Analysis and Refinement of Cross-lingual Entity Linking

Taylor Cassidy[1], Heng Ji[1], Hongbo Deng[2], Jing Zheng[3], Jiawei Han[2]

[1]Computer Science Department and Linguistics Department
Queens College and Graduate Center, City University of New York, New York, NY, USA
{taylorcassidy64,hengji}@gmail.com
[2] Computer Science Department
University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA
{hbdeng,hanj}@illinois.edu
[3] SRI International, Menlo Park, CA, USA
zj@speech.sri.com

Abstract. In this paper we propose two novel approaches to enhance cross-lingual entity linking (CLEL). One is based on cross-lingual information networks, aligned based on monolingual information extraction, and the other uses topic modeling to ensure global consistency. We enhance a strong baseline system derived from a combination of state-of-the-art machine translation and monolingual entity linking to achieve 11.2% improvement in B-Cubed+ F-measure. Our system achieved highly competitive results in the NIST Text Analysis Conference (TAC) Knowledge Base Population (KBP2011) evaluation. We also provide detailed qualitative and quantitative analysis on the contributions of each approach and the remaining challenges.

1 Introduction

The main goal of the Knowledge Base Population (KBP) track at the Text Analysis Conference (TAC) is to gather information about an entity that is scattered among the documents in a large collection, and then use the extracted information to populate an existing English knowledge base (KB). Previous KBP tasks were limited to mono-lingual processing; however, for certain entities, a lot of information is only available in documents written in a foreign language. To address this issue KBP2011 [12] included a new cross-lingual entity linking (CLEL) task in which queries from both Chinese and English are clustered, and whether each cluster corresponds to a KB entry is determined. The English KB used for this task is a subset of Wikipedia. Each KB entry consists of the title, infobox, and full text of a Wikipedia article.

There are two conventional ways to extend mono-lingual entity linking systems to the cross-lingual setting: (1) Apply a Source Language (SL) mono-lingual entity linking (MLEL) system to link SL entity mentions to SL KB entries, and then link the SL KB entry to the corresponding Target Language (TL) KB entry; (2) Apply machine translation (MT) to translate the SL document into the TL, and then apply a TL MLEL system to link entity mentions in the translated document to TL KB entries. These pipelines essentially convert CLEL to MLEL. However, these approaches are limited by their core components: approach (1) relies heavily on the existence of an

SL KB whose size is comparable to the TL KB, as well as the existence of a reliable mapping between the two KB. Thus, this approach is not easily adaptable to low-density languages. Approach (2) relies on MT output, and as such it will suffer from translation errors, particularly those involving named entities (NE).

In order to both enhance the portability and reduce the cost of cross-lingual entity linking, we have developed a novel re-ranking approach which requires neither MT nor a source language KB. Our research hypothesis is that the query entity mentions ("queries" from here on) can be disambiguated based on their "collaborators" or "supporters"; namely, those entities which co-occur with, or are semantically 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 affiliations with co-occurring entities: "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 mine and disambiguate entities.

In our second enhancement we adapt the distributional [10] and "One Sense Per Discourse" [9] hypotheses to our task: we hypothesize that queries sharing topically-related contexts tend to link to the same KB entry, and we consider the KB entry denoted by a query to be its *sense*, while treating a set of documents discussing the same topic as a discourse. Topic modeling provides a natural and effective way to model the contextual profile of each query [15]. Identical or highly similar entity mentions in a single coherent latent topic tend to express the same sense, and thus should be linked to the same KB entry. For example, a query "*Li Na*" is associated with a sports topic cluster represented by, {*tennis, player, Russia, final, single, gain, half, male, ...*}, and an identical query, "*Li Na*", is associated with a politics topic cluster represented by {*Pakistan, relation, express, vice president, country, Prime minister, ...*}; thus, they probably refer to two different entities. We also observe that entities play a significant role in distinguishing topics. Based on these observations, our second CLEL enhancement employs a topic modeling method with a biased propagation (in which both entities and documents are assigned to topic clusters), to the Chinese source documents. In doing so, we implicitly assume consistency of results among entities in each topic cluster based on our second hypothesis: "one entity per topic cluster".

