Cross-Language Entity Linking in Maryland during a Hurricane

Paul McNamee and **James Mayfield**
HLTCOE & Applied Physics Laboratory
Johns Hopkins University
{paul.mcnamee,james.mayfield}@jhuapl.edu

Douglas W. Oard
iSchool & UMIACS
University of Maryland, College Park
oard@umd.edu

Tan Xu
iSchool
University of Maryland, College Park
tanx@umd.edu

Wu, Ke
Computer Science Department
University of Maryland, College Park
wuke@cs.umd.edu

Veselin Stoyanov
HLTCOE
Johns Hopkins University
ves@cs.jhu.edu

David Doermann
UMIACS
University of Maryland, College Park
doermann@umiacs.umd.edu

Abstract

Our team from the JHU HLTCOE and the University of Maryland submitted runs for all three variants of the TAC-KBP entity linking task. For the monolingual tasks, we essentially mirrored our HLTCOE TAC-KBP 2010 submission, making only modest changes to accommodate differences in 2011, namely the requirement to cluster NIL responses, and the change in evaluation measure. However, our work on the cross-lingual task was significantly more involved, requiring development of robust, multi-phased transliteration software, use of techniques in cross-language information retrieval, and reliance on a Chinese-to-English statistical machine translation system. In this paper we describe our work for the 2011 evaluation and the results we obtained.

1 Introduction

The JHU HLTCOE has participated in the TAC Knowledge Base Population exercise since its inception in 2009. In the first year we focused our efforts on the entity linking task and also made a submission for the slot filling task. In 2010 we only submitted results for the entity linking task, and we made a concerted effort to streamline our system, making it both conceptually simpler, and more computationally efficient.

For this year's effort we again submitted results to the entity linking task. Our monolingual English runs were largely based on our 2010 system, with some adaptation to account for the new requirement to cluster NIL responses. However, our cross-lingual results required significant efforts in Chinese-to-English transliteration, cross-language information retrieval, and statistical machine translation.

In addition to technical challenges presented by the task, we also had to cope with the effects of Hurricane Irene (see Figure 1). Irene slammed into the mid-Atlantic area during the evaluation period, and our facility lost electrical power for approximately 48 hours. The local utility, Baltimore Gas & Electric (BGE), had to restore power to 822,000 Maryland customers. The HLTCOE was in the last 10% of customers to have power restored.

In Section 2 we highlight our monolingual approach. Section 3 presents our monolingual results. In Section 4 we describe the customizations that were required to address the cross-language problem. Section 5 presents our results on the Chinese-English cross-language task.

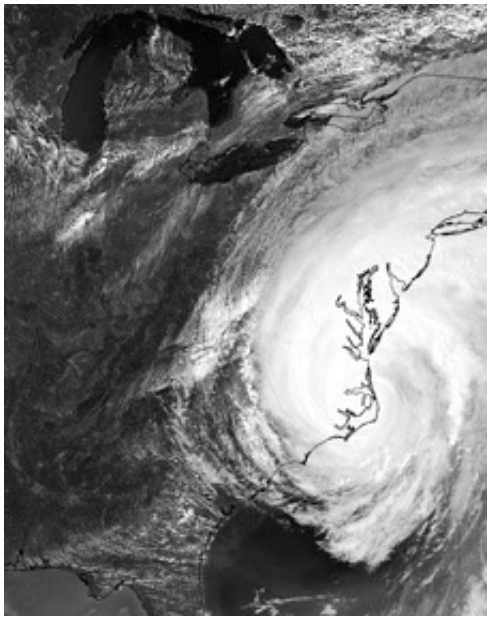

Figure 1: Satellite image of Hurricane Irene making landfall on August 27th (NASA/Goddard). Irene was classified as a Category 3 major hurricane, responsible for 55 deaths in the eastern United States and an estimated $10 billion in damage.

2 Monolingual Approach

Our approach to entity linking breaks the problem down into three main parts: *candidate identification*, *candidate ranking*, and *NIL clustering*. Candidate identification quickly identifies a small set of KB nodes that with high probability contain the correct answer, if it is present. Candidate ranking then considers each candidate in greater detail, producing a ranked list. The top candidate, which may be NIL (*i.e.,* absence from the KB) is selected. Finally, new to the 2011 task is a requirement to do cross-document coreference clustering for all queries deemed to be absent from the KB. We give a description of each of these steps in this section; more complete details of our English entity linking approach, including descriptions of all of the features used and performance on the TAC-KBP datasets can be found in (McNamee, 2010).

