understand that in both mentions of "Ford" refer to Gerald Ford. These differences and the choice of using a D2W component as an inference driver for an entity-linking system has dictated the structure of the letter. Namely our entity-linking system is composed of the following steps:

1) Identify the mentions in the text which correspond to the query. We experimented with two approaches. A *Naive Mention Identification* simply marks all the instances of the query form in the text, while a *Named Entity Mention Identification* maps the query form to all the named entities containing the form. For example, let us consider the query ( *QID, "Ford", "The Ford Library is named after Gerald Ford"* ). The naive approach would mark the underlined surface forms as an input for GLOW: *"The Ford Library is named after Gerald Ford"* . The named entity matching approach would mark the underlined surface forms as an input for GLOW: *"The Ford Library is named after Gerald Ford"* .

2) Disambiguation - this is a straightforward application of the GLOW system. We note that the GLOW system assigns each mention a disambiguation along two confidence scores: the ranker score and the linker score. Roughly speaking, the ranker score indicates the confidence that the selected disambiguation is more appropriate than the discarded disambiguation, while the linker score is the confidence that the selected disambiguation is indeed the correct one. The linker was trained to assign -1 scores to all surface forms which cannot be mapped to the entry base.

3) Selecting a single disambiguation based on thresholding as described in (Ratinov, 2011).

### 4.3 System Combination

We applied a new ranking scheme proposed by (Chen and Ji, 2011), collaborative ranking (CR), to combine these baseline systems. The CR framework identifies collaborators for each query based on an agglomerative clustering approach and a graph-based clustering approach, and integrates the strengths from multiple collaborators of a query (Micro Collaborative Ranking) and the strengths from multiple ranking algorithms (Macro Collaborative Ranking). Table 1 summarizes the individual runs submitted for the evaluation.

| System | | Description |
|---|---|---|
| CUNY System | CUNY1 | ListNet Ranking |
| | CUNY2 | MaxEnt Ranking |
| | CUNY3 | SVM Ranking |
| | CUNY4 | TfIdf Ranking |
| | CUNY5 | Entity Ranking |
| UIUC System | UIUC1 | Without Query Expansion, Without Thresholding |
| | UIUC2 | WithQuery Expansion, Without Thresholding |
| | UIUC3 | With Query Expansion, With Thresholding |
| Combined System | Combined1 | CUNY1 + CUNY2 + UIUC1 + UIUC2 + UIUC3 |
| | Combined2 | CUNY1 + CUNY2 + CUNY3 + CUNY4 + UIUC1 + UIUC2 + UIUC3 |
| | Combined3 | CUNY1 + CUNY2 + CUNY3 + CUNY4 + CUNY5 + UIUC1 + UIUC2 + UIUC3 |

Table 1: Mono-lingual Entity Linking System Combination

## 5 Cross-lingual Entity Linking

### 5.1 System Overview

Figure 1 depicts the overall pipeline of our cross-lingual entity linking system.

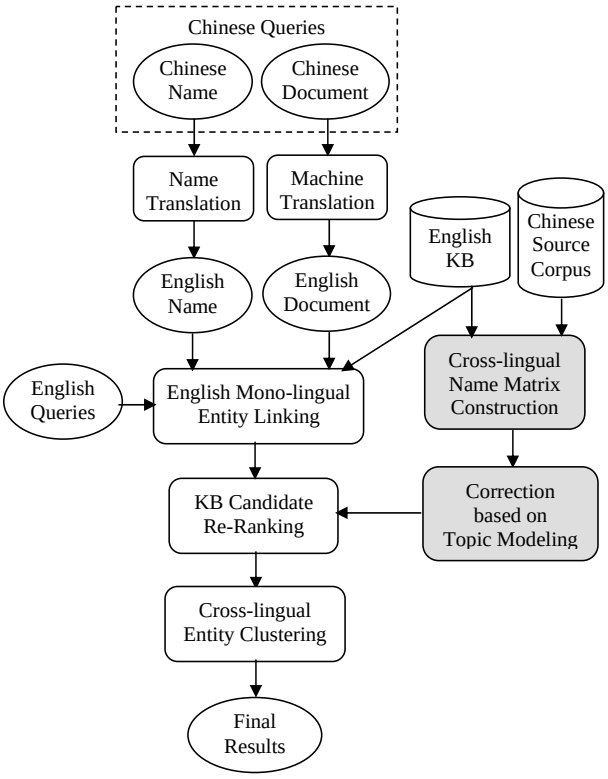


Figure 1: Cross-lingual Entity Linking System Overview

We have developed a baseline approach consisting of name translation, machine translation and mono-lingual entity linking, which belongs to the "Pipeline A" category as summarized in the KBP2011 overview paper (Ji et al., 2011).