2 Related Work

Although CLEL is a new task in the KBP track, similar efforts have been published in recent papers [18], but with evaluation settings and query selection criteria that are quite different (precision and recall are calculated on a by-token, as opposed to a by-cluster basis; their queries are selected automatically by propagating NE output from English source documents to parallel documents in other languages via automatic word alignment, while KBP CLEL queries were manually selected to cover many ambiguous entities and name variants). Almost all CLEL systems participating in the KBP2011 track (e.g. [19, 21, 7]) 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 [8], find similar sentences [2], or extract bi-lingual terminology [6]. Some recent name pair mining work has been based on aligning Multi-lingual Wikipedia Pages [22], Infoboxes [17], and web co-occurrence based networks [23]. To the best of our knowledge, our re-ranking approach is the first work to apply unsupervised cross-lingual name pair mining to enhance entity linking. In addition, [20] used unambiguous concept mentions to bootstrap the linking of more ambiguous mentions based on Wikipedia link structure, but do not incorporate more fine-grained relationships between entities.

[15] applied topic modeling for the Web People Search task [1]. We extended this idea from the mono-lingual to the cross-lingual setting. The topic modeling method we use treats "entity mention" and "document" as node types in a heterogeneous network, where the topic distribution for a document is based on both its overall content as well as the topic distributions of the entity mentions it contains, which are completely derived from the topic distributions of the documents that contain them.

3 Task Definition

We are addressing the CLEL task of the NIST TAC KBP2011 evaluations [12]. Given a Chinese or English query that consists of a name string - which may refer to a person (PER), organization (ORG) or geo-political entity (GPE, a location with a government) - and a source document ID, a system is required to provide an English KB entry ID to which the name string refers. Queries for which no such KB entry exists are classified as NIL. Co-referring queries must be clustered (including those classified as NIL), and each cluster must be assigned a unique ID. KBP2011 used a modified B-Cubed metric (B-Cubed+) [12] to evaluate entity clusters.

4 System Overview

Figure 1 depicts the overall pipeline of our cross-lingual entity linking system. We have developed a baseline approach consisting of state-of-the-art name translation, machine translation, and mono-lingual entity linking. The baseline system first translates a Chinese query and its associated document into English, and then applies English MLEL to link the translated query, given the translated document as context, to the English KB.

We apply a Chinese name coreference resolution system [14] to each source document in order to get name variants for a given query. Then we apply various name translation approaches including name transliteration, name mining from comparable corpora and information extraction based name re-ranking, as described in [13].

We then apply a hierarchical phrase-based machine translation system as described in [24] to translate Chinese documents to English. 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 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 the SCFG. A CKY-style decoder was used to solve the search problem.

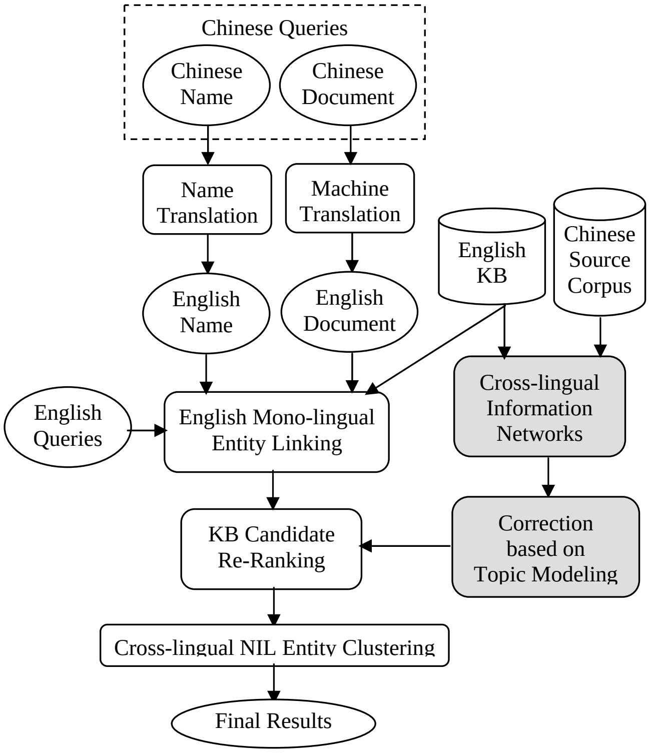

Fig. 1: Cross-lingual Entity Linking System Overview

After translating the queries and documents into English, we apply a high-performing English MLEL system [4] to link each query. This system includes two unsupervised rankers based on entity profile and document similarity features, and three supervised rankers(Maximum Entropy, Support Vector Machines and ListNet) based on surface features, document features, and profiling features (entity attributes that are extracted by a slot filling system).

