Overview of the TAC2011 Knowledge Base Population Track

Heng Ji[1], Ralph Grishman[2] and Hoa Trang Dang[3]

[1]Computer Science Department and Linguistics Department
Queens College and Graduate Center, City University of New York, New York, NY, USA
hengji@cs.qc.cuny.edu

[2] Computer Science Department, New York University, New York, NY, USA
grishman@cs.nyu.edu

[3] National Institute of Standards and Technology, Gaithersburg, MD, USA
hoa.dang@nist.gov

Abstract

In this paper we give an overview of the Knowledge Base Population (KBP) track at TAC 2011. The main goal of KBP is to promote research in discovering facts about entities and expanding a structured knowledge base with this information. Compared to KBP2010, we extended the Entity Linking task to require clustering of mentions of entities not already in the KB ('NIL queries'). We also introduced two new tasks - (1) Cross-lingual Entity Linking: Given a set of multi-lingual (English and Chinese) queries, the system is required to provide the ID of the English KB entry to which each query refers and cluster NIL queries without KB references; and (2) Temporal Slot Filling: given an entity query, the system is required to discover the start and end dates for any identified slot fill. KBP2011 has attracted many participants (over 65 teams registered, among which 35 teams submitted results). In this paper we provide an overview of the task definition, annotation issues, successful methods and research challenges associated with each task in KBP2011.

1 Introduction

The main goal of the Knowledge Base Population (KBP) track at Text Analysis Conference (TAC) is to gather information about an entity that is scattered among the documents of a large collection, and then use the extracted information to populate an existing knowledge base (KB). KBP is done through two separate sub-tasks - Entity Linking and Slot Filling. For both tasks, the system is given a query consisting of a name and a document in which this name appears. For Entity Linking, it must cluster the queries and decide whether this cluster corresponds to an entry in a KB and, if so, which one. For Slot Filling, the system must determine from a large source collection of documents the values of specified attributes ('slots') of the entity, such as the age and birthplace of a person or the top employees of a corporation.

This is the third year that we are conducting a KBP evaluation. In total 35 teams submitted results for one or both sub-tasks. Compared to KBP2010 (Ji et al., 2010; Ji and Grishman, 2011), we introduced two new tasks - Cross-lingual Entity Linking and Temporal Slot Filling. In 2009 and 2010, all of the KBP tasks were limited to monolingual processing. However, for certain queries, many slot fills can only be discovered from documents in foreign languages. Therefore we introduced a new cross-lingual entity linking task, to link a given entity from a Chinese document to an English KB. In addition, the information obtained from KBP2009 and 2010 was viewed as static, ignoring the temporal dimension that is relevant to many types of slots. While this is a reasonable simplification in many situations, it is unsatisfactory for applications that require some awareness of the time span during which a fact was valid. Therefore we introduced another new task of temporal slot filling - a slot

filling system needs to discover the start and end dates for any identified slot fill. To summarize, the following improvements were made this year:

- Defined a new task, Cross-lingual Entity Linking, and prepared its annotation guideline and training corpora;
- Defined a new task, Temporal Slot Filling, and prepared its annotation guideline and training corpora;
- Added clustering of entity mentions without KB entries into the Entity Linking task, and developed a new scoring metric incorporating NIL clustering;
- Made systematic corrections to the slot filling guidelines and data annotation;
- Defined a new task, Cross-lingual Slot Filling, and prepared its annotation guideline, in anticipation of future evaluations.

The rest of this paper is structured as follows. Section 2 describes the definition of each task in KBP2011. Section 3 briefly summarizes the participants. Section 4 highlights some annotation efforts. Section 5, 6, 7 and Section 8 summarize the general architecture of each task's systems and evaluation results, and provide some detailed analysis and discussion. From each participant, we only select the best submission without Web access for comparison. Section 9 discusses the experimental results and sketches our future work.

2 Task Definition and Evaluation Metrics

This section will summarize the tasks conducted at KBP 2011. More details regarding data format and scoring software can be found in the KBP 2011 website[1].

2.1 Overview

The overall goal of KBP is to automatically identify salient and novel entities from multiple languages, link them to corresponding Knowledge Base (KB) entries (if the linkage exists) in a target language, then discover attributes about the entities (extract temporal spans about the attributes if there exist dynamic changes), and finally expand the KB with any new attributes. Given a source language S and a target language T, Figure 1 depicts the general architecture of current KBP tasks.

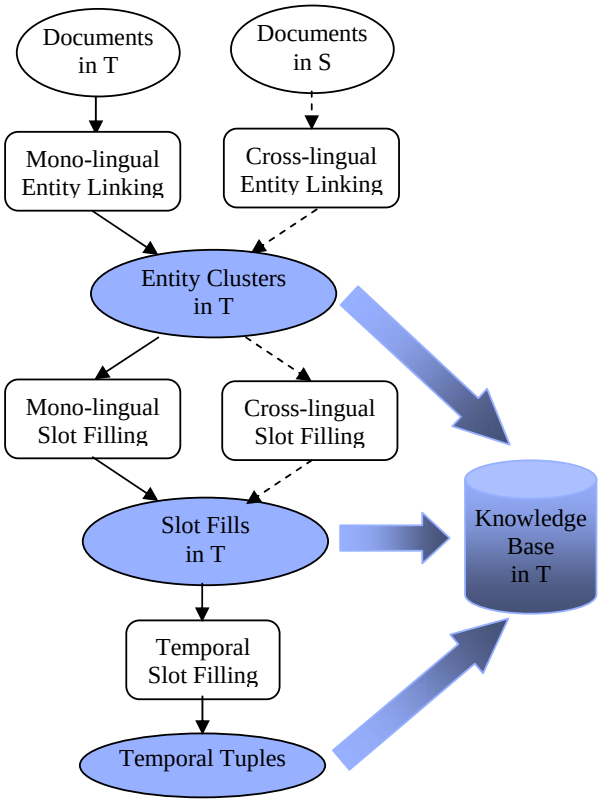

Figure 1: Overview of KBP Tasks

For example, the actor "*James Parsons*" became famous after he got an Emmy Award on August 29, 2010. A user may be interested in reading an accurate and concise profile (facts) about him. An Entity Linking system can link any document including "*James Parsons*" to the corresponding KBP entry, or determine that no corresponding entry exists (for the real application this means a new Wikipedia entry needs to be constructed about this person). This process involves both name disambiguation (e.g. the actor "*James Parsons*" should not be linked to the lawyer "*James A. Parsons*" or the judge "*James B. Parsons*") and name variant clustering ("*James Parsons*" = "*Jim Parsons*" for the actor).

In addition, a slot filling system is required to discover the values of pre-defined attributes about

[1] http://nlp.cs.qc.cuny.edu/kbp/2011/

"*James Parsons*". For example, if "*University of Houston*" is extracted as his "*school_attended*" attribute, and this fact does not exist in the KB yet, the system should add this information to expand the KB. Furthermore, certain information about many entities, such as local government agencies (e.g. "*Independent Commission Against Corruption*") or politicians (e.g. "*Chen Shui-bian*"), can only be discovered and enriched from foreign languages. Therefore cross-lingual entity linking and slot filling are needed in order to extract more complete information.

Finally, it has been estimated that one of every fifty lines of database application code involves a date or time value (Snodgrass et al., 1998). In fact, many statements in text are temporally qualified. For example, most of the slot types change over time and thus can be temporally bounded (e.g. for person-related attributes such as place of residence, schools attended, job title, employer, membership in organizations, spouse; for organizations top employees/members, number of employees). Temporal Information Extraction is also of significant interest for a variety of NLP applications such as Textual Inference (Baral et al., 2005), Multi-document Text summarization (Elhadad et al., 2002) and Template Based Question Answering (Schockaert et al., 2006). While the extraction of temporal arguments for relations and events has recently received the attention of the TempEval community (Verhagen et al., 2007; Pustejovsky and Verhagen, 2009), it focused on extracting temporal relations from individual documents. Therefore we introduced temporal slot filling into the KBP framework, to identify time spans for slot fills (e.g. "*James Parsons*" lived in "*San Diego, CA*" from 1999 to 2001).

For the evaluation an initial (or reference) KB derived from Wikipedia Infoboxes is provided to the systems, along with a large collection of source documents. As in KBP 2010, participants can submit up to three runs for each task. They were asked to make at least one run subject to certain resource constraints, primarily that the run be made as a 'closed' system, namely one which does not access the Web during the evaluation period.

2.2 Mono-lingual Entity Linking Task

2.2.1 Task Definition

In the Entity Linking task, given a 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 background document ID, the system is required to provide the ID of the KB entry to which the name refers; or NIL if there is no such KB entry. In addition, an entity linking 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.

For example, some training queries are as follows:

```
⟨query id=``EL000434"⟩
   ⟨name⟩Brentwood⟨/name⟩
   ⟨docid⟩eng-WL-11-174588-12938415⟨/docid⟩
⟨/query⟩

⟨query id=``EL000441"⟩
   ⟨name⟩Brentwood⟨/name⟩
   ⟨docid⟩eng-WL-109-174581-12950841⟨/docid⟩
⟨/query⟩

⟨query id=``EL000445"⟩
   ⟨name⟩Brentwood⟨/name⟩
   ⟨docid⟩eng-WL-109-174581-12950796⟨/docid⟩
⟨/query⟩

⟨query id=``EL000446"⟩
   ⟨name⟩Brentwood⟨/name⟩
   ⟨docid⟩eng-WL-11-174588-12938647⟨/docid⟩
⟨/query⟩

⟨query id=``EL000449"⟩
   ⟨name⟩Brentwood⟨/name⟩
   ⟨docid⟩eng-WL-11-174588-12938245⟨/docid⟩
⟨/query⟩
```

The query "EL000441" should be linked to the KB entry "E0144449" ("*Brentwood School (Los Angeles, California, U.S.)*"); the query 'EL000449" should be linked to the KB entry "E0735022" ("*Brentwood, California, U.S.*"), and the queries and "EL000434", "EL000445" and "EL000446" have no corresponding KB entries and should be clustered into a cluster with a unique ID "NIL0004".

For the regular entity linking task, the system may consult the text from the Wikipedia pages associated with the KB nodes. However, in a more

realistic setting, when a salient and novel entity appears in news or web data, there may not be many Wikipedia texts to utilize. Therefore as in KBP 2010, an optional "no wikitext" entity linking task was conducted in 2011, in which the systems can only use the attributes in the KB; this corresponds to the task of updating a structured KB with no 'backing' text.

2.2.2 Scoring Metric

We apply a modified B-Cubed (Bagga and Baldwin, 1998) metric (called B-Cubed+) to evaluate these clusters. Let us use the following notation: $L(e)$ and $C(e)$: the category and the cluster of an entity mention e, $SI(e)$ and $GI(e)$: the system and gold-standard KB identifier for an entity mention e. We can define the correctness of the relation between two entity mentions e and e' in the distribution 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. 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:

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

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

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

As we will see later in the entity linking performance in Figure 3, the scores based on B-cubed+ F-Measure highly correlates with the scores based on the old micro-averaged accuracy metric in KBP2009 and KBP2010 (the correlation is about 0.99).

2.3 Cross-lingual Entity Linking

In KBP2011 we extend entity linking to a cross-lingual setting, in which the queries come from both English and Chinese.