In addition, we made two significant enhancements as follows. For each query, we applied the mono-lingual entity linking system to generate the top N KB node candidates, making use of multiple similarity scores including document-level tf.idf score, fuzzy name matching similarity based on edit distance, along with the multiple ranking scores as described in section 4.1. For Chinese person queries, we categorize them into Chinese names or non-Chinese names, selecting a different N for each category since some common Chinese last names (e.g. "Li", "Wang") can retrieve too many KB node candidates.

Then we developed a novel joint approach of translating and disambiguating entities through cross-lingual name co-occurrence matrix construction (section 5.3). From this matrix we can extract a uniformed score about semantic context similarity between each pair of query document and KB article. This context similarity score is then linearly combined with the mono-lingual entity linking scores based on weights optimized from the 2011 training data set. Finally, we applied a new topic modeling approach with biased propagation (Deng et al., 2011) to ensure the consistency of entity linking results within each topic cluster (section 5.4).

### 5.2 Query and Document Translation

The baseline system first translates a Chinese query and its associated document into English, and

then runs English MLEL to link the translated query and document to the English KB. We added the following new components to adapt the mono-lingual system to a cross-lingual setting.

We apply a Chinese name coreference resolution system (Ji et al., 2005) to each source document in order to get name variants for a given query. Then we applied various name translation approaches as described in (Ji et al., 2009) to translate these name variants. In addition, we mined a lot name pairs offline from multi-lingual Wikipedia Infoboxes using an unsupervised learning method as described in (Lin et al., 2011).

We applied the SRI hierarchical phrase-based machine translation system as described in (Zheng et al., 2009) to translate source documents. The system is based on a weighted synchronous context-free grammar (SCFG). All SCFG rules are associated with a set of features that are used to compute derivation probabilities under a log-linear model. The features include:

- Relative frequency in two directions
- Lexical weights in two directions
- Phrase penalty
- Hierarchical rule penalty
- Glue rule penalty

In addition to these rule-related features, we also use target language model score and target sentence lengths. The scaling factors for all features are optimized by minimum error rate training (MERT) to maximize BLEU score. Given an input sentence in the source language, translation into the target language is cast as a search problem, where the goal is to find the highest-probability derivation that generates the source-side sentence, using the rules in our SCFG. The source-side derivation corresponds to a synchronous target-side derivation and the terminal yield of this target-side derivation is the output of the system. We used an SRI-developed, CKY-style decoder to solve the search problem.

### 5.3 Modeling Information Networks

As we pointed out in the introduction, there are some principle limitations of the basic approaches. In addition, some source language-specific characteristics also require us add more fine-grained contexts as ranking features. For example, when foreign politician names appear in Chinese documents, they normally only include last names. To some extent this introduces extra ambiguities to the cross-lingual setting.

When humans determine the identity of an entity mention, they would first check the "profile" of the entity, such as a person's title, origin and employer, or in which country a city entity is located.

Inspired from this intuition, we proposed and developed a novel approach to jointly mine entity translations and disambiguate entities based on entity profile comparison. Therefore the first step is to automatically extract profiles for entities. We applied a slot filling system (Chen et al., 2010) to extract entity contexts with a representation called "Information Networks" (Li et al., 2011). This approach is particularly effective to disambiguate entities with common organization names or person names. For example, many countries can have a "Supreme Court" (in Japan, China, U.S., Macedonia, etc.) or an "LDP (Liberty and Democracy Party)" (in Australia, Japan, etc.); "Newcastle University" can be located in UK or Australia; "Ji county" can be located in "Shanxi" or "Tianjin"; and many person entities share the same common names such as "Albert", "Pasha", etc.

However, not all entities have explicit profiles presented in the source documents. For example, Table 2 presents the various types of entity contexts that may help disambiguate entities. In addition, some global context such as document creation time will be helpful for entity disambiguation. For example, for an entity with the common name "Shaofen Wan", if we know its associated document was from the 1992 news, then it most likely refer to the member of "13th Central Committee of the Communist Party of China".