2.1 Candidate Identification

As a KB may contain a large number of entries, we prefer to avoid brute force comparisons between the query and all KB entities. To identify the entries that might reasonably correspond to the input named entity, we rely on a set of fast name matching techniques. In the past we have found that it is possible to achieve high recall without resorting to contextual features. We create indexes for the names in the KB to support fast lookup of potential matches. The specific techniques that we use include:

- Exact match of query and candidate names
- Acronym matching
- Known alias or nickname lookup
- Number of character 4-grams in common between query and candidate
- Sum of IDF-weighted words in common between query and candidate[1]

These methods are similar to methods used in the database community, sometimes known as *blocking* (Whang et al., 2009) or *canopies* (McCallum et al., 2000). In tests on the TAC-KBP 2009 test collection, this approach achieved 97.1% recall. For only 2.9% of the queries, the proper KB referent for the query was not one of the candidates. These cases were particularly challenging because they involved ambiguous organization names or obscure personal nicknames. However, our recall was lower in 2011 than in past years and we only considered correct responses for 2094 of 2250 (93.1%) of evaluation queries. Of these 156 non-recoverable errors, 86 were from non-newswire, and only 19 of the 156 contained a space character (*i.e.,* contained more than 1 name fragment). In hindsight, we should have paid more attention to single name queries, which were not a large fraction of previous evaluation sets.

[1]Inverse document frequency weights enable us to effectively match, for example, Q: Mary Elizabeth Surratt and KB: Mary Surratt, since *Surratt* is a highly discriminating term even though *Mary* is not.

2.2 Candidate Ranking

The second phase in our system is to score each viable candidate using supervised machine learning. We used a learning-to-rank formulation and made use of the SVM$^{\text{rank}}$ tool to train a model for ranking candidates (Joachims, 2002). We used a linear kernel and set the slack parameter C to be 0.01 times the number of training examples. The cost function used to optimize the learning is based on the number of steps required to elevate the correct candidate to the topmost rank.

In our system absence from the knowledge base is treated as a distinct ranked candidate, the so-called NIL candidate. NIL prediction is integrated into the process by including features that are indicative of no other candidate being correct. Considering absence as a ranked candidate eliminates the need to select a threshold below which NIL will be returned.

The classes of feature functions we use include:

- Name matching features between the query name (Q_{name}) and KB candidate (KB_{name})
- Text comparisons between the query document (Q_{doc}) and the text associated with the KB candidate
- Relation features, chiefly evidence from relations in the KB being evidenced in the Q_{doc}
- Co-occurring entities, detected by running named entity recognition (NER) on the Q_{doc} and finding matching names in the candidate's KB entry
- Features pertaining to the entity type of the KB candidate
- Indications that no candidate is correct and that NIL is therefore the appropriate response

2.3 NIL Clustering

Our approach to NIL clustering employs the pairwise model (Bagga and Baldwin, 1998) and reuses features utilized for entity linking. We train a pairwise classifier to label pairs of candidates as coreferent or not coreferent. The classifier uses the same two-stage approach and the same features used for entity linking. We process each candidate sequentially by first identifying a subset of previously seen candidates that are potential matches. We then build a feature vector for each named entity in this subset as compared to the NIL entity that we want to cluster. Finally, we use our classifier to score each such named entity pair and put the current named entity in the same cluster as the counterpart with which it scored highest. If the highest scoring pair is under a pre-defined threshold, we put the named entity in a new cluster.

This step also employs a linear SVM, but since the task is classification, we rely on SVM$^{\text{light}}$ (Joachims, 2002) instead of SVM$^{\text{rank}}$. We tuned the C parameter on some held out training data, settling on a value of 0.3.

We also ran a baseline condition that combined two NIL-predicted queries if their names matched exactly. This exact match baseline worked surprisingly well.