For example, some training queries are as follows:

```
⟨query id=''EL_CLCMN_03011"⟩
   ⟨name⟩李娜⟨/name⟩
   ⟨docid⟩XIN_CMN_20050429.0146⟨/docid⟩
⟨/query⟩

⟨query id=''EL_CLCMN_03012"⟩
   ⟨name⟩李娜⟨/name⟩
   ⟨docid⟩XIN_CMN_20080211.0135⟨/docid⟩
⟨/query⟩

⟨query id=''EL_CLCMN_03013"⟩
   ⟨name⟩李娜⟨/name⟩
   ⟨docid⟩XIN_CMN_19991116.0016⟨/docid⟩
⟨/query⟩

⟨query id=''EL_CLENG_03959"⟩
   ⟨name⟩Li Na⟨/name⟩
   ⟨docid⟩AFP_ENG_20070116.0014⟨/docid⟩
⟨/query⟩

⟨query id=''EL_CLENG_03960"⟩
   ⟨name⟩Li Na⟨/name⟩
   ⟨docid⟩AFP_ENG_20080815.0791⟨/docid⟩
⟨/query⟩

⟨query id=''EL_CLENG_03961"⟩
   ⟨name⟩Li Na⟨/name⟩
   ⟨docid⟩NYT_ENG_20080803.0109⟨/docid⟩
⟨/query⟩

⟨query id=''EL_CLENG_03962"⟩
   ⟨name⟩Li Na⟨/name⟩
   ⟨docid⟩AFP_ENG_20080919.0491⟨/docid⟩
⟨/query⟩

⟨query id=''EL_CLENG_03963"⟩
   ⟨name⟩Li Na⟨/name⟩
   ⟨docid⟩AFP_ENG_20080829.0072⟨/docid⟩
⟨/query⟩

⟨query id=''EL_CLENG_03964"⟩
   ⟨name⟩Li Na⟨/name⟩
   ⟨docid⟩APW_ENG_20080814.0998⟨/docid⟩
```

```
⟨/query⟩
```

A cross-lingual entity linking system should cluster the queries "EL_CLCMN_03012" and "EL_CLCMN_03013" and link the cluster to the KB entry "E0026964" (the track cyclist), and cluster the remaining queries and link that cluster to another KB entry "E0128750" (the tennis player). The B-Cubed+ metric is also applied to evaluate cross-lingual entity linking systems.

2.4 Regular Mono-lingual Slot Filling Task

2.4.1 Task Definition

The goal of Slot Filling is to collect from the corpus information regarding certain attributes of an entity, which may be a person or some type of organization. Each query in the Slot Filling task consists of the name of the entity, its type (person or organization), a document (from the corpus) in which the name appears (to disambiguate the query in case there are multiple entities with the same name), its node ID (if the entity appears in the knowledge base), and the attributes which need not be filled. Attributes are excluded if they are already filled in the reference data base and can only take on a single value. An example query is

```
⟨query id="SF114"⟩
   ⟨name⟩Masi Oka⟨/name⟩
   ⟨docid⟩eng-WL-11-174592⟨/docid⟩
   ⟨enttype⟩PER⟨/enttype⟩
   ⟨nodeid⟩E0300113⟨/nodeid⟩
   ⟨ignore⟩per:date_of_birth
per:age per:country_of_birth
per:city_of_birth⟨/ignore⟩
⟨/query⟩
```

Along with each slot fill, the system must provide the ID of a document which supports the correct-ness of this fill. If the corpus does not provide any information for a given attribute, the system should generate a NIL response (and no document ID). The sets of attributes are listed in Table 1.

For each attribute we indicate the type of fill and whether the fill must be (at most) a single value or can be a list of values. Since the overall goal is to augment an existing KB, two types of redundancy in list-valued slots must be detected and avoided. First, two fills for the same entity and slot must refer to distinct individuals. Second, if the knowledge base already has one or more values for a slot, items in the system output must be distinct from those already in the knowledge base. In both cases, it is not sufficient that the strings be distinct; the fills must refer to distinct individuals. For example, if the knowledge base already has a slot fill "William Jefferson Clinton", the system should not generate a fill "Bill Clinton" for the same slot.

2.4.2 Scoring Metric

As is the case with IR (document retrieval) evaluations, it is not feasible to prepare a comprehensive slot-filling answer key in advance. Because of the difficulty of finding information in such a large corpus, any manually-prepared key is likely to be quite incomplete. Instead (as for IR) we pool the responses from all the systems and have human assessors judge the responses. To increase the chance of including answers which may be particularly difficult for a computer to find, LDC did prepare a manual key which was included in the pooled responses.

Each response is rated as correct, inexact, redundant, or wrong. A response is inexact if it either includes part of the correct answer or includes the correct answer plus extraneous material. No credit is given for inexact answers. Two types of redundant answers are flagged for list-valued slots. First, a system response may be equivalent to an answer in the reference knowledge base; this is considered incorrect. Second, two system responses for the same attribute may be equivalent; in the latter case, only the first of a set of equivalent answers is marked correct. (This is implemented by assigning each correct answer to an equivalence class, and only giving credit for one member of each class.)

Given these judgments, we can count:

Correct = total number of non-NIL system output slots judged correct

System = total number of non-NIL system output slots

Reference = number of single-valued slots with a correct non-NIL response + number of equivalence classes for all list-valued slots

Person Slots			Organization Slots		
alternate_names	Name	List	alternate_names	Name	List
date_of_birth	Value	Single	political/religious_affiliation	Name	List
age	Value	Single	top_members/employees	Name	List
country_of_birth	Name	Single	number_of_employees/members	Value	Single
stateorprovince_of_birth	Name	Single	members	Name	List
city_of_birth	Name	Single	member_of	Name	List
origin	Name	List	subsidiaries	Name	List
date_of_death	Value	Single	parents	Name	List
country_of_death	Name	Single	founded_by	Name	List
stateorprovince_of_death	Name	Single	founded	Value	Single
city_of_death	Name	Single	dissolved	Value	Single
cause_of_death	String	Single	country_of_headquarters	Name	Single
countries_of_residence	Name	List	stateorprovince_of_headquarters	Name	Single
stateorprovinces_of_residence	Name	List	city_of_headquarters	Name	Single
cities_of_residence	Name	List	shareholders	Name	List
schools_attended	Name	List	website	String	Single
title	String	List			
member_of	Name	List			
employee_of	Name	List			
religion	String	Single			
spouse	Name	List			
children	Name	List			
parents	Name	List			
siblings	Name	List			
other_family	Name	List			
charges	Name	List			

Table 1: Slot Types for KBP2011 Regular Slot Filling

$$Precision = \frac{Correct}{Reference}$$
$$Recall = \frac{Correct}{System}$$
$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

The F score is the primary metric for regular slot filling system evaluation.

2.5 Temporal Slot Filling Task

2.5.1 Task Definition

In KBP2011 we also added a new task of temporal slot filling. The goal of this new task is to add limited temporal information to selected slots in the regular slot filling output. We limited temporal information to the following *time intensive* slot types:

- per:spouse
- per:title
- per:employee_of
- per:member_of
- per:cities_of_residence
- per:stateorprovinces_of_residence
- per:countries_of_residence
- org:top_employees/members

There are two versions of the task, the full temporal task and the diagnostic temporal task. For

the full temporal task, the system is given a query file just as for the regular slot filling task, and is expected to generate a slot filling output augmented with temporal information as described below. Therefore for the full task, slot fills and temporal information must be gathered across the entire corpus. For the diagnostic temporal task, the system is given two files, a query file and a slot file. The slot file includes the same form as the output of a run for the regular slot filling task: each line specifies a query, a slot, a slot value, and a document supporting that slot value. The system should determine the temporal information for each specified slot value, based only on the information in the document specified in the slot file. The output for the full temporal task is scored through system output pooling, like the regular slot filling task. The diagnostic temporal task is based on a set of slot fills tagged through manual annotation, and is scored automatically.

2.5.2 Representation of Temporal Information

Temporal information can be scattered across documents (e.g. in one document "John joined Microsoft in Sept. 1990" and in another document "Microsoft renewed his contract yesterday "), and expressed with different granularities (e.g. "He started working for Microsoft in the 90s", "He began his contract in September of this year"). According to our task definition, we assume that: (i) events are not discontinuous in time; (ii) the temporal information is distributed across several documents; and (iii) both the gold standard and system outputs can contain uncertainty. This uncertainty can be due to the variable levels of granularity of temporal information (e.g. years, months) or to the reasoning based on temporal order relations ("He worked for Microsoft before working for Apple").

Given the previous assumptions the representation model should consider temporal ranges for both the beginning and ending points. For simplicity, we assume that uncertainty follows uniform distributions over time ranges. Our representation model consists of a tuple $< t_1, t_2, t_3, t_4 >$, which represents the set S of possible beginnings and endings of an event such that:

$$S = \{< t_{init}, t_{end} > | (t_1 < t_{init} < t_2) \wedge (t_3 < t_{end} < t_4)\}$$

In other words, t_1 and t_3 represent the lower bounds for the beginning and ending points respectively, while t_2 and t_4 represent the upper bounds.

This temporal representation model can represent well temporal aggregation, temporal relations between event and time (Allen, 1983), and temporal relations between two events when one of the events is anchored in time. It also provides a straightforward method to detect inconsistencies when aggregating temporal information in a tuple.

The main limitation of assuming that events are continuous is that our representation model is not able to capture some relations such as regularly recurring events ("each Friday"), some fuzzy relations ("lately", "recently") that are encoded with the SET type in TimeML (Pustejovsky et al., 2003), durations where neither endpoint is known ("he worked for IBM for 7 years"), relations between slots ("she married Fred two years after moving to Seattle"), slot values which are true over multiple disjoint intervals ("Cleveland was President from 1885 to 1889 and from 1893 to 1897") and the same slot value ("President") affiliated with different entities ("Mary was the President of Student Union from 1998 to 2003 and the President of Woman Sports Association from 2002 to 2005").

Table 2 presents some examples of 4-tuple representation, assuming the publication date of the text is January 1, 2001.

2.5.3 Scoring Metric

We define a metric $Q(S)$ that compares a system's output $S =< t_1, t_2, t_3, t_4 >$ against a gold standard tuple $S_g =< g_1, g_2, g_3, g_4 >$, based on the absolute distances between t_i and g_i:

$$Q(S) = \frac{1}{4} \sum_i \frac{1}{1 + |t_i - g_i|}$$

The absence of a constraint in t_1 or t_3 is treated as a value of -∞; the absence of a constraint in t_2 or t_4 is treated as a value of +∞. The unit of each tuple element is counted based on years.

Overall system scores are computed the same way as for regular slot filling (see section 2.4.2) except that, in computing the value of correct, we take the sum over all *correct* slot fills of S(slot).

Assume the set of gold standard tuples is $\{G^1, G^2, ..., G^N\}$, and the set of system output tuples is $\{S^1, S^2, ..., S^M\}$, where each G^i is a four tuple for the i-th slot fill in gold standard $<$

Document text	T1	T2	T3	T4
Chairman Smith	-	20010101	20010101	-
Smith, who has been chairman for two years	-	19990101	20010101	-
Smith, who was named chairman two years ago	19990101	19990101	19990101	-
Smith, who resigned last October	-	20001001	20001001	20001031
Smith served as chairman for 7 years before leaving in 1991	19840101	19841231	19910101	19911231
Smith was named chairman in 1980	19800101	19801231	19800101	-

Table 2: 4-tuple Representation Example

$g_1, g_2, g_3, g_4 >$, each S^j is a four tuple for the j-th slot fill in system output $< t_1, t_2, t_3, t_4 >$, each element is associated with an instance of unique slot fill and scored independently. Then we can get the following Precision, Recall and F-measure scores:

$$Precision = \frac{\sum_{S^i \in C(S)} Q(S^i)}{M}$$

$$Recall = \frac{\sum_{S^i \in C(S)} Q(S^i)}{N}$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Where $C(S)$ is set of all instances in system output which have correct slot filling answers, and $Q(S)$ is quality value of S. Therefore for temporal SF task, a correct slot fill with temporal information gets credit $Q(S)$ (instead of 1 as in regular slot filling task).

3 Participants Overview

Table 3 summarizes the participants for each task. Over 65 teams registered for KBP 2011 (not including the RTE-KBP Pilot task (Bentivogli et al., 2010)), among which 35 teams submitted results. Table 4 shows the number of participants and submissions compared to KBP2009 and KBP2010.

4 Data Annotation

The details of the data annotation for KBP are presented in a separate paper by the Linguistic Data Consortium (Li et al., 2011).

The English text corpus is unchanged from KBP2010, consisting of 1,286,609 newswire documents, 490,596 web documents, and 683 other documents. For cross-lingual entity linking we used approximately one million news documents from Chinese Gigaword. The English reference Knowledge Base was also unchanged, consisting of 818,741 nodes derived from an October 2008 dump of English Wikipedia. Compared to KBP2009 and KBP2010, we annotated new evaluation data for all tasks, and training data for two new tasks. Table 5 summarizes the new KBP2011 training and evaluation data provided for participants.