We construct a large entity supporting matrix that stores the similarity between each entity mention in any Chinese source document and each English KB entry. Each row represents a named entity mention and each column represents a KB entry. The matrix is initialized by assigning to each cell a score measuring the similarity between the corresponding name mention and KB entry. We use a score

| Context Types | Examples | | | | |
|---|---|---|---|---|---|
| | Query | KB Node | Key Context | Context Sentence | Context Sentence Translation |
| Co-occurrence | 塞维利亚 (Sevilla) | Sevilla, Spain | 西班牙 (Spain) | **西班牙**两名飞行员 15 日举行婚礼，从而成为西班牙军队中首对结婚的同性情侣。婚礼在**塞维利亚**市政厅举行。 | Two pilots had their wedding in **Spain** on 15th, and so they became the first homosexual couple who got married in Spanish troops. The wedding was held in **Sevilla** city hall. |
| | 民主进步党 (Democratic Progressive Party) | Democratic Progressive Party, Bosnia | 波士尼亚 (Bosnia) | **波士尼亚**总理塔奇克的助理表示：“…。”由于…。 另外，伊瓦尼奇表示，在中央政府担任要职的两名他所属的**民主进步党**党员也将辞职。 | The assistant of **Bosnia** Premier Taqik said “…”. Because ... . In addition, Ivanic said, two **Democratic Progressive Party** members who held important duties in the central government… |
| Part-whole Relation | Fairmont | Fairmont, West Virginia | WV | Verizon coverage in **WV** is good along the interstates and in the major cities like Charleston, Clarksburg, **Fairmont**, Morgantown, Huntington, and Parkersburg. | - |
| | 曼彻斯特 (Manchester) | Manchester, New Hampshire | 新罕布什尔州 (New Hampshire) | 曼彻斯特 (新罕布什尔州) | Manchester (New Hampshire) |
| Employer/ Title | 米尔顿 (Milton) | NIL1 | 巴西(Brazil); 代表 (representative) | **巴西**政府高级代表**米尔顿** | **Milton**, the senior **representative** of **Brazil** government |
| | | NIL2 | 厄瓜多尔皮钦查省 (Pichincha Province, Ecuador); 省长 (Governor) | **厄瓜多尔皮钦查省省长**米尔顿 | Milton, the **Governor** of **Pichincha Province, Ecuador** |
| Start-Position Event | 埃特尔 (Ertl) | NIL3 | 智利 (Chilean) 奥委会 (Olympic Committee) 选为 (elected) 主席 (chairman) | **智利**击剑联合会领导人**埃特尔**今晚被**选为**该国**奥委会**新任**主席** | The leader of **Chilean Fencing Federation** Ertl was **elected** as the new **chairman** of this country’s **Olympic Committee** tonight. |
| Affiliation | 国家医药局 (National Medicines Agency) | NIL4 | 保加利亚 (Bulgarian) | **保加利亚国家医药局** | **Bulgarian National Medicines Agency** |
| Located Relation | 精细化工厂 (Fine Chemical Plant) | NIL6 | 芜湖市 (Wuhu City) | **芜湖市精细化工厂** | **Fine Chemical Plant** in **Wuhu City** |

Table 2: Information Networks Examples for Entity Disambiguation

based on the transliteration similarity score used in (You et. al., 2010) which is based on the minimum edit distance between the pinyin form of the Chinese source name and the English KB name, topic modeling results, and the similarity between the document and the corresponding KB entry article. All name strings and KB article titles were processed: punctuation, whitespace, parenthetical qualifiers, and any text after the first comma was removed. The matrix is iteratively updated based on the values of the cells corresponding with the pairs of the respective neighbors of the name mention and

KB entry. Neighbors of a Chinese name mention are the other name mentions in the associated source document that are involved in some relations, events, attributes or co-occur in the same sentences after coreference resolution. KB entry neighbors are calculated based on Wikipedia article link information: any two KB entries are considered neighbors if a link to one appears on the Wikipedia page of the other, or both. After a sufficient number of iterations, the updated supporting matrix can be used for clustering and name translation mining.

The similarity matrix, $R$, is updated using the update algorithm from (You et. al., 2010)

$$R_{ij}^{t+1} = \lambda * [\sum_{(u,v)_k \in B^t(i,j,\theta)} \frac{R_{uv}^t}{2^k}] + (1-\lambda) R_{ij}^0 \quad (1)$$

$R_{ij}^{t+1}$ denotes the score of **( ( nm, documentID)**$_i$**, KBid**$_j$ **)**, $B^t(i,j,\theta)$ denotes the set of pairs of the form **( ( name mention string, documentID)**$_{N(i)}$**, KBid**$_{N(j)}$ **)**, where $N(x)$ denotes the set neighbors of $x$ whose values in $R_{ij}^t$ exceed $\theta$, and $(u,v)_k$ is the pair with the $k_{th}$ highest score at iteration $t$. $\lambda$ is a linear interpolation parameter controls to what extent the scores of the neighbor pairs with high similarity scores may contribute to updating a cell's score.