Name	Source	Total	PER	ORG	GPE
coe09train	HLTCOE	1496	539	618	458
kbp09train	NIST/LDC	119	47	66	6
kbp09eval	NIST/LDC	3904	627	2710	567
kbp10train	NIST/LDC	747	182	189	376
kbp10eval	NIST/LDC	2250	751	750	749
kbp11train	NIST/LDC	665	168	219	278
		9181	2267	4486	2428

Table 1: Sources of English training data.

3 Monolingual Results

We trained our English entity linking models using 9181 exemplars from internally-produced sources, as well as the KBP 2009-2010 training and evaluation data, plus the 2011 training data. Our sources are listed in Table 1.

As our focus this year was on the cross-language task, we used our 2010 entity linking system for the English tasks. We submitted five English runs; three for the condition that allowed use of KB article text, and two for the no-KB-text condition. None of our runs made any use of live Web access.

Run	P@1	B^3P	B^3R	B^3F_1
hltcoe1	0.772	0.730	0.750	0.740
hltcoe2	0.772	0.724	0.748	0.736
hltcoe3	0.728	0.681	0.701	0.691

Table 2: Scores from NIST. Scores for English entity-linking task.

3.1 Allowed to use KB article text

The three runs using KB article text can be distinguished by two factors: (a) use of statistical vs. exact-match NIL clustering, and (b) use of an *augmented* knowledge base. The augmented KB was created by including Wikipedia articles from a July 2011 snapshot of English Wikipedia. If our entity linker predicted that one of the Wikipedia articles which is not contained in the official TAC KBP KB is the correct entry, then we would return NIL as our response. This technique was used by good effect by the IIIT Hyderabad team in the 2010 evaluation (Varma et al., 2010).

The three runs are:

- hltcoe1: exact match NIL clustering with regular KB
- hltcoe2: statistical NIL clustering with regular KB
- hltcoe3: statistical NIL clustering with augmented KB

Summary results for the three conditions are given in Table 2. The regular KB appears to have been more successful than the augmented version, which was a surprise. The statistical NIL clustering performed just a little below the exact-match baseline.

According to summary information provided by NIST, the best reported B^3F_1 score for webless runs on this task is 0.846, with a median score of 0.716. Our top score was 0.740.

3.2 Not allowed to use article text

Our two runs for the no-KB-article-text condition were:

- hltcoe1: exact match NIL clustering with regular KB

Run	P@1	B^3P	B^3R	B^3F_1
hltcoe1	0.749	0.707	0.720	0.714
hltcoe2	0.749	0.702	0.717	0.710

Table 3: Scores from NIST. Scores for English entity-linking-no-wikitext task.

- hltcoe2: statistical NIL clustering with regular KB

Summary results for these runs are given in Table 3. Consistent with the other condition, the two NIL clustering approaches performed very similarly, though the exact-match baseline again scored just slightly higher.

According to summary information provided by NIST, the best reported B^3F_1 score for webless runs on this task is 0.714, with a median score of 0.521. Our top score was 0.714, so it appears our system obtained top marks on this condition.

4 Cross-Language Approach

We have recently developed a test set for cross-language entity linking in 21 languages, and this afforded us the opportunity to test and refine our methods on other languages before attempting the KBP 2011 task (Mayfield et al., 2011; McNamee et al., 2011). To accomplish the cross-language task, we relied on name transliteration to match the Chinese entities to English names, and we used statistical machine translation and cross-language information retrieval to transform document contexts into English equivalents.

Figure 2 illustrates our approach.

4.1 Transliteration

Translating query names into the same language as used in KB is the first hurdle that our cross-lingual entity linking system needs to tackle. For the specific case of this year's task, we need to convert Chinese names into English. Machine transliteration is an important technique to accomplish this, especially when dealing with proper nouns, such as names of people, places, and organizations, where pronunciation is often preserved between corresponding name pairs in

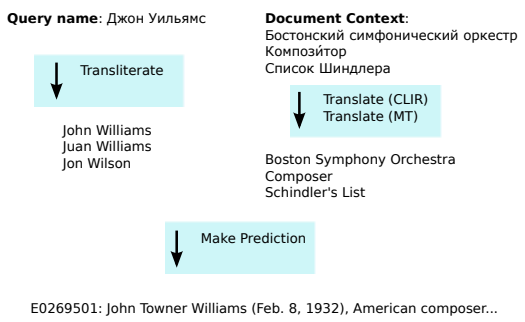

Figure 2: Transforming a non-English query to support matching against the English knowledge base using name transliteration and translated context.

different languages. Therefore, we developed a two-phased system for the purpose of Chinese-to-English transliteration.