Task	Training			Evaluation		
	PER	GPE	ORG	PER	GPE	ORG
MLEL	-	-	-	750	750	750
CLEL	817	685	660	824	642	710
SF	-	-	-	50	-	50
TSF	40	-	10	80	-	20

Table 5: New Data Prepared for KBP2011

A manual annotation for the 2010 slot-filling task (prepared by LDC) was included along with the pooled system outputs and the pooled slot fills were then manually assessed; the assessors did not know which fills came from the manual annotation. When the manual annotation was then scored against the assessments, the precision was only 70.1%. In other words, about one-third of the manual fills were considered incorrect by the assessors. Some errors could be attributed to revisions in the guidelines in preparation for assessment, but more generally the low precision reflected underspecification of the slot fill guidelines, particularly for some of the most common slots. To address these problems, several refinements were made to the slot filling guidelines for 2011, including:

- **org:top_members/employees**: made clear that

Team Name	Organization	Entity Linking		Slot Filling	
		Mono-lingual	Cross-lingual	Regular	Temporal
CEA_LVIC	Atomic Energy and Alternative Energies Commission			√	
CMCRC	Capital Markets Cooperative Research Centre	√			
COGCOMP	University of Illinois at Urbana Champaign	√			
CUNY_BLENDER	City University of New York	√	√		√
CUNY_UIUC_SRI	City University of New York, University of Illinois at Urbana Champaign and SRI International	√	√		
DAI	Technische Universität Berlin / DAI-Labor	√			
DMIR_INESCID	Instituto Superior Técnico / INESC-ID Lisboa	√			
ECNU	East China Normal University	√	√	√	
HIT	Harbin Institute of Technology	√			
HITS	Heidelberg Institute for Theoretical Studies		√		
HLTCOE	Johns Hopkins University Human Language Technology Center of Excellence	√	√		
ICL_KBP	Institute of Computational Linguistics, Peking University	√		√	
IIRG	University College Dublin	√		√	√
ILPS	University of Amsterdam			√	
KBP2011PKUTM	Institute of Computer Science and Technology, Peking University	√			
LCC	Language Computer Corporation	√	√		
LSV	Saarland University		√	√	
MS_MLI	Microsoft Research	√			
MSRA	Microsoft Research Asia	√			
NLPR_TAC	Institute of Automation, Chinese Academy of Sciences	√			
NUSchime	National University of Singapore	√			
NYU	New York University			√	
PolyUCOMP	Department of Computing, The Hong Kong Polytechnic University	√		√	
PRIS	Pattern Recognition and Intelligent System Lab, Beijing University of Posts and Telecommunications	√		√	
REACTION	Lasige, Faculty of Sciences, University of Lisbon	√	√		
SIEL_IIITH	International Institute of Information Technology, Hyderabad	√			
STANFORD1	Stanford University			√	√
STANFORD2	Stanford University		√		
STANFORD_UBC	Stanford University and University of the Basque Country	√	√		
THUNLP	Tsinghua University	√	√		
UALBERTA	University of Alberta			√	
UBC	University of the Basque Country			√	
UNED	NLP & IR Group, National Distance Education University			√	√
USFD2011	University of Sheffield	√		√	√
WBSG	Free University of Berlin	√			

Table 3: Overview of KBP2011 Participants

Task / Participants/Year		Entity Linking			Slot Filling			
		Mono-lingual		Cross-lingual	Regular	Surprise	Temporal	
		Regular	Optional				Full	Diagnostic
#Teams	2009	13	-	-	8	-	-	-
	2010	16	7	-	15	5	-	-
	2011	**22**	**8**	**11**	14	-	**5**	**4**
#Submissions	2009	35	-	-	16	-	-	-
	2010	46	20	-	31	6	-	-
	2011	**53**	15	**27**	31	-	**11**	**7**

Table 4: Number of KBP Participants and Submissions

top members of departments or subsidiaries are not top members of the parent organization; a top employee should have decision-making authority over the entire organization

- **per:age**: age must be explicitly given in text (may not be calculated)
- **per:residence**: must be lexically supported ("resides", "grew up", etc.)
- **per:title**: more guidance was provided regarding which modifiers should be included
- **per:member_of** and **per:employee_of**: guidance was added for some special cases

The net result was that the precision of the manual annotation for 2011 improved to 86.2%.

In addition, some of the earlier manual annotations (used as training data) were corrected to more closely follow the guidelines.

5 Mono-lingual Entity Linking

5.1 Approach Overview

5.1.1 General Architecture

A typical KBP2011 mono-lingual entity linking system architecture is summarized in Figure 2. It includes five steps: (1) query expansion - expand the query into a richer set of forms using Wikipedia structure mining or coreference resolution in the background document; (2) candidate generation - finding all possible KB entries that a query might link to; (3) candidate ranking - rank the probabilities of all candidates; (4) NIL detection and clustering - detect the NILs which got low confidence at matching the top KB entries from step (3), and group the NIL queries into clusters. Table 6 summarizes the systems which exploited different approaches at each step. In the following subsections we will highlight the new and effective techniques used in entity linking.

5.1.2 Ranking Features

In entity linking, query expansion techniques are alike across systems, and KB node candidate generation methods normally achieve more than 95% recall. Even after we introduced the new NIL clustering component in this year's evaluation, systems achieved very high performance in clustering itself. Therefore, the most crucial step is ranking the KB candidates and selecting the best node. It's encouraging to see many new and interesting ranking features have been invented during each year's evaluation. Table 7 summarizes the road map of typical ranking features used in mono-lingual entity linking systems in KBP2009, KBP2010 and KBP2011.

5.2 Evaluation Results

5.2.1 Overall Performance

The results of mono-lingual regular entity linking and optional entity linking systems are summarized in Figure 3 and Figure 4.

5.2.2 Comparison with KBP2010 and Human Annotators

Table 8 shows the number of unique names for 2250 queries for different entity types in KBP2010 and KBP2011 evaluations.

There are two principal challenges of entity linking: the same entity can be referred to by more than one name string and the same name string

Methods		System Examples	System Ranking Range		
			NIL	Non-NIL	All
Query Expansion	Wikipedia Hyperlink Mining	CUNY(Cassidy et al., 2011), NUSchime(Zhang et al., 2011)	[2,4]	[3,8]	[3, 4]
	Source document coreference resolution	CUNY(Cassidy et al., 2011)	[4]	[8]	[4]
	Statistical Model	NUSchime(Zhang et al., 2011)	[2]	[3]	[3]
Collaborative Clustering	All entities in the query document	MS_MLI	[5]	[1]	[2]
	Graph based clustering to find collaborators	CUNY(Cassidy et al., 2011)	[4]	[8]	[4]
Candidate Generation	Document semantic analysis and context modeling	CUNY(Cassidy et al., 2011), LCC	[1, 4]	[2, 8]	[1, 4]
	IR	CUNY(Cassidy et al., 2011)	[4]	[8]	[4]
Candidate Ranking	Unsupervised Similarity Computation (e.g. VSM)	CUNY(Cassidy et al., 2011)	[4]	[8]	[4]
	Supervised Classification + Ranking	NUSchime(Zhang et al., 2011), CUNY(Cassidy et al., 2011), DMIR_INESCID, LCC, MS_MLI, DAI, HLTCOE, KBP2011PKUTM, PolyUCOMP	[1, 19]	[2, 19]	[1, 19]
	Rule-based	LCC, HIT	[1, 14]	[2, 5]	[1, 11]
	Global Graph-based Ranking	CUNY(Cassidy et al., 2011), CMCRC	[4, 9]	[6, 8]	[4, 6]
	Rules	HIT	[14]	[5]	[11]
NIL Clustering	Pairwise supervised classification and transitive closure detection	DMIR_INESCID, HLTCOE	[3, 11]	[11, 13]	[9, 12]
	Hierarchical agglomerative clustering	NUSchime(Zhang et al., 2011), LCC	[1, 2]	[2, 3]	[1, 3]
	Graph based clustering	NUSchime(Zhang et al., 2011)	[2]	[3]	[3]
	Topic Modeling	NUSchime(Zhang et al., 2011), PolyUCOMP	[2, 19]	[3, 19]	[3, 19]
	Name String Matching	CUNY(Cassidy et al., 2011), HLTCOE, CMCRC, ICL_KBP, USFD, COGCOMP	[3, 21]	[6, 21]	[4, 21]
	Longest mention with within-document Coreference	CMCRC	[9]	[6]	[6]
	Linking to larger KB and mapping down	CMCRC	[9]	[6]	[6]
	Polysemy and synonymy based clustering	DAI	[16]	[16]	[16]

Table 6: Mono-lingual Entity Linking Method Comparison

Year	All	Person	Organization	Geo-political
2010	752	310	288	194
2011	1325	458	467	403

Table 8: Number of Unique Names in Mono-lingual Entity Linking Evaluation Data Sets

can refer to more than one entity. We defined two

difficulty criteria as follows:

Feature Category	Feature Type	Feature Description
Name	Spelling match	Exact string match, acronym match, alias match, string match based on edit distance, ratio of longest common subsequence to total string length, name component match, first letter match for abbreviations, organization suffix word match
	KB link mining	Name pairs mined from KB text redirect and disambiguation pages
	Name Gazetteer	Organization and geo-political entity abbreviation gazetteers
Document surface	Lexical	Words in KB facts, KB text, query name, query text.
		Tf.idf of words and ngrams
	Position	Query name appears early in KB text
	Genre	Genre of the query text (newswire, blog, …)
	Local Context	Lexical and part-of-speech tags of context words
Entity Context	Type	Query entity type, subtype
	Relation	Entities co-occurred, or involved in some attributes/relations/events with the query
	Coreference	Coreference links between mentions in the source document and the KB text
Profile		Slot fills of the query, KB attributes
Concept		Ontology extracted from KB text
Topic		Topics (identity and lexical similarity) for the query text and KB text
KB Link Mining		Attributes extracted from the hyperlink graphs (in-links, out-links) of the KB article
Popularity	Web	Top KB text ranked by search engine and its length
	Frequency	Frequency in KB texts

Table 7: A Road Map of Typical Ranking Features in Mono-lingual Entity Linking

$$ambiguity = \frac{\#name\ strings\ referring\ to\ more\ than\ one\ cluster}{\#name\ strings}$$

$$variety = \frac{\#clusters\ expressed\ by\ more\ than\ one\ name\ string}{\#clusters}$$

Table 9 shows some statistics about ambiguity and variety for the queries in KBP 2010 and 2011 evaluations (750 persons, each year). It is worth noting that the number of *(KB+NIL)* is larger than the number of *All* since some queries (with the same name string) could result in either KB ids or NIL ids.

From Table 8 and Table 9 we can roughly estimate that KBP2011 includes more ambiguous Non-NIL entities than KBP2010. For example, KBP2011 evaluation data set includes a highly ambiguous query "University" which refers to "University of Minnesota".