For a the query name mention corresponding with row $i$, the column $j$ containing the highest score of the row is selected as the target entity. Since the matrix can be populated with any score that compares a name mention to a KB entry, updating the matrix can be used as a re-ranker based on entity co-occurrence for any scoring metric.

### 5.4 Topic Modeling

We applied an entity-driven topic modeling approach described in our recent work (Deng et al., 2011). For each source document and its associated background metadata, we extract English and Chinese named entities using a bi-lingual entity extraction system (Ji and Grishman, 2006). For Chinese data we applied the Tsinghua word segmenter (Wan and Luo, 2003) for pre-processing. The entity extraction system consists of a Hidden Markov Model (HMM) tagger augmented with a set of post-processing rules. This new topic model approach directly incorporates the heterogeneous background information with topic modeling in a unified way. The underlying intuition is that multi-typed entities should be treated differently along with their inherent textual information and the rich semantics of the relationships. For example, the topic distribution of an entity without explicit text information (e.g., person $u_l$) depends on the topic distribution of the documents that mention $u_l$. On the other hand, the topic of a document $d_j$ is also correlated with its mentioned entities to some extend, but, most importantly, its topic should be principally determined by its inherent content of the text. To incorporate both the textual information and the relationships between documents and multi-typed entities, we defined a biased regularization framework by adding the regularization terms to the log-likelihood along with their constraints.

The number of topics was estimated based on the percentage of clusters per query in the training data. After extracting topic clusters, we applied majority voting among the queries which have the same name spelling and belong to the same topic cluster, to ensure them to link to the same KB entry.

### 5.5 NIL Clustering

We have adopted a simple substring matching based approach to NIL clustering. In addition, we applied a within-document Chinese coreference resolution system (Ji et al., 2005) and some abbreviation gazetteers to expand each query (e.g. "魁北克/Quebec") into a cluster of coreferential names ("魁北克, 魁北克集团/Quebec, Quebec group") for matching.

## 6 Experiments

### 6.1 Data

The English source collection includes 1,286,609 newswire documents, 490,596 web documents, and 683 other documents. The Chinese source collection includes approximately one million news documents from Chinese Gigaword. The English reference Knowledge Base consists of 818,741 nodes derived from an October 2008 dump of English Wikipedia (Ji et al., 2011). We used KBP 2009-2011 Entity Linking training data sets and the KBP 2009-2010 evaluation data sets to develop our

systems, and then conduct blind test on KBP2011 Entity Linking evaluation data sets. The detailed data statistics are summarized in Table 3.

| Corpus | | # Queries | | |
|---|---|---|---|---|
| | | Person | Organization | GPE |
| Mono-lingual | Training | 1868 | 3960 | 1816 |
| | Evaluation | 750 | 750 | 750 |
| Cross-lingual | Training | 817 | 660 | 685 |
| | Evaluation | 824 | 710 | 642 |

Table 3: Data sets

### 6.2 MLEL Results

The B-cubed+ Precision (P), Recall (R) and F-Measure (F) results of mono-lingual entity linking systems (combined1, combined2, combined3) are summarized in Table 4, for persons (PER), organizations (ORG), geo-political entities (GPE), and overall queries (ALL) respectively.

| Entity Type | System | P | R | F |
|---|---|---|---|---|
| PER | Combined1 | 0.741 | 0.783 | 0.761 |
| | Combined2 | 0.727 | 0.769 | 0.748 |
| | **Combined3** | **0.746** | **0.786** | **0.766** |
| GPE | **Combined1** | **0.761** | **0.768** | **0.764** |
| | Combined2 | 0.691 | 0.691 | 0.691 |
| | Combined3 | 0.69 | 0.691 | 0.691 |
| ORG | **Combined1** | **0.764** | **0.800** | **0.781** |
| | Combined2 | 0.737 | 0.779 | 0.758 |
| | Combined3 | 0.749 | 0.791 | 0.770 |
| ALL | **Combined1** | **0.758** | **0.784** | **0.771** |
| | Combined2 | 0.724 | 0.748 | 0.736 |
| | Combined3 | 0.737 | 0.761 | 0.749 |

Table 4: Mono-lingual Entity Linking Evaluation Results (%)

As we add more rankers from Combined1 to Combined3, the performance for PER and ORG entity types were significantly improved. But the entity-based ranker and profile-based ranker hurt the performance for GPE entities mainly because global features are dominant for geo-political entities.