The first phase is dictionary-lookup, which provides us computational speed and high transliteration accuracy. Since Chinese queries are extracted from newswire, and the target KB is derived from Wikipedia, we compiled a composite dictionary from the following four sources of Chinese-English name transliterations (see Table 4). The sources include two which are created by Xinhua News Agency, one from the Linguistic Data Consortium (LDC), which was originally generated by using Xinhua newspaper data, and a fourth which was created by extracting interlanguage links from Chinese Wikipedia to English Wikipedia using data available on August 20^{th}, 2011. The resulting dictionary consists of a 1-to-many mapping.

Because our dictionary of people's names includes both full name entries and also family and given names in isolation, our dictionary lookup procedure was to first look for exact full-name matches, and if that failed, to assume that the query is composed of two name parts. Therefore, we split the query into all possible combinations of two sub-strings, which must fulfill the requirement that both substrings can be found in the people's name dictionary. Then, because the cross-product of all combined name possibilities could be large we use the Google bi-gram data to select the most frequent English names given the query (Brants and Franz, 2006) when we desired a single best transliteration to work with.

For the KBP 2011 evaluation data, 77.2% of all queries can be resolved by dictionary lookup. However since the transliteration method is independent of document context, we also computed the frequency that unique query names could be found in our dictionary, which is 77.7%.

If a query name cannot be found in our dictionary, then a second phase is applied, where a generative transliteration approach is adopted. In this phase, we firstly built 17 transliteration models from entries in our people's name dictionary using software tool developed by Irvine et al. (2010). Since the software is an orthographic-based transliteration system, we use Pinyin (the official transcription system from Chinese characters to Roman alphabets) to transform each Chinese name in order to train the models.[2]

The reason that 17 models are used is because the people's name dictionary provides the origin of each name, which we assume can represent a distinguishable transliteration rule with Chinese Pinyin. For example, the name "侯赛因" (English: Hussein), when it is used in Arabic, can also be transliterated as "Hassine", or "Choesin" in Indonesia, or "Guseein" in Kazakhstan, to name just a few.

We selected 17 origin languages, which each contain at least 10,000 entries. Queries are first pinyinized, and then transliterated from pinyinized Chinese to English using these models. For our final submission, with regarding to these out-of-dictionary queries, we use standard pinyin as the "best" transliteration, and all 17 transliterations as an expansion set of additional viable transliterations. The overall procedure of our transliteration system can be illustrated by using the example in Table 5.

[2]The Pinyin algorithm we used is a Perl library available at: `http://search.cpan.org/~fayland/Lingua-Han-PinYin-0.15/lib/Lingua/Han/PinYin.pm`

Dictionary Name	Source	Size
1. Names of the world's peoples (Guo, 2007)	Xinhua News Agency	676,871
2. Place names of the world (Zhou, 2008)	Xinhua News Agency	177,322
3. Chinese<->English Name Entity Lists v 1.0 (organization, corporate) (Huang, 2005)	LDC	122,344
4. Chinese-English cross-lingual name pairs	Chinese Wikipedia	427,678

Table 4: Chinese-English Name Transliteration Dictionaries

Query	English	Dictionary	Translit	Origin
布里斯托尔	Bristol	bristol	N/A	N/A
瑟琳娜	Selina	N/A	selinna	Pinyin
			tharina	English
			corinne	French
			selino	Spanish
			selina	Italian
			selina	Russian
			serina	German
			serino	Serbia
			serina	Sweden
			srinno	Finland
			selino	Turkish
			selino	Arabic
			serina	Indian
			selina	Czech
			serino	Roman
			srinna	Hungary
			sinnaya	Japanese
			sak-rimnak	Korean

Table 5: Chinese transliteration example. 布里斯托尔 is found in the dictionary, and "Bristol" is returned. 瑟琳娜 is not in our dictionary, and standard Pinyin (selinna), and other language-of-origin-specific transliterations are returned.