Difficulty	Year	All	NIL	Non-NIL
Ambiguity	2010	12.9	9.3	5.7
	2011	13.1	7.1	12.1
Variety	2010	2.1	1.7	2.5
	2011	1.6	0.9	2.4

Table 9: Difficulty Measures in Mono-lingual Entity Linking Evaluation Data Sets (%)

Table 10 shows the comparison between the averaged human annotators and some top systems that participated in both KBP2010 and KBP2011[2]. For a fair comparison, we also asked these top systems to run their KBP2011 system on 2010 data set. From Table 10 we can see that KBP2011 systems perform generally worse on 2011 data set than 2010 data set (comparing the third column and the fourth column). However, comparing the performance on the same KBP2010 data set, we can see almost all systems achieved significant improvement in 2011 (comparing the second column and the third column). In fact, LCC

[2]the human annotator performance was tested on a subset of KBP2010 evaluation data including 200 queries

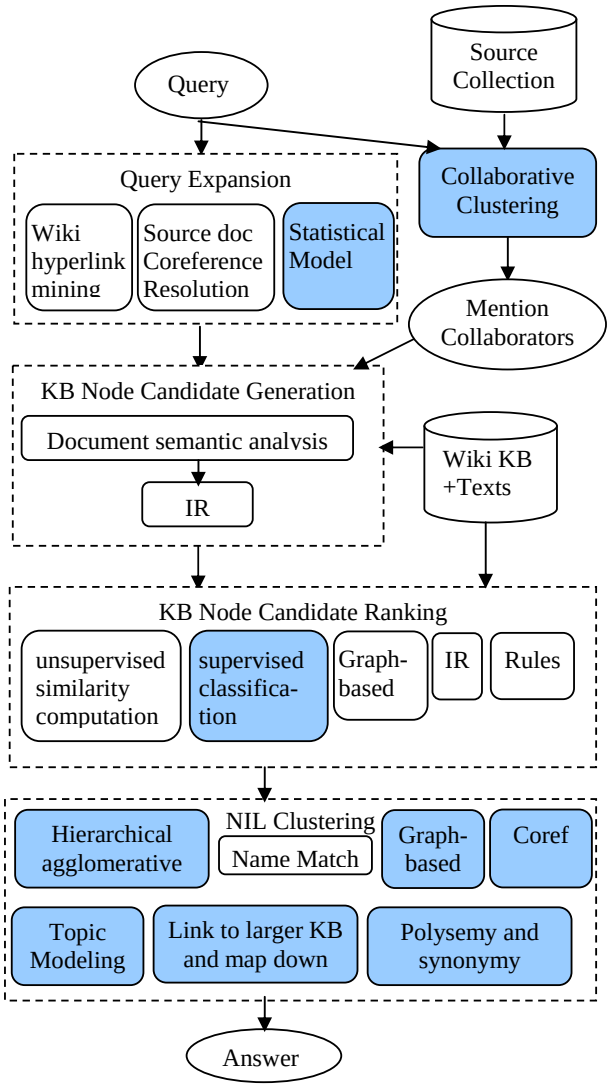

Figure 2: General Mono-lingual Entity Linking System Architecture

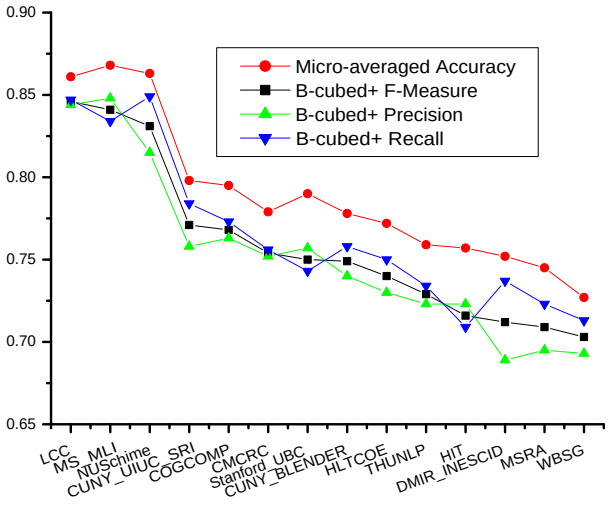

Figure 3: Performance of Top 14 Mono-lingual Regular Entity Linking Systems (B-cubed+ F-Measure above 70%)

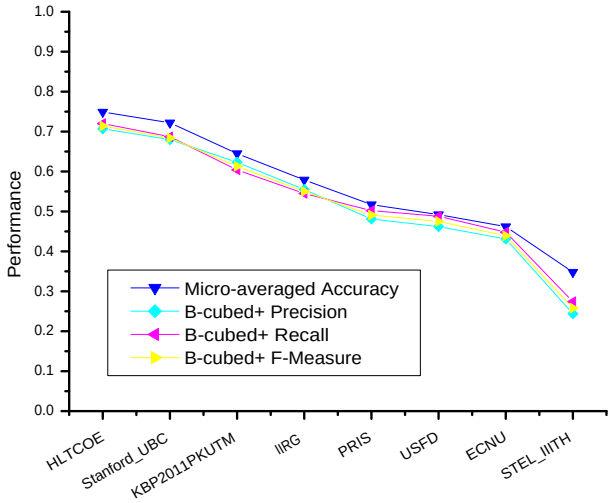

Figure 4: Performance of Mono-lingual Optional Entity Linking Systems

team closely approaches human annotators.

5.2.3 Performance of Various Entity Types

Figure 5 shows the F-measure scores of the top 14 systems on various entity types. We can see that systems generally performed the best on person entities, and the worst on geo-political entities. However the rank of overall performance is not consistent with the rank of individual types. For example, the top 3 system (NUSchime) achieved the best performance on person entities.

5.2.4 Performance of NIL and Non-NIL Clustering

Since NIL clustering is the main new component in 2011 Entity Linking, we also evaluate NIL clustering and Non-NIL clustering separately. It's not fair to use B-cubed+ metric to compare NIL clustering and Non-NIL clustering because it requires KB entry linking for Non-NIL clusters but not for NIL clusters. Therefore, we remove this requirement in B-cubed+ and use the regular B-cubed F-measure to measure clustering

Participants	2010 Systems /Human on 2010 Data	2011 Systems/ Human on 2010 Data	2011 Systems/ Human on 2011 Data
LCC	0.858	0.898(Estimated)	0.861
NUSchime	0.794	0.878	0.863
Stanford_ UBC	0.800	0.844	0.79
CUNY	0.693	0.834	0.778
CMCRC	0.819	0.844	0.779
HLT-COE	0.815	0.815	0.772
Human	0.902	0.902(Estimated)	-

Table 10: Entity Linking Micro-Averaged Accuracy Comparison across Systems and Human Annotators

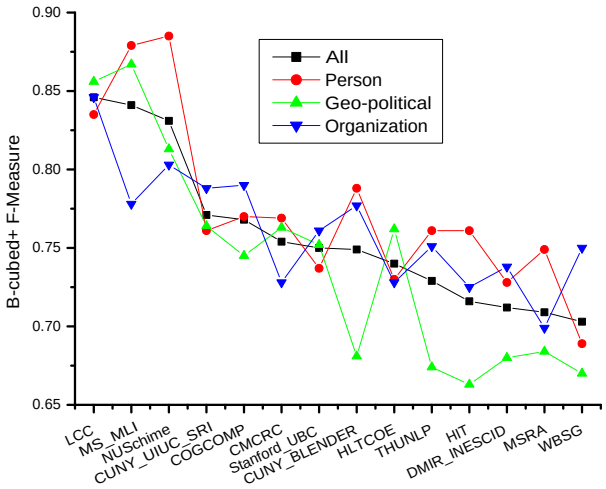

Figure 5: Mono-lingual Entity Linking Performance Comparison on Entity Types

performance. Figure 6 and 7 compares the B-cubed scores of NIL clustering and Non-NIL clustering from the top systems. We can see that the ranking of NIL clustering performance doesn't correlate with that of the overall performance, possibly because some systems focus more research efforts on NIL clustering than others. In addition, only 7.1% of the NIL queries are ambiguous, and there are not many name variants (many variation pairs only differ in capitalization forms such as "EL ALTO" vs. "El Alto", "CARMEL" vs. "Carmel") , so the baseline clustering algorithm based on name string matching or within-document coreference resolution(e.g. CUNY, COGCOMP, CMCRC (Radford et al., 2011)) can obtain reasonably high performance. Since NIL queries can be grouped based on different features, hierarchical clustering approaches that consisted of multiple steps were adopted by the top teams including LCC and NUSchime. For example, LCC system used a three-step process of grouping likely matches, clustering within those groups, and merging the final clusters.

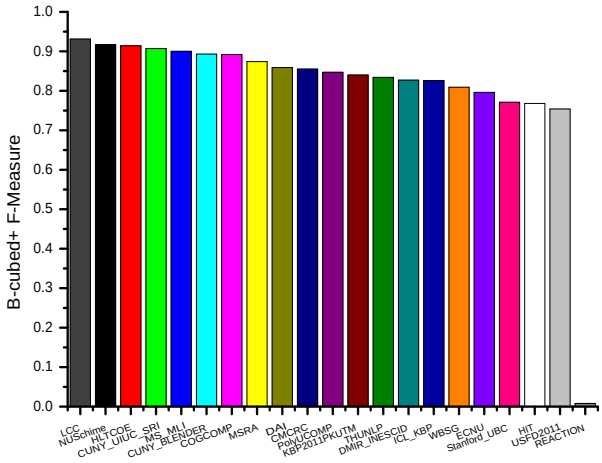

Figure 6: NIL Clustering Performance in Mono-lingual Entity Linking

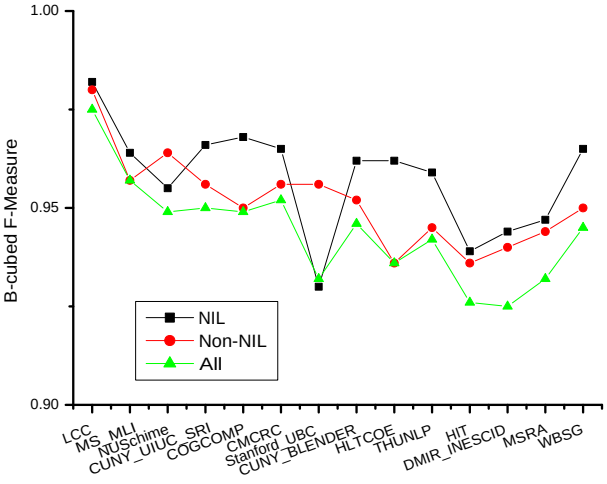

Figure 7: Mono-lingual Entity Linking Performance Comparison on NIL Clusters and Non-NIL Clusters

5.3 What's New and What Works

We have seen encouraging advances in mono-lingual entity linking since KBP was

launched in 2009. Many new and interesting approaches have been published by participants at major NLP conferences (Dredze et al., 2010; Zhang et al., 2010; Zheng et al., 2010; Chen and Ji, 2011; Gottipati and Jiang, 2011; Han and Sun, 2011; Ploch, 2011; Zhang et al., 2011; Zhang et al., 2011). In this section we will only highlight some new techniques that have been used by multiple systems in KBP2011.

5.3.1 Statistical Name Variant Expansion

Almost all entity linking systems exploit name variant expansion for the queries based on Wikipedia structure mining or coreference resolution in the background document. However, previous systems used rule-based methods so they cannot capture complicated acronyms such as swapped or missed acronym letters (e.g. "CCP" vs. "Communist Party of China"; "MD" vs. "Ministry of Defence"), multiple letters from expansion ("MINDEF" vs. "Ministry of Defence"). NUSchime system trained a statistical classifier to detect name variants and achieved 15.1% accuracy improvement over state-of-the-art acronym expansion methods. NLPR system (Zhang et al., 2011) and CUNY system mined acronyms using some common patterns such as "(A)" (e.g. "All Basotho Convention (ABC)").

5.3.2 Topic Features

Most context-based ranking features follow the distributional hypothesis (Harris, 1954), namely that queries sharing the same contexts tend to link to the same KB entry. If we consider one KB entry of a query as a certain "sense" of the query name, 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). An entity in a coherent latent topic tends to express the same sense and thus should be linked to one KB entry. For example, one query "*Li Na*" from a sports topic cluster represented by "{*tennis, player, Russia, final, single, gain, half, male, ...*}" and the other query "*Li Na*" from another politics topic cluster represented by "{*Pakistan, relation, express, vice president, country, Prime minister, ...*}" are likely referring to two different entities. In addition, each KB article can be considered as an entity-level semantic topic.

This general idea can be traced back to some initial attempts at using topic modeling for entity linking in KBP2009. Janya system (Srinivasan et al., 2009) represented a document as a vector of topic clusters, with each topic cluster including a group of topically-related common noun words. This year several entity linking systems exploited topic features for ranking candidates. NUSchime system (Zhang et al., 2011) treated Wikipedia as a document collection with multiple topical labels, and learned the posterior distribution over words for each topical label (Wikipedia category). MS_MLI system extracted topic from Wikipedia categories, Wikipedia list pages and lexicosyntactic pattern matches. Some other systems applied topic modeling results - both topic cluster IDs and topic vocabularies - as features. LCC system exploited a robust context modeling method which used common, low ambiguity topics extended from (Milne and Witten, 2008). DMIR_INESCID system used Latent Dirichlet Allocation (LDA) topic model to extract features including topic vector similarity, topic match and topic divergence (Anastacio et al., 2011). HLT-COE system used Wikipedia concept categories derived from Wikitology system (Syed et al., 2008) to assign topic features. CUNY-UIUC-SRI system (Cassidy et al., 2011) incorporated topic information in a more indirect way. They 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.

5.3.3 New Ranking Algorithms

All of the top systems in KBP2011 use supervised learning to rank KB node candidates. Support Vector Machines based Ranking (SVMRank) and Maximum Entropy (MaxEnt) based Ranking are still the most popular methods. In the meanwhile many new learning algorithms have been introduced this year such as Random Forests (e.g. THUNLP, DMIR_INESCID) and ListNet(CUNY). A nice summary of pros and cons about these ranking algorithms when applied to entity linking is presented in (Chen and Ji, 2011); they observed that ListNet achieved the best performance compared to

seven other ranking algorithms including SVMRank and MaxEnt using the same features and resources.