### 6.3 CLEL Results

The results of cross-lingual entity linking systems are summarized in Table 5. We can see that the approaches using the cross-lingual name co-occurrence matrix and topic modeling have significantly improved the results for Chinese queries, especially for PER and GPE entities. Cross-lingual entity linking performs significantly worse on Chinese PER queries when compared with monolingual English entity linking, mainly because the translation of PER names is the most challenging among three entity types (Ji et al., 2009). Nevertheless, we found that for some Chinese names, their Chinese spellings are much less ambiguous than English spellings because the mapping from Chinese character to pinyin is multiple-to-one. Therefore Chinese documents can actually help link a cross-lingual cluster to the correct KB entry, which is the reason some small gains were acheived in the F-measure for English queries.

| Entity Type | Query Language | System | P | R | F |
|---|---|---|---|---|---|
| PER | English | Baseline | 74.0 | 68.0 | 70.9 |
| | English | Enhanced | **76.0** | **72.1** | **74.0** |
| | Chinese | Baseline | 37.6 | 41.9 | 39.6 |
| | Chinese | Enhanced | **64.9** | **72.9** | **68.7** |
| | All | Baseline | 45.7 | 47.7 | 46.7 |
| | All | Enhanced | **67.3** | **72.7** | **69.9** |
| GPE | English | Baseline | 82.1 | 79.5 | 80.8 |
| | English | Enhanced | **81.8** | **82.0** | **81.9** |
| | Chinese | Baseline | 73.5 | 74.9 | 74.2 |
| | Chinese | Enhanced | **83.3** | **83.7** | **83.5** |
| | All | Baseline | 76.7 | 76.6 | 76.7 |
| | All | Enhanced | **82.7** | **83.0** | **82.9** |
| ORG | English | Baseline | 77.5 | 80.7 | 79.1 |
| | English | Enhanced | **79.0** | **84.7** | **81.7** |
| | Chinese | Baseline | 68.1 | 83.7 | 75.1 |
| | Chinese | Enhanced | **68.9** | **85.8** | **76.4** |
| | All | Baseline | 71.7 | 82.6 | 76.7 |
| | All | Enhanced | **72.7** | **85.4** | **78.5** |
| ALL | English | Baseline | 78.2 | 76.9 | 77.6 |
| | English | Enhanced | **79.2** | **80.4** | **79.8** |
| | Chinese | Baseline | 56.3 | 63.2 | 59.6 |
| | Chinese | Enhanced | **71.0** | **79.6** | **75.1** |
| | All | Baseline | 63.3 | 67.6 | 65.4 |
| | All | Enhanced | **73.6** | **79.9** | **76.6** |

Table 5: Cross-lingual Entity Linking Evaluation Results (%)

### 6.4 Analysis

#### 6.4.1 Difficulty Level Analysis

In Figure 2 we present the distribution of 1481 Chinese queries in the KBP2011 CLEL evaluation corpus which need different techniques, according to their difficulty levels. The percentage numbers are approximate because some queries may rely on the combination of multiple types of features.

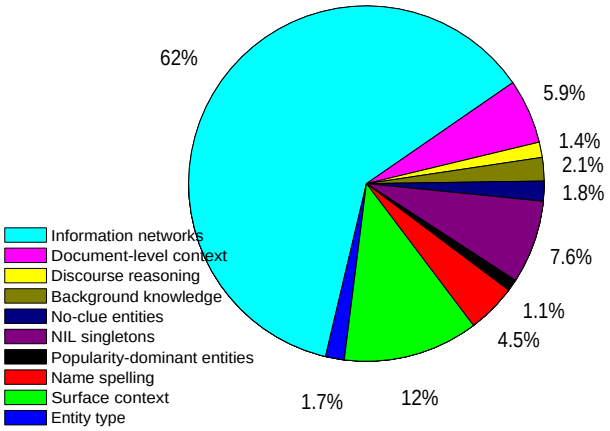


Figure 2: Distribution of 2011 CLEL queries according to difficulty levels

***NIL singletons***: About 7.6% queries are singleton entities (e.g. "中绿集团/Zhonglv Group", "丰华中文学校/Fenghua Chinese School") which don't have corresponding KB entries. Therefore the NIL detection step must form singleton clusters without considering context information.