We then developed a novel joint approach for translating and disambiguating entities through cross-lingual information network construction (section 5). From the information networks we can extract a context similarity score for each query, KB entry pair. This context similarity score is then combined with the MLEL scores (i.e. the results of applying MLEL to MT output) based on weights optimized from the KBP2011 training data set. In addition, we applied a new entity-driven topic modeling approach with biased propagation [5], to ensure the consistency of entity linking results within each topic cluster (section 6).

Finally, we implemented a simple substring matching based approach to NIL clustering. For Chinese queries, we apply a within-document Chinese coreference resolution system and some abbreviation gazetteers to expand each query (e.g. “魁北克/Quebec”), yielding a cluster of coreferential names (“魁北克, 魁北克集团/Quebec, Quebec group”) for greedy matching.

5 Information Networks for CLEL

5.1 Motivations

As we pointed out in the introduction, both basic approaches to CLEL present problems. In addition, there are some characteristics specific to Chinese that can be addressed by adding 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.

Some entity mentions are more difficult to translate than others due to referential ambiguity. However, entity mentions can be disambiguated based on co-occurring entity mentions that are less ambiguous. When a human determines the referent of a query, one strategy is to first construct its *profile* from the text. This might include its title, origin, employer or social affiliations in the case of a person, or location and capital city in the case of a country, etc. To the extent that the corresponding relationships between queries and co-occurring entity mentions are significant, we expect them to be reflected in the KB structure (as relations between the target KB entry and other KB entries); thus, a query can be disambiguated by comparing a profile extracted from its surrounding text to profiles of candidate target KB entries, given in terms of the Wikipedia link structure, info boxes, and relations expressed in the KB entry’s text. This method is reliable to the extent that the profile entity mentions are unambiguously associated with their own KB entries, and relations expressed in text are in fact expressed in the KB. If these conditions are met, unambiguous entity mentions can bootstrap disambiguation of more difficult cases in their profiles. Inspired by this intuition, we propose a novel approach to jointly mine entity translations and disambiguate entities based on entity profile comparison.

We exploit a representation called “Information Networks” [16] to model the profile for each query. This approach is effective for disambiguating queries with common organization names or person names to the extent that the query’s profile is readily inferred from the context, and the profiles of competing target KB entries for a given query don’t overlap. For example, if a query such as “*Supreme Court*”, “*LDP (Liberty and Democracy Party)*”, or “*Newcastle University*” has a country entity mention in its profile, it is fairly easy to disambiguate after comparing query profiles with candidate KB entry profiles. In practice, however, the extent to which entity profiles are explicitly presented varies. Table 1 presents the various types of contexts that may help disambiguate entities.

5.2 Information Networks Construction

For a given Chinese query, we refer to the other entity mentions in the associated source document that are associated with the query as its *neighbors*. Here, association can consist of either an automatically extracted relationship or simple co-occurrence (note

Table 1: Information Networks Examples for Entity Disambiguation

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：… 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 …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**

that co-occurrence is determined after coreference resolution). We apply a state-of-the-art bi-lingual (English and Chinese) IE system [11, 3] to extract relations and events defined in the NIST Automatic Content Extraction Program (ACE 2005) program [1]. Each IE system includes tokenization/word segmentation, part-of-speech tagging, parsing, name tagging, nominal mention tagging, entity coreference resolution, time expression extraction and normalization, relation extraction, and event extraction. Names are identified and classified using a Hidden Markov Model. Nominals are identified using a Maximum Entropy (MaxEnt)-based chunker and then semantically classified using statistics from the ACE training corpora. Entity coreference resolution, relation extraction, and event extraction are also based on MaxEnt models, incorporating diverse lexical, syntactic,

[1] `http://www.itl.nist.gov/iad/mig/tests/ace/`

semantic, and ontological knowledge. In addition, we apply a state-of-the-art slot filling system [4] to identify KBP slot values for each person or organization entity that appears as a query in the source documents. This system includes a bottom-up pattern matching pipeline and a top-down question answering (QA) pipeline.

For a given KB entry, we determine its neighbors by first applying the above extraction techniques to the associated Wikipedia article, and then by utilizing Wikipedia article link information: any two KB entries are considered neighbors if a link to one KB entry appears in the text (Wikipedia page) of the other.