5.3.4 Query Classification

In KBP2010 and KBP2011, the number of queries is equally distributed in three entity types (persons, organizations and geo-political entities). However, as we can see from Figure 5, different entity types have quite different number of unique names. Therefore, it's likely to be beneficial to tune parameters (e.g. clustering thresholds) or train classifiers for each entity type separately. Many systems in KBP2011 (e.g. DMIR_INESCID) classified queries into three entity types first and trained models specifically for each entity type. CUNY system (Cassidy et al., 2011) found that entity context and profile based features significantly improved the performance of person and organization queries but hurt the performance of geo-political entities because global features are dominant for geo-political entities.

In addition, different queries may have very different number of KB node candidates according to their confusability. MSRA system classified queries into two types (single-candidate vs. multi-candidate) and then applied different models for these two types.

5.3.5 Go Beyond Single Query and Single KB Entry

Entity Linking has been designed in a "top-down" fashion, namely a set of target entities is provided to each system in order to simplify the task and its evaluation. However, in order to make entity linking systems more useful for some particular applications - such as assisting scientific paper reading by providing Wikipedia pages for any terminology that is referred to - a "bottom-up" task by identifying and linking all entities in the source documents will be more desirable. The introduction of NIL clustering component this year helps promote this extension. In fact, the prototype of UIUC system - "Wikification" (Ratinov et al., 2011) - aims to link all possible concepts to their corresponding Wikipedia pages, through combining local clues and global coherence of the joint cross-linking assignment by analyzing Wikipedia link structure and estimating pariwise article relatedness.

In KBP 2010, the WebTLab system and the CMCRC system extracted all entities in the context of a given query, and disambiguated all entities at the same time. Following this idea, MS_MLI system in 2011 conducted entity linking for all entities from each document, enriched knowledge base, and then reanalyzed the entities by using the new knowledge base adjusted to ensure the global feature consistency for the query entity. CUNY system (Chen and Ji, 2011; Cassidy et al., 2011) further extended this idea to cross-document level by constructing "collaborators" for each query and exploiting the global context from the entire collaborator cluster for each query.

5.4 Remaining Challenges

In our previous paper (Ji and Grishman, 2011) we summarized the remaining challenges in KBP2010 mono-lingual entity linking systems. Some of these challenges have been successfully addressed in KBP2011 systems. For example, in 2009 and 2010 most systems had to rely on Web access (e.g. ranking of Wikipedia pages from search engine) to estimate the popularity of a candidate KB node, while such submissions with Web access are not considered as official runs during the KBP evaluation. In contrast, in 2011 many systems have attempted some offline approaches to compute popularity. For example, (Han and Sun, 2011) computed popularity based on the distribution of the candidate KB entry name in a large corpus.

However, some other typical challenges still remain. For most systems, GPE is still the most difficult entity type. For some small location names (e.g. "*Del Rio*"), a system will need to acquire background knowledge (e.g. "*Del Rio*" is part of "*Gruene, Texas*") in order to disambiguate them. The introduction of the NIL clustering component in KBP2011 also brings some new challenges at identifying name variants. For example, in order to obtain the name variant pair of "*NU*" and "*Aviva*", a system will need to extract the org:alternative_names slot from the following text:

> "Hi there everbody, As I live in the Anglia Area it was on the news, so high flyng new director of **Aviva** stated the company was a global one, and

the business needs to reflect that, hence the possible name dropping of **Norwich Unio**. Or let's be honest, is that the reason, or more likely the **NU** name is being dragged through the mud over this."

We expect the idea of going beyond single queries and documents as described in 5.3.5 can help improve entity linking performance for these difficult cases.

6 Cross-lingual Entity Linking

6.1 General Architecture

There are two basic approaches to cross-lingual entity linking as depicted in Figure 8:

- **Pipeline A** (Name Translation and MT + English Entity Linking): Translate a Chinese query and its associated document into English, and then run English mono-lingual entity linking to link the translated query and document to English KB (such as HLT-COE system (McNamee et al., 2011) and CUNY baseline system (Cassidy et al., 2011)).
- **Pipeline B** (Chinese Entity Linking + Cross-lingual KB linkages): Apply Chinese Entity Linking to link a Chinese query to Chinese KB, and then use cross-lingual KB linkages to map the Chinese KB node to English KB node (such as LCC system (Monahan et al., 2011) and HITS system (Fahrni and Strube, 2011) that used external hyperlinks, image similarity and templates).

From the overall performance shown later in section 6.2.1 it's hard to tell which pipeline is better. Each method has its own limitations in terms of quality and portability. Pipeline A essentially converts the problem to mono-lingual entity linking and may suffer from name translation and document translation errors, while Pipeline B heavily relies on the existence of source language KB and thus is not easily adaptable to other low-density languages.

6.2 Evaluation Results

6.2.1 Overall Performance

The results of cross-lingual entity linking systems are summarized in Figure 9 and Figure 10.

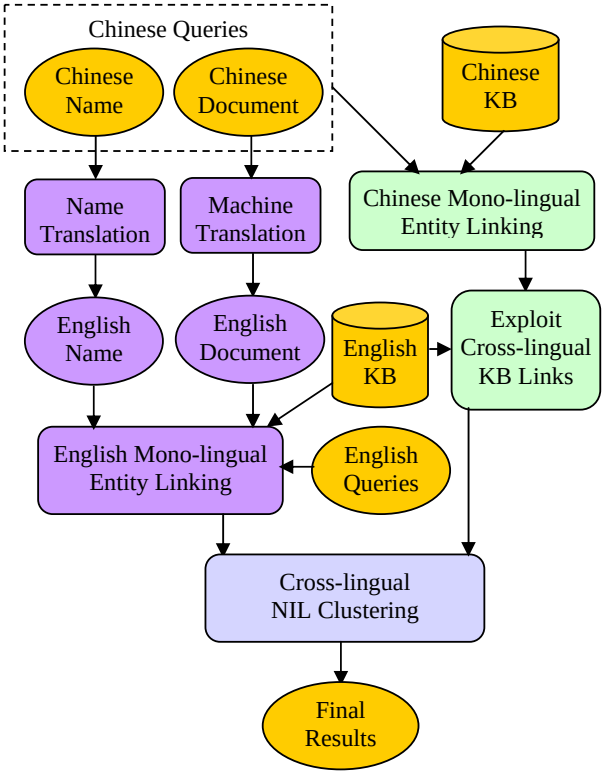

Figure 8: General Cross-lingual Entity Linking System Architecture

ECNU submitted both regular and optional runs, surprisingly their optional entity linking system without using Wikipedia documents achieved 4.9% higher F-measure than their regular system.

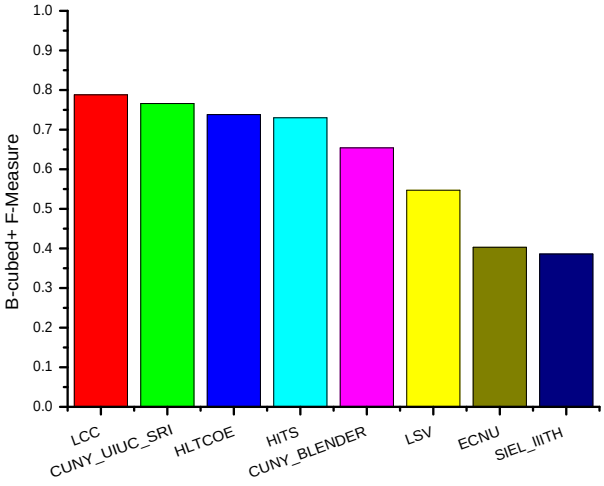

Figure 9: Performance of Cross-lingual Regular Entity Linking Systems

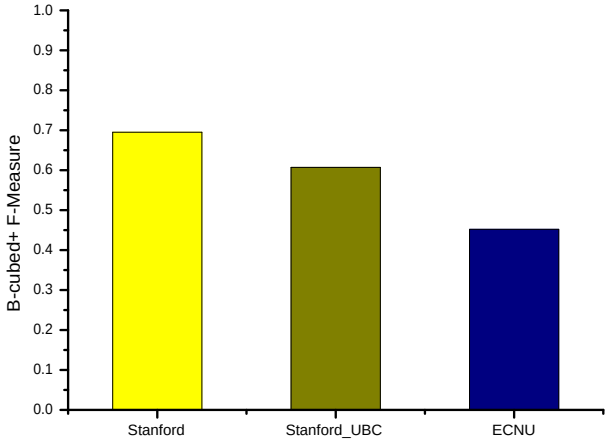

Figure 10: Performance of Cross-lingual Optional Entity Linking Systems

6.2.2 From Mono-lingual to Cross-lingual

This is the first year for KBP to add the cross-lingual track. Thanks to LDC's sufficient training data with good quality control, many teams were able to quickly develop a new system or adapt their mono-lingual system for the cross-lingual entity linking task. Table 11 shows the comparison of difficult query distributions in the training data sets. We can see that cross-lingual training data includes higher percentage of ambiguous names than mono-lingual data. We didn't conduct systematic inter-annotator agreement study as we did for mono-lingual entity linking last year (Ji et al., 2010), but we did notice a few annotation errors by checking some queries randomly. If time and funding permit in KBP2012, we should check human annotation performance and do better annotation quality control for this task.

Difficulty	Year	All	NIL	Non-NIL
Ambiguity	Mono-lingual	12.9	5.7	9.3
	Cross-lingual	20.9	14.0	28.6
Variety	Mono-lingual	2.1	2.5	1.7
	Cross-lingual	1.6	2.4	0.9

Table 11: Difficulty Measures in Entity Linking Training Data Sets (%)

It is worth investigating what kinds of challenges have been brought to entity linking because of language barriers. The top cross-lingual entity linking systems (LCC, CUNY_UIUC_SRI) can be ranked at top 4 and 5 in the mono-lingual track as shown in Figure 3, better than most mono-lingual entity linking systems. For fair comparison, we summarize the performance of Chinese queries and English queries separately in Figure 11, Figure 12, Figure 13 and Figure 14.

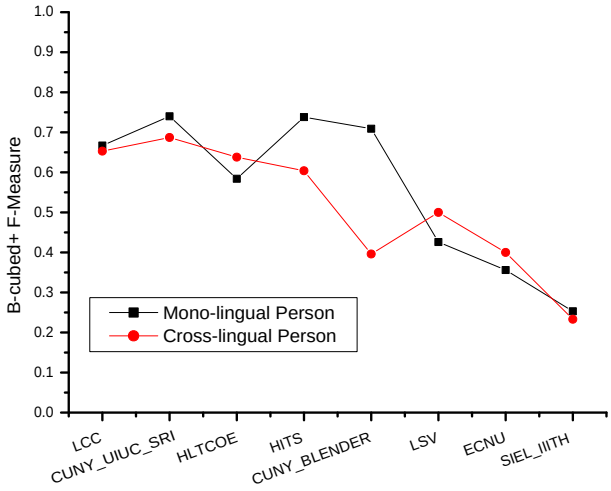

Figure 11: Performance Comparison of Mono-lingual and Cross-lingual Entity Linking (Persons)

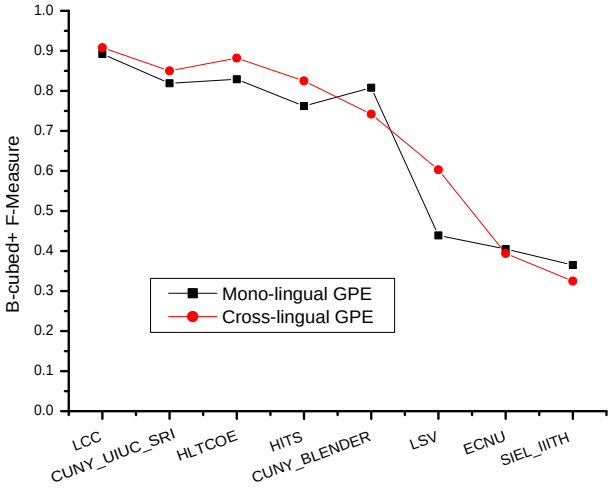

Figure 12: Performance Comparison of Mono-lingual and Cross-lingual Entity Linking (Geo-political)