***Popularity-dominant entities***: A few (1.1%) queries are popular entities, such as "路透社/ Reuters", and so they can be correctly linked based on popularity features alone.

***Name spelling***: 4.5% queries can be disambiguated by their full names that appear in the source documents. For example, "莱赫. 卡钦斯基/ Lech Aleksander Kaczynski" and "雅罗斯瓦夫. 卡钦斯基/ Jaroslaw Aleksander Kaczynski","田中角荣/ Kakuei Tanaka" and "田中真纪子/ Makiko Tanaka" can be disambiguated based on their firstnames; while

***Surface context*** 12% queries can be disambiguated based on some lexical features or string matching based name coreference resolution. For example, for a query "亚行/Asian Development Bank" that appears in the title of a document, a CLEL system simply needs to recognize its full name "亚州开发银行/Asian Development Bank" in order to link it to the correct KB entry.

***Entity type***: For 1.7% queries, entity type classification is crucial. For example, if we know "沙巴/Sabah" is a geo-political entity instead of a person in the source document, we can filter out many incorrect KB candidates.

***Information networks***: As we have discussed in Table 2, many entities (62% of the evaluation queries) can be linked based on their context information networks. Such information is particularly effective for those entities that may be located/affiliated in many different locations. For example, almost every city has a "交通广播电台/Traffic Radio", and every country has a "联邦法院/Federal Court", so it's important to identify the other context entities with which the query entities are associated. Information networks can be very helpful to disambiguate some highly ambiguous geo-political names if we can identify their higher-level context entities. For example, there are many different KB candidates for a query with the common name "海得拉巴/ Hyderabad"; we can correctly disambiguate the query if we know which place (e.g. " Andhra Pradesh") the query is part of.

***Document-level context***:

Document-level contexts, including what can be induced from topic modeling, are important for disambiguating uncommon entities (e.g. when"哈姆斯/Harms" refers to "Rebecca Harms" instead of the most frequent "Healing of Harms" ). In addition, when an uncommon query includes a nick name such as "何伯/He Uncle", a CLEL system must analyze the whole document to find useful context features. For example, for the following two entities with the same name"何伯/He Uncle" , which are in the in the same city "Hong Kong", we will need to discover that one query refers to "a man with surname He", while the other refers to "He Yingjie".

> **document 1**: "其中,81岁姓何老翁昨趁假期,与友一行9人在大屿山东涌翔东路出发行山,至下午2时56分,一行人途至莲花山山顶附近,何伯不慎失足跌倒,跌伤头部流血,幸受伤仍清醒,由同行报警。/Among them, **the 81 years old**

**man with last name He**, ..., ..., **He Uncle** fell down..."

**document 2**: "有位何伯,在7月27日香港演艺界举行的忘我大汇演上捐出了3400万港元,不露面,不扬名。此人是香港烟草之大股东、良友基金创办人何英杰。/there is a person named **He Uncle**, donated .... This person is **He Yingjie**, who is the founder of ...".

***Discourse reasoning***: A few queries require cross-sentence shallow reasoning to resolve. For example, in a document including a query "三沙镇/Sansha Town", most sentences only mention explicit contexts about "三沙港/Sansha Port" (e.g. it's located in "Fujian Province") , so we need to propagate these contexts to disambiguate the query, based on the assumption that "Sansha Port" is likely to be located in "Sansha Town".

***Background knowledge***: About 2% queries require background knowledge to translate and disambiguate. For example, "梁泰龙" should be transalted into a Korean name "Jonathan Leong" (and thus refer to the Korean) or a Chinese name "Liang Tailong", depending on his nationality mentioned explicitly or implicitly in the source documents.

***No-clue entities***: There are also some very challenging queries in the evaluation set. Most of them are some entities which are not involved in any central topics of the source documents, therefore they are not linked to any KB entries and also there are no explicit contexts we can find to cluster them. For example, some news reporters such as "张小平/Xiaoping Zhang" and some ancient people such as "包拯/Bao Zheng" were selected as queries.