5.3 Information Networks based Re-Ranking

As alluded to above, a query's neighbors may refer to the neighbors of its referent in the KB. Therefore, low baseline scores may be boosted based on the high scores of neighbor pairs. In particular, when choosing between two KB referents for a given query, we want to give more weight to the KB entry whose KB neighbors are likely to be the intended referents of the context neighbors of the query in question. The baseline system generates N-Best KB entries for each query, with a confidence value for each hypothesis. For each link type (ACE relation, ACE event, KBP attribute or co-occurrence) in the information networks of a query and a candidate KB entry, we counted the number of matched context entity pairs, and used these statistics as additional features for re-ranking. Together with the baseline confidence values, these features are sent to a supervised re-ranker based on Maximum Entropy, which was trained using the KBP2011 training data.

6 Topic Modeling for CLEL

The information networks we constructed capture each query's local (within-document) context but fail to incorporate global (cross- document) context. A document in which an entity is mentioned will normally contain only a small subset of the information that could, in principle, be used to distinguish it from other entities. One way to alleviate this problem would be to simply construct links between entity mentions irrespective of document boundaries; however, this would likely do more harm than good due to noise introduced by ambiguous names. To capture entities' global context we apply an entity-driven topic modeling framework adapted from [5].

The underlying intuition, when applied to the task at hand, is that the topic of a document is based primarily on its own explicit content, but is influenced to some extent by the topic of each entity contained therein, each of which is determined based on the topic of each document in which it appears. To incorporate both the textual information and the relationships between documents and entities, we use a biased regularization framework in which regularization terms are added to the log-likelihood topic distribution, and are subject to the constraint that the probability of an entity having a given topic is equal to the mean of the probabilities that each of its containing documents have that topic. A regularization term for a given entity type represents the difference between the probability that a document has a given topic and the mean of the probabilities associated with each entity it contains having that topic. A loss function is defined as the difference between the topic probabilities for documents and those of the entities they contain, which is minimized via generalized expectation-maximization. Finally, each document and each entity is considered a member of the topic cluster whose topic it's most strongly associated with. As the regularization parameter approaches 0 the model is reduced to standard probabilistic latent semantic analysis.

For each source document we extract its metadata, as well as English and Chinese named entities, using a bi-lingual named entity extraction system [14] which consists of a Hidden Markov Model (HMM) tagger augmented with a set of post-processing rules. 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 that they each link to the same target KB entry. Thus, two queries with the same namestring can be linked to different KB entries only if they have different topics.

7 Experiments

7.1 Data

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. We used the KBP 2011 Cross-lingual Entity Linking training data set to develop our systems, and then conducted a blind test on KBP2011 Cross-lingual Entity Linking evaluation data set. The detailed data statistics are summarized in Table 2.

Table 2: Data sets

Corpus		# Queries		
		Person	Organization	GPE
Training	English	168	253	243
	Chinese	649	407	441
Evaluation	English	183	269	303
	Chinese	641	441	399

7.2 Overall Performance

Performance on the cross-lingual entity linking task both before and after applying our enhancements are summarized in Table 3. Source language information networks and topic modeling have significantly improved the results for Chinese queries, especially for the person (PER) and geo-political (GPE) types. Performance on PER queries is significantly worse for Chinese than for English, mainly because the translation of PER names is the most challenging among the three entity types; however, our enhancements were particularly beneficial for this category in which our system acheived the highest score. On the other hand, we found that for some Chinese names, their Chinese mentions are actually less ambiguous than their English mentions because the mapping from Chinese character to pinyin is many-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 achieved in the F-measure for English queries. The overall F-measure was improved from 65.4% to 76.6%.

Table 3: Cross-lingual Entity Linking Evaluation Results (%)

Entity Type	Chinese						English					
	Baseline			Enhanced			Baseline			Enhanced		
	P	R	F	P	R	F	P	R	F	P	R	F
PER	37.5	42.0	39.6	**65.1**	**73.1**	**68.9**	74.7	73.3	74.0	**76.3**	**76.1**	**76.2**
GPE	73.5	74.9	74.2	**83.3**	**83.9**	**83.6**	82.1	81.2	81.6	**82.1**	**82.3**	**82.2**
ORG	68.3	83.9	75.3	**69.7**	**85.7**	**76.8**	77.5	81.0	79.2	**80.3**	**84.9**	**82.5**
ALL	56.3	63.4	59.6	**71.0**	**79.8**	**75.1**	78.4	79.0	78.7	**79.9**	**81.7**	**80.8**

7.3 Discussion

In Figure 2 we present the distribution of 1,481 Chinese queries in the KBP2011 CLEL evaluation corpus in terms of the various techniques needed to disambiguate them as well as their difficulty levels. The percentage numbers are approximate because some queries may rely on a combination of multiple strategies.