There are many more Chinese queries (1481) than English queries (695), so the comparison is not completely fair. However, the scores can give us

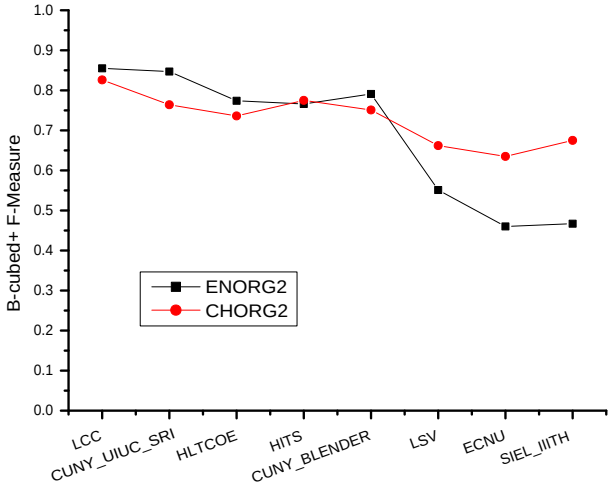

Figure 13: Performance Comparison of Mono-lingual and Cross-lingual Entity Linking (Organizations)

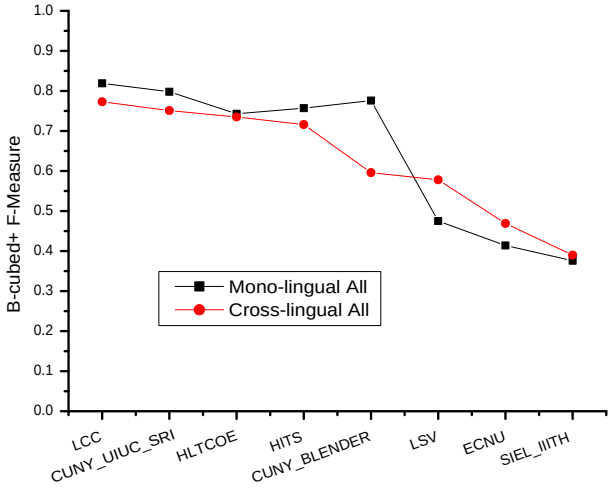

Figure 14: Performance Comparison of Mono-lingual and Cross-lingual Entity Linking (All)

some general idea about the level of cross-lingual entity linking compared to mono-lingual task. The mean score of cross-lingual queries is only 2.65% lower than that of mono-lingual queries.

We can see that for person entities, cross-lingual entity linking performs significantly worse than mono-lingual entity linking, mainly because the translation of person 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. In contrast, for organizations and geo-political entities, cross-lingual entity linking performance is very close to mono-lingual entity linking. Interestingly, in the mono-lingual entity linking evaluation, person performance is generally the best among three entity types, but for the English queries of cross-lingual entity linking task, English person queries generally have the worst performance. This may be caused by the imbalanced NIL and Non-NIL query distribution among English queries: only 25 of the 183 English person queries are NILs.

6.2.3 Performance of NIL and Non-NIL Clustering

Figure 15 compares the scores of NIL clustering and Non-NIL clustering in cross-lingual entity linking. There is a large gap between mono-lingual and cross-lingual NIL clustering performance, but the gap is much smaller for Non-NIL queries, possibly because in cross-lingual NIL clustering the systems can only exploit document-level information and translation as features. The detailed challenges about cross-lingual NIL clustering will be discussed later.

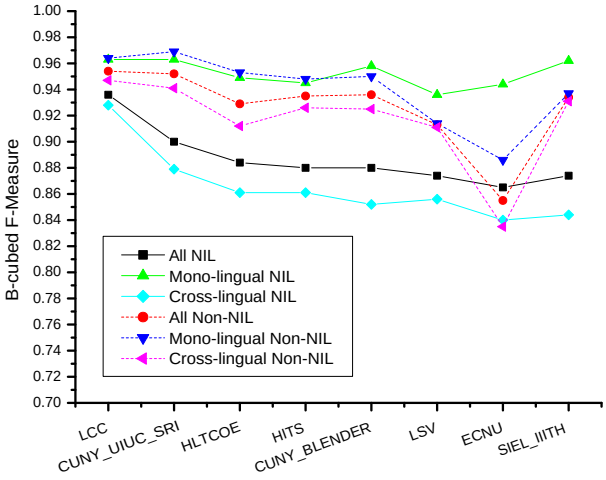

Figure 15: Cross-lingual Entity Linking Performance Comparison on NIL Clusters and Non-NIL Clusters

6.3 What's New and What Works

6.3.1 Entity Context Modeling

As we pointed out in section 6.1, there are some principle limitations of the basic approaches. In addition, some source language specific characteristics also require us to add more fine-grained contexts as ranking features. Therefore, in order to enhance both the portability and reduce the cost of cross-lingual entity linking, some teams have attempted to skip the steps of full MT and/or source language KB and cross-lingual Wikipedia linkages. 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 which country a city entity is located. Inspired from this intuition, CUNY constructed a cross-lingual name co-occurrence matrix for joint translating and disambiguating entities. Their 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 with the queries.

For each source document, the neighbor entities can be extracted from attribute and relation extraction. This approach is particularly effective to disambiguate entities with common organization names or person names. For example, many countries can have "Supreme Court" (in Japan, China, U.S., Macedonia, etc.) or "LDP (Liberty and Democracy Party)" (in Australia, Japan, etc.); "Newcastle University" can be located in UK or Australia; Many person entities share the same common names such as "Albert", "Pasha", etc.; "Ji county" can be located in "Shanxi" or "Tianjin". Also some person entities can share the same Chinese name spelling (e.g. "比利时") but different English name spelling ("Liang Tailong" vs. "Jonathan Leong") even if they come from different nationalities (China vs. Korea). However, not all entities have explicit profiles presented in the source documents. Table 12 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 entity with a common name "Shaofen Wan", if we know its associated document was from 1992 news, then it's likely refer to the member of "13th Central Committee of the Communist Party of China".

For example, three different entities with the same name spelling "阿尔伯特/Albert" can be disambiguated by their 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". For each Wikipedia article, the neighbor entities were generated from the incoming and outgoing links, as well as links in re-direct pages. DAI's mono-lingual entity linking system (Ploch, 2011) used similar ideas by encoding an entity context feature based on co-occurred entities and a link context feature based on Wikipedia incoming and outgoing links.

6.3.2 Name Translation

Query name translation is a crucial step in Pipeline A since an incorrect translation may lead the linking failure at the first step of English KB candidate generation. Some common name translation examples are as follows: "麦克金蒂/McGinty" is mistakenly translated into "Kim Jong-il"; "路易斯/Lewis" is mistakenly translated into "Luiz"; "莫科/Moco" is mistakenly translated into "Mo Kel". In these cases an English entity linking system will mistakenly link the query to a wrong KB entry. The top Pipeline A systems all benefited from hybrid name translation approaches in addition to basic machine translation. CUNY system applied a name translation system as described in (Ji et al., 2009) that includes four main components, a name tagger, translation lists, a transliteration engine, and a context-based ranker. HLT-COE system applied a progressive approach that includes dictionary lookup, Pinyinizaiton, and language-of-origin-specific transliteration.

6.4 Remaining Challenges

6.4.1 Difficulty Level Categorization

In Figure 16 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.

NIL singletons: About 7.6% queries are singleton entities (e.g. "中绿集团/Zhonglv Group", "丰华

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 12: Entity Context Examples

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

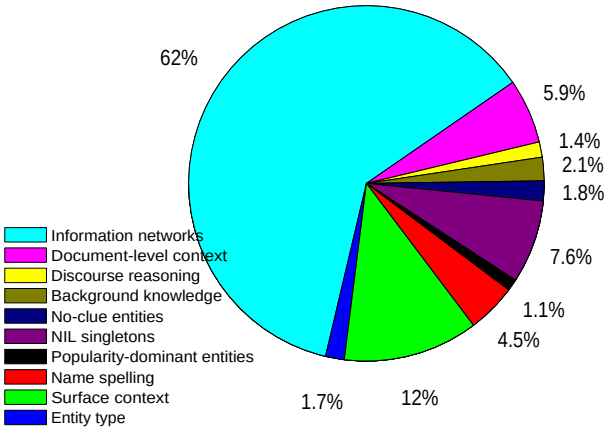

Figure 16: Distribution of 2011 CLEL queries according to difficulty levels

"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 12, 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 Person Name Translation

As discussed in section 6.2.2, person name translation is another important reason to degrade cross-lingual entity linking performance compared to the mono-lingual task. More advanced person name transliteration and re-ranking methods should be explored. Almost all systems in KBP2011 separated the processes of name translation and machine translation. A more integrated name-aware MT approach may be beneficial to enhance this task.

Figure 17 summarizes the distribution of 641 Chinese person queries in the KBP2011 CLEL evaluation corpus which need various translation techniques.

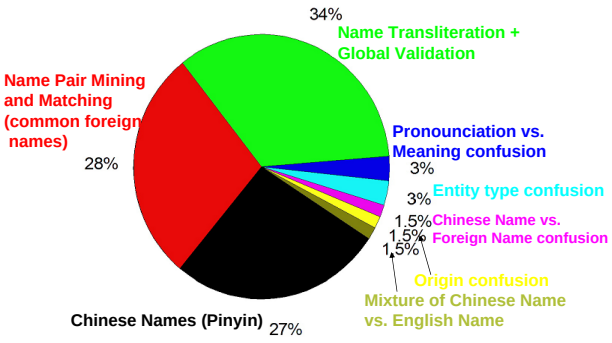

Figure 17: Distribution of person queries according to translation techniques

Among the 641 query names, only 27% are Chinese person names and can be translated by simple pinyin conversion (e.g. "王其江 (Wang Qijiang), 吴鹏(Wu Peng), 倪寿明 (Ni Shouming), 李娜 (Li Na)"). 28% are common foreign names that can be found in manually or automatically constructed bi-lingual name dictionaries (e.g. "伊莎贝拉 (Isabella), 斯诺(Snow), 戴夫(Dave), 林肯(Lincoln), 理查森 (Richardson) 亚当斯(Adams) 惠灵顿 (Wellington), 戴维斯(Davis), 理查(Richard)"). 34% uncommon foreign names may receive many different transliteration hypotheses (e.g. "皮耶" can be transliterated into "Piye", "Pierre", etc.) and so require global context validation to select the best hypothesis (e.g. the person's title, nationality). The remaining 10.5% names are extremely difficult ones. For example, 3% foreign names can represent common words (e.g. "拉索/Lasso" can mean "cable") and so it requires a system to decide whether to translate it based on pronunciation or meaning. 6% queries should receive different translations depending on their entity types (e.g. "魏玛" should be translated into "Weima" when it's a person and "Weimar" when it's a geo-political entity) or origins. Finally, some person names (especially person entities from Hongkong and Taiwan) are mixed with Chinese last names and English first names (e.g. "王菲/Wang Fei" should be translated into "Faye Wang").

6.4.3 Cross-lingual NIL 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 Regular Slot Filling

7.1 Approach Overview

A typical KBP2011 regular slot filling system architecture is depicted in Figure 18. Many of the systems used distant supervision to learn patterns or train classifiers, since little annotated training data was available but there are large data bases of closely corresponding relations. Several other approaches, including question-answering strategies and hand-coded rules, were also represented. The best results were obtained by combining a distant-supervision-trained classifier with selected hand-coded local rules. A lot of efforts have been attempted to enhance the basic distant supervision framework by techniques such as lexical pattern clustering (Jean-Louis et al., 2011) and coreference resolution.

7.2 Evaluation Results

7.2.1 Overall Performance

Figure 19 presents the performance of regular slot filling systems in KBP2011.

In comparison, the manual slot fills prepared by LDC and incorporated into the response pool, when scored against the assessment results, had a precision of 86%, recall of 73%, and F measure of 79%. Thus the gap between human and system performance remains large. An analysis of last year's slot filling data indicated that there was no single dominant cause of failure, but rather a number of contributing factors, including errors in named entity recognition, coreference analysis, recognition of implicit arguments, recognition of significant page layout (where relations are indicated by placement on the page rather than linguistic markers), and limited inference capabilities (Min and Grishman, 2012).