#### 6.4.2 Cross-lingual NIL Entity Clustering

NIL clustering was particularly difficult in this CLEL evaluation. Topic modeling helped improve clustering NIL queries in most cases, providing evidence superior to what could be provided using local lexical features. However, for some queries with common names (e.g. "Li Na", "Wallace"), it's still possible for them to refer to different entities even if the source documents involve the same topic. For example, two source documents included two different entities with the same name "莫里西/Molish" and similar topics about "analysis of life length/death".

Another other significant challenge is when a person entity has different titles during different times. For example, we need to incorporate temporal slot filling in order to group "众议院情报委员会主席高斯/Gauss, the chairman of the Intelligence Committee " and "美国中央情报局局长高斯/The U.S. CIA director Gauss" into the same entity cluster, or to group "中国著名作家王蒙/The famous Chinese writer Wang Meng" and "前文化部长王蒙/Wang Meng, the former head of the Culture Department" into the same entity cluster.

## 7 Conclusions and Future Work

In KBP2011, temporal slot filling was our main research focus. Therefore we have spent very limited efforts at developing the cross-lingual entity linking system, whose evaluation took place during the Irene hurricane week. However, we achieved very promising results in both mono-lingual and cross-lingual tracks. In addition, we were able to explore some novel approaches including cross-lingual information networks modeling and topic modeling for this new cross-lingual entity linking task. In our ongoing work, we are developing name-aware machine translation system by tightly integrating name translation into machine translation training and decoding phases. In the future, we will add more global evidence into information networks, such as temporal document distributions. We are also interested in extending the name co-occurrence matrix into three languages (e.g. the triangle links among English, Chinese and Japanese).

## Acknowledgments

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF- 09-2-0053 (NS-CTA), the U.S. NSF CAREER Award under Grant IIS-0953149, the U.S. NSF EAGER Award under Grant No. IIS-1144111, the U.S. DARPA Broad Operational Language Translations program and PSC-CUNY Research Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official

policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## References


Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding Similar Sentences Across Multiple Languages in Wikipedia. *Proc. EACL2006*.

Eytan Adar, Michael Skinner, and Daniel S. Weld. 2009. Information Arbitrage across Multi-lingual Wikipedia. In *Second ACM International Conference on Web Search and Data Mining (WSDM'09), Barcelona, Spain, February 2009*, February.

Javier Artiles, Andrew Borthwick, Julio Gonzalo, Satoshi Sekine and Enrique Enrique Amigo. 2010. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Task. Proc. Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2010).

Sisay Fissaha Adafre and Maarten de Rijke. 2011. Language-Independent Identification of Parallel Sentences Using Wikipedia. *Proc. WWW2011*.

Gosse Bouma, Sergio Duarte, and Zahurul Islam. 2009. Cross-lingual alignment and complettion of wikipedia templates. In *The Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*.

Gerard de Melo and Gerhard Weikum. 2010. Untangling the cross-lingual link structure of wikipedia. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden*.

Amit Bagga and Breck Baldwin. 1998 Algorithms for Scoring Coreference Chains. *Proc. Resources and Evaluation Workshop on Linguistics Coreference*.

Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai and H. Li. 2007. Learning to rank: from pairwise approach to listwise approach In *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pages 129-136.

Zheng Chen, Suzanne Tamang, Adam Lee, Xiang Li, Marissa Passantino and Heng Ji. 2010. Top-down and Bottom-up: A Combined Approach to Slot Filling. *Lecture Notes in Computer Science*, Volume 6458, pp. 300-309.

Z. Chen, S. Tamang, A. Lee, X. Li, W.-P. Lin, M. Snover, J. Artiles, M. Passantino and H. Ji. 2010. CUNYBLENDER TAC-KBP2010 Entity Linking and Slot Filling System Description. In *Proceedings of Text Analytics Conference (TAC2010)*.

Zheng Chen and Heng Ji. 2011. Collaborative Ranking: A Case Study on Entity Linking. *Proc. EMNLP2011*.

Hongbo Deng, Jiawei Han, Bo Zhao, Yintao Yu and Cindy Xide Lin. 2011. Probabilistic Topic Models with Biased Propagation on Heterogeneous Information Networks. In *KDD submission*, 2011.

M. Dredze, P. McNamee, D. Rao, A. Gerber and T. Finin. 2010. Entity Disambiguation for Knowledge Base Population. *Proc. COLING 2010*.

Maike Erdmann, Kotaro Nakayama, Takahiro Hara and Shojiro Nishio. 2009. Improving the Extraction of Bilingual Terminology from Wikipedia. *ACM Transactions on Multimedia Computing Communications and Applications*.