4.2 Statistical Machine Translation

4.2.1 Chinese to English SMT

We used a hierarchical phrase-based machine translation system(Chiang, 2007) to translate Chinese documents into English. Translation grammars were extracted using the suffix-array variant of Hiero (Lopez, 2007). Decoding (the process of searching for the best translation given the input) was done using *cdec* (Dyer et al., 2010). We extracted translation grammars from the non-UN portions and non-HK Hansards portions of the NIST OpenMT '08 training corpora[3] and used the following features:

1. SCFG translation rule score
2. SCFG translation rule arity (number of non-terminals)
3. Language model score (with a 3-gram English language model trained on the training data and English Gigaword)
4. Word penalty
5. Rule-based translation for numbers

The feature weights were tuned on the MT03 development data using minimum error rate training (MERT) (Och, 2003) and the final weights achieved a BLEU score of 35.51 on the MT02 development data.

Approximately 6% of Chinese characters in query documents in the evaluation collection remained untranslated because of vocabulary limitations in our training data.

Different from typical machine translation tasks, for TAC-KBP we desired to retain the span markings of query entities in the translation. However, the translation model which we were working with is not able to preserve such information through translation; neither is it capable of enforcing a contiguous span of words in the source language being translated into a contiguous span of words in the target language. To overcome these limitations, we translated the document without markings and restored the span markings with the information from translation derivation trees using the following heuristic:

1. Visit the derivation tree nodes recursively in a top-down order;
2. If the node is a leaf node, mark if and only if its source span intersects with the span we wish to restore;
3. For any internal node,
   (a) If its source span is a subset of the span we wish to restore, mark the transla-

[3] http://www.itl.nist.gov/iad/mig/tests/mt/2008/doc/mt08_constrained.html

tion of this whole node and there is no need to visit its children;

(b) If its source span does not intersect with the span we wish to restore, leave the translation of this whole node unmarked and there is no need to visit its children;

(c) Otherwise, visit each children.

4.2.2 Enhancing MT Output

The core machine translation system produces output with two problems. First, it is unable to translate some Chinese words, and these are left as Chinese strings in the MT output. We reasoned that a large percentage of these out-of-vocabulary (OOV) terms were likely to be named entities, and that transliterating them might improve the quality of the resulting text for the purpose of context matching. We therefore looked for any blank-delimited string in the MT output that began with a Unicode character residing in a CJK block, applied our transliterator to the string, and replaced the original string with the transliteration.

The second difficulty with the MT output is that it is all lower case. While this is not a problem for the retrieval aspects of the subsequent processing (which typically downcase all input anyway), it plays havoc with named entity recognition. We use the Ratinov and Roth NER system (2009) out-of-the-box. Because English NER systems such as this one cue heavily off of capitalization, performance on the MT output tends to be poor. To ameliorate the problem, we attempt to automatically re-introduce case into the output text. Good solutions to capitalization are extant (Wang et al., 2006); for uncased source languages such as Chinese the usual approach is to train a language model to guess the proper target language case. As a simple approximation to this, we simply replaced each word in the MT output with the most frequently occurring capitalization of that word in the Google unigram data.[4] We did not measure NER performance on the resulting rewritten files, but qualitative assessment suggested that recall was significantly improved without significantly damaging precision. Extrinsic evaluation showed that entity linking on the NIST training data improved by 2% when capitalization was restored in this way.

[4] LDC2006T13

4.3 MT Example

In Table 6 we show Chinese text from document XIN_CMN_20081230.0139 and corresponding English MT output with OOVs subsequently transliterated and after case restoration was conducted.

4.4 CLIR

To match contexts when the query document and KB are in different languages we treated cross-language context linking as a Cross-Language Information Retrieval (CLIR) problem in which the query is created from the words in the vicinity of mentions of the query name and the documents are the text associated with the Knowledge Base Entry. We adopted Probabilistic Structured Queries (PSQ) (Darwish and Oard, 2003), the key idea of which is to treat alternate translations of a query term as synonyms and to weight the contributions of each "synonym" using a statistical translation model from the same statistical machine translation system described below.