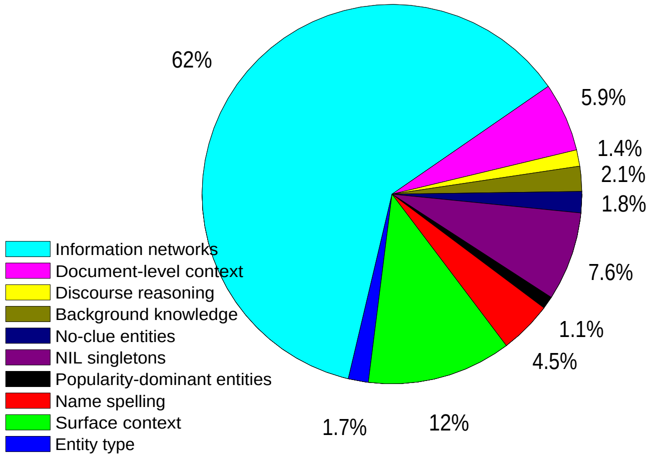

Fig. 2: Distribution of CLEL queries according to difficulty levels

– ***(1) Easy Queries***

NIL singletons: About 7.6% of the queries are singleton entities (e.g. "中绿集团/Zhonglv Group", "丰华中文学校/Fenghua Chinese School"), in that they only appear in one query and do not have a corresponding KB entry.

Name spelling: 4.5% of the queries can be disambiguated because their full names 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 first names.

– ***(2) Queries Linked by Baseline Methods***

Surface context: 12% of the queries can be disambiguated based on 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" later in the document in order to link it to the correct KB entry.

Popularity-dominant entities: A few (only 1.1%) of the queries are popular entities, such as "路透社/ Reuters"; such queries can be correctly linked using popularity features alone.

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.

– ***(3) Queries Linked by Enhanced Methods***

Information networks: As we have discussed in Table 1, many entities (62% of the evaluation queries) can be linked based on contextual information networks. Such information is particularly effective for those entities that may be located in or affiliated with 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 highly ambiguous geo-political names if we can identify higher-level context entities that subsume them. For example, there are many different KB candidates for the 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.

Topic Modeling: 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", as opposed to "Healing of Harms" which is more likely on a relative frequency basis). 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" based on their associated topic distributions.

document 1: "其中,81岁姓何老翁昨趁假期,...何伯不慎失足跌倒.../Among them, **the 81 year old man with last name He**, ..., ..., **He Uncle** fell down..."

document 2: "有位何伯,...此人是...创办人何英杰。/there is a person named **He Uncle**, ... This person is **He Yingjie**, who is the founder of ...".

- ***(4) Remaining Difficult Queries***

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”, and that it’s located in “Fujian Province”. These contexts must be combined, under the assumption that “Sansha Port” is likely to be located in “Sansha Town”, in order to disambiguate the query,

Background knowledge: About 2% queries require background knowledge to translate and disambiguate. For example, if “梁泰龙” refers to a Korean person then the English translation is “Jonathan Leong”, but if the name refers to a Chinese person the translation should be “Liang Tailong”. Thus, the correct translation of a persons name may depend on his nationality, which might be revealed explicitly or implicitly in the source documents.

No-clue entities: Some challenging queries are not involved in any central topics of the source documents, and as a result systems tend not to link them to any KB entries; in addition, their mentions have no significant context in common. 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.

8 Conclusions and Future Work

In this paper we described a high-performing cross-lingual entity linking system. This system made use of some novel approaches - aligning Chinese source and English KB based information networks and entity-driven topic modeling - to enhance a strong baseline pipeline previously used for this task. In the future, we will add more global evidence into information networks, such as temporal document distributions. We are also interested in incorporating additional source languages (e.g. the triangle links among English, Chinese and Japanese).