Angela Fahrni and Michael Strube. 2011. HITS' Cross-lingual Entity Linking System at TAC2011: One Model for All Languages. *Proc. TAC2011*.

Elena Filatova. 2009. Multilingual Wikipedia, Summarization, and Information Trustworthiness. *Proc. SIGIR2009 Workshop on Information Access in a Multilingual World*.

William A. Gale, Kenneth W. Church and David Yarowsky. 1992. One Sense Per Discourse. *Proc. DARPA Speech and Natural Language Workshop*.

Zellig Harris. 1954. Distributional Structure. *Word* , 10(23):146-162.

Ahmed Hassan, Haytham Fahmy and Hany Hassan. 2007. Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora. *Proc. RANLP2007*.

Heng Ji, David Westbrook and Ralph Grishman. 2005. Using Semantic Relations to Refine Coreference Decisions. *Proc. EMNLP2005*.

Heng Ji and Ralph Grishman. 2006. Data Selection in Semi-supervised Learning for Name Tagging. *Proc. COLING/ACL 06 Workshop on Information Extraction Beyond Document*.

Heng Ji, Ralph Grishman, Dayne Freitag, Matthias Blume, John Wang, Shahram Khadivi, Richard Zens and Hermann Ney. 2009. Name Translation for Distillation. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*.

Heng Ji. 2009. Mining name translations from comparable corpora by creating bilingual information networks. In *ACL-IJCNLP 2009 workshop on Building and Using Comparable Corpora (BUCC 2009): from Parallel to Non-parallel Corpora*.

Heng Ji, Ralph Grishman and Hoa Trang Dang. 2011. An Overview of the TAC2011 Knowledge Base Population Track. *Proc. Text Analytics Conference (TAC2011)*.

T. Joachims. 1999. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

T. Joachims. 2006. Training Linear SVMs in Linear Time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Zornitsa Kozareva and Sujith Ravi. 2011. Unsupervised Name Ambiguity Resolution Using A Generative Model. *Proc. EMNLP2011 Workshop on Unsupervised Learning in NLP*.

Qi Li, Sam Anzaroot, Wen-Pin Lin, Xiang Li and Heng Ji. 2011. Joint Inference for Cross-document Information Extraction. *Proc. 20th ACM Conference on Information and Knowledge Management (CIKM2011)*.

Wen-Pin Lin, Matthew Snover and Heng Ji. 2011. Unsupervised Language-Independent Name Translation Mining from Wikipedia Infoboxes. *Proc. EMNLP2011 Workshop on Unsupervised Learning for NLP.*.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas W. Oard and David Doermann. 2011. Cross-Language Entity Linking. *Proc. IJCNLP2011*.

Paul McNamee, James Mayfield, Douglas W. Oard, Tan Xu, Ke Wu, Veselin Stoyanov and David Doermann. 2011. Cross-Language Entity Linking in Maryland during a Hurricane. *Proc. TAC2011*.

Sean Monahan, John Lehmann, Timothy Nyberg, Jesse Plymale and Arnold Jung. 2011. Cross-Lingual Cross-Document Coreference with Entity Linking. *Proc. TAC2011*.

Lev Ratinov, Dan Roth, Doug Downey and Mike Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. *Proc. ACL2011*.

Lev Ratinov. 2011. GLOW TAC-KBP2010 Entity Linking System. *Proc. TAC2011*.

Alexander E. Richman and Patrick Schone. 2008. Mining Wiki Resources for Multilingual Named Entity Recognition. *Proc. ACL2008*.

Min Wan and Zhensheng Luo. 2003. Study on Topic Segmentation Method in Automatic Abstracting System. In *Proc. Natural Language Processing and Knowledge Engineering*.

M. Yoshida, M. Ikeda, S. Ono, I. Sato, and H. Nakagawa. 2010. Person name disambiguation by boostrapping. In *SIGIR*.

Gae-won You, Seung-won Hwang, Young-In Song, Long Jiang, Zaiqing Nie. 2010. *Mining Name Translations from Entity Graph Mappings.* In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 430-439.

Jing Zheng, Necip Fazil Ayan, Wen Wang and David Burkett. 2009. Using Syntax in Large-scale Audio Document Translation. *Proc. Interspeech*.

Zhicheng Zheng, Fangtao Li, Minlie Huang and Xiaoyan Zhu. 2010. Learning to Link Entities with Knowledge Base. *Proc. NAACL2010*.