We indexed the Wikipedia articles in our test collection using a publicly available IR tool (Indri), learned isolated word translation probabilities from the parallel text using Hiero, and implemented PSQ using Indri's *#wsyn* operator. To limit statistical noise and query latency, low-probability translations ($p < 0.005$) were deleted (and remaining probabilities were renormalized). Based on initial tests on training data, we found that a query composed of the contextual window size of $\pm$ 40 terms to the left and right of the query name mention was optimal. Because we wanted a feature complementary to name matching features, we did not intentionally include the query name in the Indri query (although it may appear in its own context if it is repeated within 40 words).

(小标题)最有故事的新词汇 23岁的大耳朵“飞鱼”菲尔普斯凭他在今夏创造的一系列奇迹,为英文这门语言贡献了一个新词汇,“Phelpsian”,这个以他的名字为词根的形容词意为“前所未有的(胜利)”。 北京奥运会上,这位水中如飞、离开水却连路都走不稳的泳坛奇才参赛8项拿回8块金牌打破7项世界纪录,不仅改写了同胞斯皮茨保持的一届奥运会夺得金牌最多的纪录,还成为奥运史上收获金牌最多的运动员,斯皮茨评价这一成绩为“史诗般伟大”。 一些人开始称菲尔普斯是游泳运动历史上最伟大的全能运动员,“超人”二字显然已不足以承载人们想堆砌在他身上的赞美、感慨和惊叹。忘记“巴尔的摩子弹”的绰号吧,菲尔普斯的对手说:“他来自外星,他来自未来。”美国总统布什和刚退休的盖茨都曾到场为他助威,网坛明星小威廉姆斯也想要到他的签名,而中国某体育解说员希望奥运会结束后把菲尔普斯留下来,“让全世界的医学专家做个解剖,看看他有什么特殊的装置”	” (subhead) have the biggest story of the New Terms the 23 - year - old large ears “ flying ” Phelps on his summer in a Series of Miracle , to create the English language of the doors to a New Terms , “ Phelpsian ” , um , to his name to describe root expression for “ unprecedented (victory) ” . Beijing Olympic Games , the water Yu Fi left water but even they CA n’t walk Road from the Swimming genius eight received eight Golds seven World records , not only has rewritten compatriots Slovak Spitz maintained a record of the most Olympic Gold Medal of the harvest has also become Olympic History , the most Gold medals for the athletes , Sri Lanka Spitz spoke highly of achievements in this “ epic great ” . some people began that Phelps is Swimming in the History of the movement of the great all athletes , “ Superman ” word has obviously not enough to accommodate people want to piled on him the praise said with emotion and place . forget “ Baltimore bullets ” Phelps ’s opponents say the nicknames : “ he from extraterrestrial , the future . ” retired U.S. President George W. Bush and the Gates had little to cheer for him , Tennis Star Williams also wants to go to his signature , a Sports commentator and China hopes that the Olympic Games after the Phelps stayed , “ allowing a Anatomy , Medical experts around the World see what he has a Special device ” .
(小标题)最神奇的赛场 北京奥运会的两大标志性建筑——“鸟巢”和“水立方”在创造建筑奇迹的同时也“创造”着竞技奇迹。前者目睹了“怪鸟”博尔特的横空出世以及撑杆跳女皇伊辛巴耶娃的第24次打破世界纪录;后者的梦幻泳池则催生了“八金奇迹”菲尔普斯以及24项世界纪录和66项奥运会纪录,分别占北京奥运会产生的这两项纪录总数的近三分之二和四分之三强。 尽管专家和专业人士分析后认为“鸟巢”里诞生超多奇迹、“水立方”变成“水魔方”与跑道气场或泳池的深度等客观条件无关,在大鸟窝里奔跑跳跃或在泡泡里游泳“特别舒服”也不能直接跟好成绩挂钩,但伦敦奥组委为了避嫌还是委婉地表示,2012年伦敦奥运会的游泳比赛不会放在水深3米的泳池里进行。	(subhead) of the most amazing at Beijing Olympic Games , the two major landmarks —— “ Bird ’s nest ” and “ water Cube ” construction Miracle created in the same “ ” created a Sporting Miracle . the former witnessed “ guainiao ’95 there was ” player born the sensational Jump and pole Vault Queen Yelena Isinbayeva part 24 times the World record ; the latter ’s dream pool is the birth of “ eight Gold Miracle ” Phelps and 24 counts of World records and 66 counts of Games record , respectively Accounting for Beijing Olympic Games , a record of the two - thirds of the Total and three quarters of strong . Although experts and professionals after analysis that “ & Miracle out of the Bird ’s nest in ” “ water Cube ” into “ water Rubik’s Cube ” and runway Aura or unrelated to objective conditions , such as the depth of the Swimming pools of the Bird nest in running Jump or in the bubble in Swimming “ particularly comfortable ” cannot directly linking genhao results , However , on London ’s Olympic organizing Committee said in order to Bixian or tactful way , 2012 London Olympic Games Swimming competition will not put 3 - Meter - deep waters in the Swimming pools .