Acknowledgements

The work was supported in part by the U.S. National Science Foundation grants IIS-0953149, IIS-1144111, IIS-0905215, CNS-0931975, the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), the U.S. DARPA Broad Operational Language Translations program, the U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265. 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

1. Artiles, J., Borthwick, A., Gonzalo, J., Sekine, S., Amigo, E.E.: WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Task. In: Proc. CLEF 2010 (2010)

2. Adafre, S.F., de Rijke, M.: Language-Independent Identification of Parallel Sentences Using Wikipedia. In: Proc. WWW2011 (2011)
3. Chen, Z., Ji, H.: Language Specific Issue and Feature Exploration in Chinese Event Extraction. In: Proc. HLT-NAACL 2009 (2009)
4. Chen, Z., Ji, H.: Collaborative Ranking: A Case Study on Entity Linking. In: Proc. EMNLP2011 (2011)
5. Deng, H., Han, J., Zhao, B., Yu, Y., Lin, C.X.: Probabilistic Topic Models with Biased Propagation on Heterogeneous Information Networks. In: Proc. KDD 2011 (2011)
6. Erdmann, M., Nakayama, K., Hara, T., Nishio, S.: Improving the Extraction of Bilingual Terminology from Wikipedia. ACM Transactions on Multimedia Computing Communications and Applications (2009)
7. Fahrni, A., Strube, M.: HITS' Cross-lingual Entity Linking System at TAC2011: One Model for All Languages. In: Proc. TAC2011 (2011)
8. Filatova, E.: Multilingual Wikipedia, Summarization, and Information Trustworthiness. In: Proc. SIGIR2009 Workshop on Information Access in a Multilingual World (2009)
9. Gale, W.A., Church, K.W., Yarowsky, D.: One Sense Per Discourse. In: Proc. DARPA Speech and Natural Language Workshop (1992)
10. Harris, Z.: Distributional Structure. Word (1954)
11. Ji, H., Grishman, R.: Refining Event Extraction through Cross-Document Inference. In: Proc. of ACL-08: HLT. pp. 254–262 (2008)
12. Ji, H., Grishman, R., Dang, H.T.: An Overview of the TAC2011 Knowledge Base Population Track. In: Proc. Text Analytics Conference (TAC2011) (2011)
13. Ji, H., Grishman, R., Freitag, D., Blume, M., Wang, J., Khadivi, S., Zens, R., Ney, H.: Name Translation for Distillation. In: Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation (2009)
14. Ji, H., Westbrook, D., Grishman, R.: Using Semantic Relations to Refine Coreference Decisions. In: Proc. EMNLP2005 (2005)
15. Kozareva, Z., Ravi, S.: Unsupervised Name Ambiguity Resolution Using A Generative Model. In: Proc. EMNLP2011 Workshop on Unsupervised Learning in NLP (2011)
16. Li, Q., Anzaroot, S., Lin, W.P., Li, X., Ji, H.: Joint Inference for Cross-document Information Extraction. In: Proc. CIKM2011 (2011)
17. Lin, W.P., Snover, M., Ji, H.: Unsupervised Language-Independent Name Translation Mining from Wikipedia Infoboxes. In: Proc. EMNLP2011 Workshop on Unsupervised Learning for NLP (2011)
18. McNamee, P., Mayfield, J., Lawrie, D., Oard, D.W., Doermann, D.: Cross-Language Entity Linking. In: Proc. IJCNLP2011 (2011)
19. McNamee, P., Mayfield, J., Oard, D.W., Xu, T., Wu, K., Stoyanov, V., Doermann, D.: Cross-Language Entity Linking in Maryland during a Hurricane. In: Proc. TAC2011 (2011)
20. Milne, D., Witten, I. H.: Learning to Link with Wikipedia. In: Proc. CIKM 2008 (2008)
21. Monahan, S., Lehmann, J., Nyberg, T., Plymale, J., Jung, A.: Cross-Lingual Cross-Document Coreference with Entity Linking. In: Proc. TAC2011 (2011)
22. Richman, A.E., Schone, P.: Mining Wiki Resources for Multilingual Named Entity Recognition. In: Proc. ACL2008 (2008)
23. You, G., Hwang, S., Song, Y., Jiang, L., Nie, Z.: Mining Name Translations from Entity Graph Mappings. In: Proc. EMNLP2010 (2003)
24. Zheng, J., Ayan, N.F., Wang, W., Burkett, D.: Using Syntax in Large-scale Audio Document Translation. In: Proc. Interspeech (2009)