7.2.2 Comparison with KBP2010 and Human Annotators

An absolute comparison of this year's and last year's scores is difficult.[3] The measured recall is

[3]A comparsion with 2009 scores is even more problematic, since the systems were optimized for a different metric.

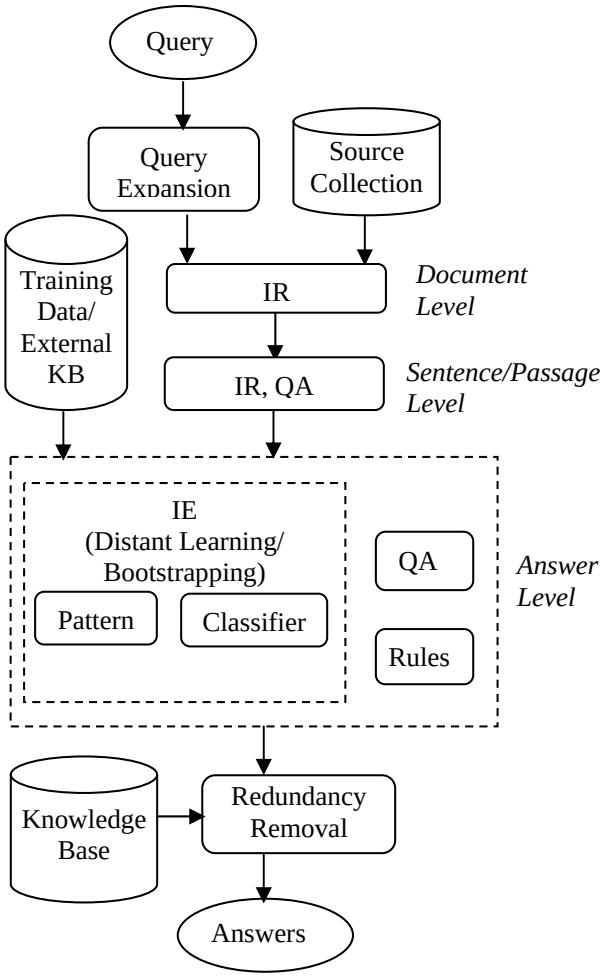

Figure 18: General Regular Slot Filling System Architecture

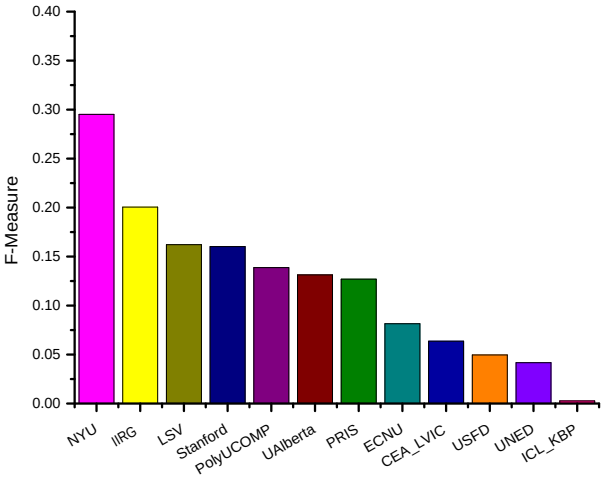

Figure 19: Performance of Regular Slot Filling Systems

relative to the set of correct responses in the pool, and so depends on the quality and quantity of the other systems in the pool in a given year (since the recall of current individual systems is quite low). Precision can be expected to be more consistent, but, as we have noted before, there are small revisions to the guidelines which will affect assessor judgments.

Nonetheless, some general comparisons are possible. The number of non-nil slot fills found during the manual annotation by LDC was almost the same for the two years (797 fills for 2010; 796 for 2011), suggesting that the basic difficulty of finding slot fills was similar. In contrast, the sizes of the sets of correct pooled responses was quite different: in 2010 there were 1057 equivalence classes filling 623 different slots, while in 2011 there were 953 equivalence classes filling 498 slots; this suggests that the 'competition' was stronger last year.

In addition we ran this year's top slot-filling system (NYU (Sun et al., 2011)) on last year's data. Even here the comparisons are somewhat problematic because last year's assessments are based on slightly different guidelines; furthermore, some system responses will be new (not in last year's pool) and will be judged by a different assessor. To reduce the number of responses for which new judgments were required, we used the scoring metric ignoring document IDs (see next section); even so, we had to judge 160 new responses. With that caveat, the performance on the two evaluation sets is shown in Table 13.

Evaluation Set	Recall	Precision	F-Measure
2010	22.4%	47.9%	30.5%
2011	27.2%	37.4%	31.5%

Table 13: NYU Regular Slot Filling System Performance

This suggests that this year's task is at least not significantly harder than last year's, and may be a bit easier.

7.2.3 Performance without Document Validation

Since some systems used the combined evidence from multiple documents (and sometimes multiple corpora) to extract slot fills and then find individual documents from the source collection to support the fills, it is worth checking the impact of the requirement to provide supporting documents. Therefore we removed the requirement of comparing supporting document IDs during scoring. More precisely, a triple [query, slot type, slot fill] is counted as correct if there is any document for which the matching entry in the pooled answer key is marked as correct. The results are presented in Figure 20. All systems obtained some gains with this metric; the largest gain (4.6%) was reported for the Stanford system, which relied on distant supervision and multiple text corpora.

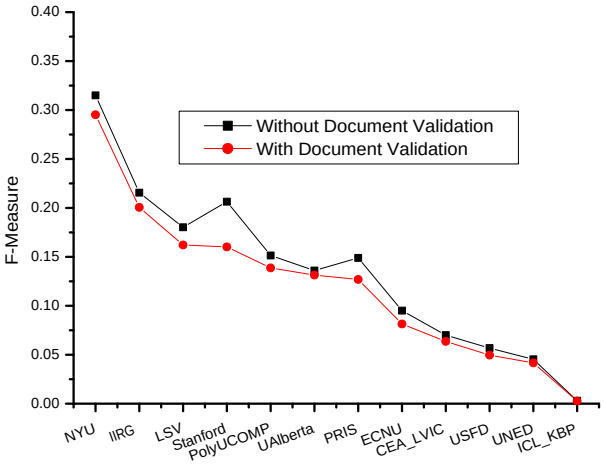

Figure 20: Impact of Document Validation

7.2.4 Slot-specific Analysis

Table 14 gives some basic statistics for the various slots for the evaluation corpus. Here 'entities' indicates how many of the 50 person queries and 50 organization queries had some non-NIL fill for this slot; 'values' indicates how many equivalence classes were found across all the entities. One can see from this table that the distribution of slot fills is quite skewed, being dominated by alternate names (10% + 4%) and slots associated with employment and membership (org:top_members/employees, per:employee_of, per:member_of; per:title; 12% + 7% + 4% + 21%). The latter are often very local relations, such as members of a single noun group ("Ford President Smith") with no explicit predicate or connective, so capturing such local relations is crucial to slot-filling success. Alternate names depend on

name coreference (e.g., recognizing acronyms), so this too is important to good performance.

slot	entities	values
org:alternate names	44	**98 (10%)**
org:city of headquarters	17	19 (1%)
org:country of headquarters	22	22 (2%)
org:dissolved	1	1 (0%)
org:founded	5	6 (0%)
org:founded_by	6	7 (0%)
org:member_of	8	11 (1%)
org:members	3	8 (0%)
org:number_of_ employees,members	6	6 (0%)
org:parents	17	24 (2%)
org:political, religious_affiliation	2	2 (0%)
org:shareholders	7	18 (1%)
org:stateorprovince _of_headquarters	16	17 (1%)
org:subsidiaries	17	**32 (3%)**
org:top_members, employees	40	**118 (12%)**
org:website	13	14 (1%)
per:age	15	16 (1%)
per:alternate_names	25	**46 (4%)**
per:cause_of_death	3	3 (0%)
per:charges	8	15 (1%)
per:children	8	17 (1%)
per:cities_of_residence	14	17 (1%)
per:city_of_birth	6	6 (0%)
per:city_of_death	1	1 (0%)
per:countries_of_residence	14	20 (2%)
per:country_of_birth	2	3 (0%)
per:country_of_death	1	1 (0%)
per:date_of_birth	3	3 (0%)
per:date_of_death	3	4 (0%)
per:employee_of	39	**71 (7%)**
per:member_of	17	**47 (4%)**
per:origin	18	23 (2%)
per:other_family	4	6 (0%)
per:parents	3	3 (0%)
per:religion	5	5 (0%)
per:schools_attended	10	16 (1%)
per:siblings	6	6 (0%)
per:spouse	7	8 (0%)
per:stateorprovince_of_birth	1	1 (0%)
per:stateorprovinces_of_ residence	11	11 (1%)
per:title	50	**201 (21%)**

Table 14: Statistics on slots in 2011 evaluation data

8 Temporal Slot Filling

8.1 Approach Overview

8.1.1 System Architecture

A typical KBP2011 full temporal slot filling system architecture is depicted in Figure 21. It starts with a regular slot filling component to extract slot fills for the given query. Interestingly, many queries (e.g. "George Bush", "Juan Carlos") in the KBP2011 temporal slot filling evaluation data set are ambiguous, which requires a temporal slot filling system to include entity disambiguation during selecting relevant documents. Then the system should apply document retrieval again to retrieve relevant documents based on the query and slot fills. Sentence retrieval should consider not only content relevance but also time-richness, namely that the sentence should include the query, slot fills, as well as some candidate time expressions. Then the remaining processing can be decomposed into two problems: (1) the classification of any temporal expression in the contexts of a query and its slot fills; and (2) temporal aggregation.

8.1.2 Temporal Classification

Temporal classification is applied to label temporal expressions that appear in the context of a particular entity and the slot value as one of the following temporal classes: 'BEGINNING', 'ENDING', 'WITHIN', 'RANGE' or 'NONE'.

Suppose the query entity is *Smith*, the slot type is *per:title*, and the slot-fill is *Chairman*, the following is a description of each class along with the corresponding four-tuple representation, assuming the document creation time is January 1st. 2001:

BEGINNING $< t_a, t_b, t_a, \infty >$

The temporal expression describes the beginning of the slot fill.

E.g. *Smith, who was named chairman two years ago*

$<$ 1999-01-01, 1999-01-01, 1999-01-01, $\infty >$

ENDING $< -\infty, t_b, t_a, t_b >$

The temporal expression describes the end of the slot.

E.g. *Smith, who resigned last October*

$< -\infty,$ 2000-10-01, 2000-10-01, 2000-10-31 $>$

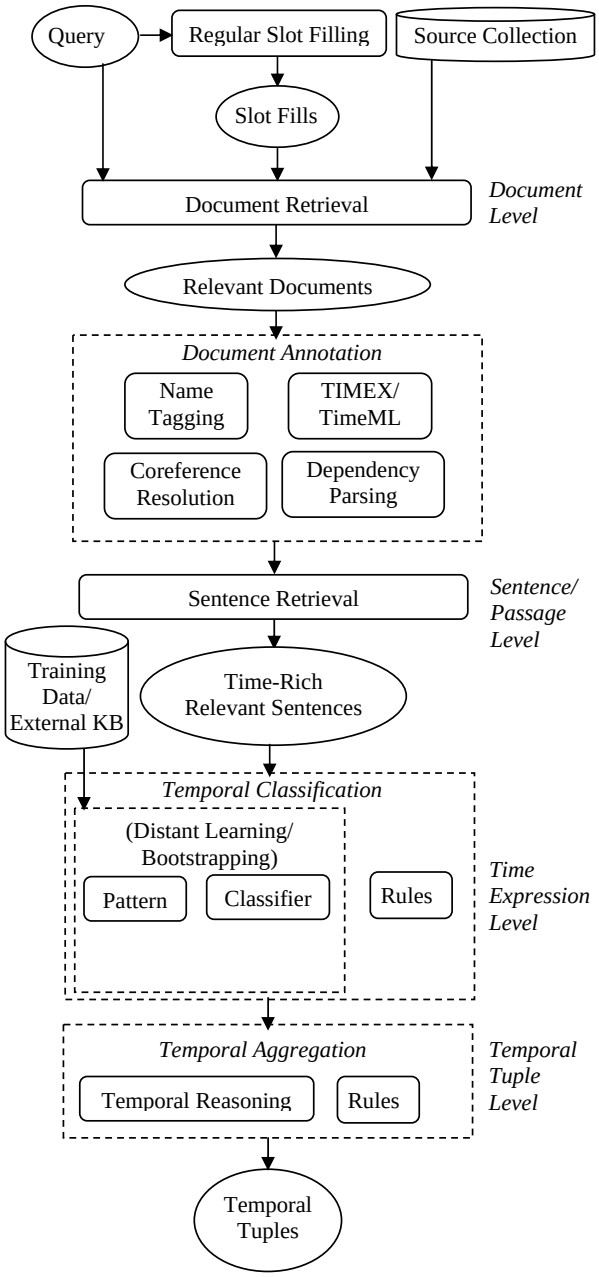

Figure 21: General Temporal Slot Filling System Architecture

WITHIN $< -\infty, t_b, t_a, \infty >$

The temporal expression describes a time at which the slot fill is valid.