Table 6: Excerpts from XIN_CMN_20081230.0139 referencing US swimmer Michael Phelps. The MT output on the right has been augmented by transliterating OOV words and performing case restoration.

4.5 NIL Clustering

Cross-language NIL clustering proceeded in a similar way as the monolingual case. We built a pairwise classifier that decides whether a pair of candidates is coreferent. We hypothesized that there are peculiarities in which entities are expressed and transliterated in each language, so our model could benefit from learning a separate classifier for each pair of languages of the candidates (*i.e.,* English-English, English-Chinese and Chinese-Chinese). However, results on the provided development data indicated that learning separate classifiers leads to less accurate model is opposed to a single classifier trained on all the data and using the one-best transliteration. Therefore, we adopted the latter approach for our submissions. We again used an exact name match approach as a baseline to compare against.

5 Chinese-English Results

For the cross-language task, queries could be English queries in English documents or Chinese queries with Chinese documents. (Also, we saw at least one case in NIST-provided training data where there was a Roman query string from an otherwise Chinese text document.) We used our English entity linking system for the English language documents, and the cross-language described in Section 4 for the Chinese documents.

We trained our Chinese system using the 1481 NIST-provided Chinese training examples, along with some in-house produced training exemplars. To create the later, we ran the BBN SERIF tool on documents from Chinese Gigaword Fourth Edition, and identified relatively high frequency entities. In total 1117 additional examples were generated, consisting of 406 persons, 403 organizations, and 308 locations. A native Chinese speaker searched online English Wikipedia for these entities, and we then automatically mapped the Wikipedia entities to the TAC KBP knowledge base. The majority of entities in the additional data were present in the KB (1002 of 1117).

We submitted three runs for the cross-language subtask. The submitted runs differ in (a) their approach to candidate identification, and (b) their method of NIL clustering (statistical, or exact-match). We had considered submitting a run that used CLIR-only, or MT-only as a method of modeling context; however, in 2-fold cross-validation experiments using our two sources of training data, we always achieved better performance using both the MT documents and the CLIR features. Neither run made any use of live Web access.

The two approaches to candidate identification were essentially a choice between a lower and a higher recall approach. Our concern was that in cases where transliteration dictionary lookup failed and we resorted to pinyinization, the pinyinized names, particularly for non-Chinese names could be fairly dissimilar to that of the English KB name for an entity, and therefore very difficult to match. For example, the surname of former Florida governor Charlie Crist, the target of queries EL_CLCMN_02317 & EL_CLCMN_02318, is written 克利斯特 and pinyin produces: "kelisite".

Our lower-recall (higher precision) method of generating candidates for a particular Chinese name is based on 1-best transliterator output, and the higher-recall approach uses all transliterator possibilities for the name. For the later method, the candidate set is built from the union of the candidate set identified for each transliteration variant. For the *Crist* example above, candidates include "christo", "crist", and "krist", all of which are much easier to match.

Our three submitted runs were:

- hltcoe1: MT+CLIR, with 1-best transliteration, our regular KB, and exact match NIL clustering
- hltcoe2: MT+CLIR, with all-transliteration-variants used for candidate identification, the regular KB, and statistical NIL clustering
- hltcoe3: MT+CLIR, with all-transliteration-variants used for candidate identification, the regular KB, and exact-match NIL clustering

Run	P@1	B^3P	B^3R	B^3F_1
hltcoe1	0.800	0.702	0.779	0.738
hltcoe2	0.790	0.708	0.703	0.705
hltcoe3	0.790	0.689	0.765	0.725

Table 7: Scores from NIST. Scores for cross-lingual entity-linking task.

Summary results for the three conditions are given in Table 7. Exact-match NIL clustering again prevailed over our statistical method. Contrary to our experiments with our training data, 1-best transliteration (used in hltcoe1) outperformed the higher recall approach. This is no doubt influenced by the fact that over three-quarters of the evaluation queries were able to be transliterated using our transliteration dictionary.

According to summary information provided by NIST concerning submissions from 10 teams, the best reported B^3F_1 score for web-less runs on this task is 0.788, with a median score of 0.675. Our top score (from hltcoe1) was 0.738.

6 Conclusions

We were pleased to see that our previous English entity linking system continued to perform well on the TAC KBP task. While we found the Chinese to English cross-language task to be quite challenging, by incorporating features to address name translation and context (*i.e.,* document) translation, we believe that our approach has been shown to be fairly adaptable to other languages. While high-quality machine translation may not be available for every language, our system can adopt either an MT or a CLIR-based solution to map document context to English.

Acknowledgments

We are very appreciative for the assistance we received from Dawn Lawrie, Chris Callison-Burch, Ann Irvine, Kristy Hollingshead, and Vlad Eidelman during this project. Ben Shayne and Scott Roberts did us a great service by expeditiously restoring our computing cluster to normal function after the prolonged electrical disruption.

References

A. Bagga and B. Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *the 36th Annual Meeting of the Association for Computational Linguistics*.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1.

D. Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Kareem Darwish and Douglas W. Oard. 2003. Probabilistic structured query methods. In *ACM SIGIR*, pages 338–344. ACM.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 7–12, Uppsala, Sweden, July. Association for Computational Linguistics.

Guorong Guo, editor. 2007. *Names of the world's peoples: a comprehensive dictionary of names in Roman-Chinese*. Xinhua News Agency, 2 edition.

Shudong Huang. 2005. Chinese <-> English Name Entity Lists v 1.0.

Ann Irvine, Chris Callison-Burch, and Alexandre Klementiev. 2010. Transliterating from all languages. In *AMTA*.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Knowledge Discovery and Data Mining (KDD)*.

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 976–985, Prague, Czech Republic, June. Association for Computational Linguistics.

James Mayfield, Dawn Lawrie, Paul McNamee, and Douglas W. Oard. 2011. Building a cross-language entity linking collection in twenty-one languages. In *Cross-Language Evaluation Forum (CLEF)*.

Andrew McCallum, Kamal Nigam, and Lyle Ungar. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In *Knowledge Discovery and Data Mining (KDD)*.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas W. Oard, and David Doermann. 2011. Cross-language entity linking. In *International Joint Conference on Natural Language Processing (IJCNLP-2011)*, Chiang Mai, Thailand, November. Association for Computational Linguistics.

Paul McNamee. 2010. HLTCOE efforts in entity linking at TAC KBP 2010. In *Text Analysis Conference (TAC)*, Gaithersburg, Maryland, November.

F.J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado, June. Association for Computational Linguistics.

Vasudeva Varma, Praveen Bysani, Kranthi Reddy, Vijay Bharath Reddy, Sudheer Kovelamudi, Srikanth Reddy Vaddepally, Radheshyam Nanduri, Kiran Kumar N, Santhosh Gsk, and Prasad Pingali. 2010. IIIT Hyderabad in Guided Summarization and Knowledge Base Population. In *Text Analysis Conference (TAC)*, Gaithersburg, Maryland, November.

Wei Wang, Kevin Knight, and Daniel Marcu. 2006. Capitalizing machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 1–8, New York City, USA, June. Association for Computational Linguistics.

Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina. 2009. Entity resolution with iterative blocking. In *SIGMOD 2009*, pages 219–232. ACM.

Dingguo Zhou, editor. 2008. *Place names of the world: a comprehensive dictionary of place names in Roman-Chinese*. Xinhua News Agency, 1 edition.