E.g. *Chairman Smith*

$< -\infty, 2001\text{-}01\text{-}01, 2001\text{-}01\text{-}01, \infty >$

RANGE $< t_a, t_a, t_b, t_b >$

The temporal expression describes a range in which the slot fill is valid.

E.g. *Smith served as chairman for 7 years before leaving in 1991*

$< 1984\text{-}01\text{-}01, 1984\text{-}12\text{-}31, 1991\text{-}01\text{-}01, 1991\text{-}12\text{-}31 >$

NONE $< -\infty, \infty, -\infty, \infty >$

The temporal expression is unrelated to the slot fill.

E.g. *Last Sunday Smith had a party with his friends*

$< -\infty, \infty, -\infty, \infty >$

Various classification approaches have been explored, including SVM and kernel methods (CUNY system and UNED system), pattern matching (IIRG system) and heuristic rules (Stanford system and USFD system).

8.1.3 Temporal Aggregation

The 4-tuple representation provides a convenient way to do rule-based temporal aggregation. In order to produce the final 4-tuple for each entity/slot value pair, a system can sort the set of the corresponding classified temporal expressions according to the classifier's prediction confidence. We can initialize a 4-tuple to $< -\infty, +\infty, -\infty, +\infty >$ and then iterate through that set, aggregating at each point the temporal information as indicated by the predicted label (see Section 8.1.2). Given two four-tuple T and T', the following equation can be used for aggregation:

$$T \wedge T' =$$
$$< max(t_1, t'_1), min(t_2, t'_2), max(t_3, t'_3), min(t_4, t'_4) >$$

At each step the tuple is modified only if the result is consistent (i.e. $t_1 \leq t_2$, $t_3 \leq t_4$, and $t_1 \leq t_4$).

8.2 Evaluation Results

Preliminary scores were computed for the temporal slot-filling systems in the diagnostic task, using preliminary temporal annotations of slot fills found through manual search at the LDC (the annotations contain known errors, so scores may change after additional quality control). For comparison we also include three baselines: (1). DCT-WITHIN: use the document creation time (DCT) as 'WITHIN' for each query; (2). SENT-WITHIN: if there is any time expression in the context sentence of the query and slot fill (with entity coreference resolution)

then label it as 'WITHIN', otherwise label DCT as 'WITHIN'. (3). SENT-NONE: use 'INFINITY' for each tuple element.

8.2.1 Diagnostic System Performance

Figure 22 presents the performance of temporal slot filling systems in the diagnostic task. This year, the queries were selected from time-rich documents - there are 366 WITHIN labels, 230 BEGINNING labels and 119 ENDING labels for 100 queries. Therefore it's difficult to beat these three baselines. USFD system didn't beat the DCT-WITHIN baseline, while only CUNY and IIRG systems outperformed the SENT-WITHIN baseline.

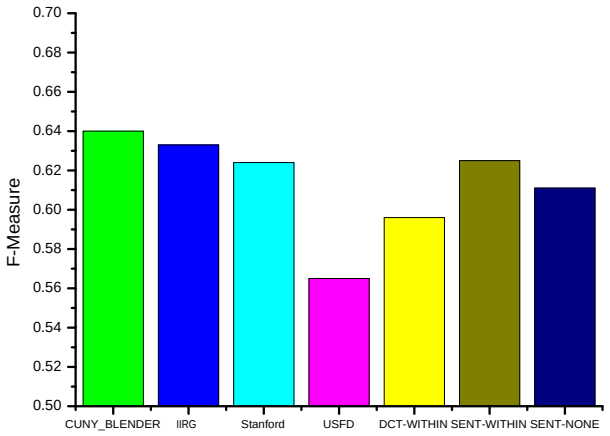

Figure 22: Performance of Temporal Slot Filling Systems in Diagnostic Task

8.2.2 Full System Performance

Figure 23 presents the performance of temporal slot filling systems in the full task. Compared to the baselines, full task performance is more promising than diagnostic task. Using CUNY regular slot filling system to detect the slot fills, CUNY's full temporal system achieved 16.7% higher F-measure than DCT-WITHIN baseline and 14.4% higher F-measure than SENT-WITHIN baseline.

Figure 24 shows that the performance of regular slot filling generally correlates with full temporal slot filling for each system. Not surprisingly, it indicates that a major error source of full temporal slot filling is the noise produced by regular slot filling. In general each system obtained much worse performance on the full task than the diagnostic task

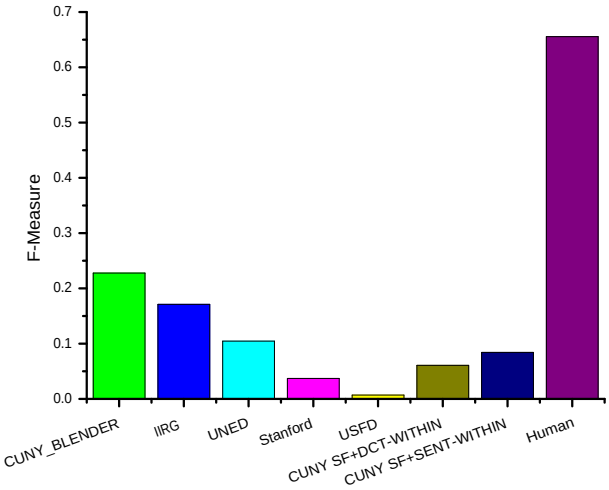

Figure 23: Performance of Temporal Slot Filling Systems and Human Annotators in Full Task

because of regular slot filling errors.

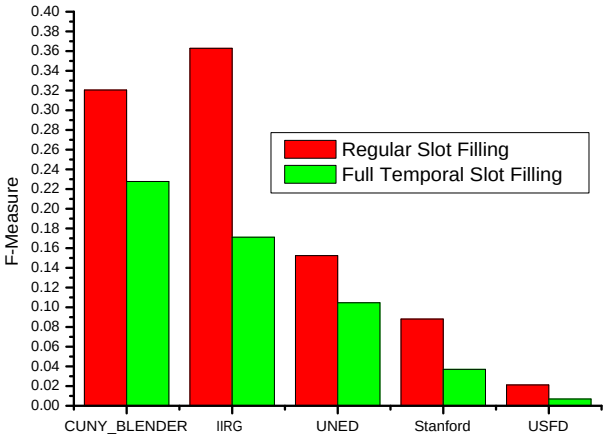

Figure 24: Impact of Regular Slot Filling on Full Temporal Slot Filling

8.2.3 Performance Comparison on Slot Types

Figure 25 and 26 present the performance of temporal slot filling systems on various slot types.

In the diagnostic task, all systems achieved much better results on "cities_of_residence" than other slot types, because one evaluation document includes a lot of short lists from which the queries and temporal 'WITHIN' answers are extracted:

> "EUROPE DMITRI MEDVEDEV Prime minister of Russia 42 Moscow, Russia

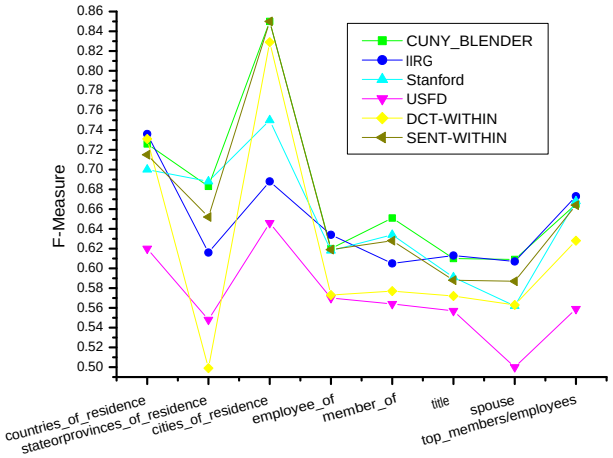

Figure 25: Performance of Diagnostic Temporal Slot Filling Systems on Various Slot Types

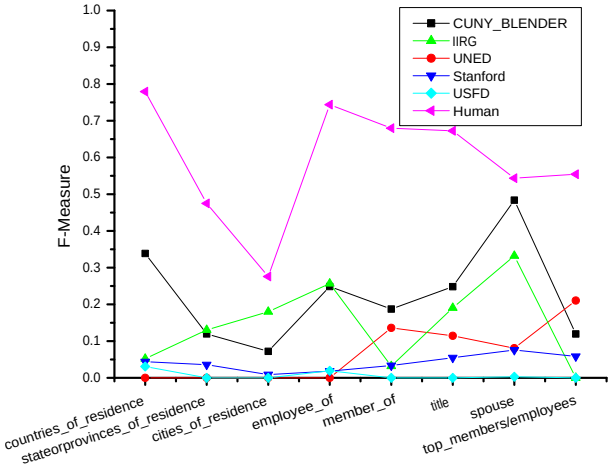

Figure 26: Performance of Full Temporal Slot Filling Systems on Various Slot Types

...
NICOLAS SARKOZY President of France 52 Paris, France
...
LEWIS HAMILTON Racecar driver 22 Stevenage, England
...
JAVIER BARDEM Actor 38 Madrid, Spain
...
LIONEL MESSI Soccer player 20 Barcelona, Spain
...
ASIA PERVEZ MUSHARRAF President of Pakistan 64 Islamabad,Pakistan
...
ZHOU XIAOCHUAN Governor of China's central bank 59 Beijing, China
...".

It is also problematic to choose training documents which include a lot of short sentences with very rich temporal information and simple patterns. Such documents are not representative of the entire source collection and not useful for learning new features:

"Tom LaSorda, president and CEO, Sept. 2005-Aug. 2007
Dieter Zetsche, president and CEO, Nov. 2000- Sept. 2005
James P. Holden, president and CEO, Oct. 1999-Nov. 2000
Thomas T. Stallkamp, president, Jan. 1998-Dec. 1999
...
Lee A. Iacocca, chairman and CEO, Sept. 1979-Dec. 1992 (president from Nov. 1978-Sept. 1979)
Eugene A. Cafiero, president, Oct. 1975-Nov. 1978
John J. Riccardo, chairman, Oct. 1975-Sept. 1979 (president, Jan. 1970-Oct. 1975)

In general, due to the limited time for annotation, for the diagnostic task KBP2011 temporal slot filling training and evaluation documents include a much higher concentration of explicit time information than regular newswire documents, which makes it very difficult to beat the WITHIN baselines as described in the above subsection. In the future, we aim to choose more representative documents for the diagnostic task of temporal slot filling.

For the full task, some systems got zero scores for some slot types. In general top system performance is worse than human annotators, except for "cities_of_residence" and "spouse" slots.

8.3 Discussions

CUNY system description paper (Artiles et al., 2011) provides a comprehensive analysis of

the successful techniques as well as remaining challenges for temporal slot filling task. In the following we only outline those important aspects which are not specific to their system:

8.3.1 What Works

Enhance Distant Supervision through Rich Annotations: As in regular slot filling, given the expensive nature of human-assessed training data for the TSF task, CUNY and Stanford systems adapted a distant supervision (Mintz et al., 2009) approach to obtain large amounts of training data from the Web without human intervention. In a TSF task, we need to annotate relations among three elements instead of two: query entity, slot fill and time expression. In order to reduce uncertainty in temporal information projection, they introduced rich annotations including name tagging, entity coreference resolution, time expression identification and normalization, dependency parsing into the distant supervision process.

Multi-level Reference Time Extraction: Add fine-grained reference date from sentence-level or sub-sentential level in addition to document creation time.

Combining Flat and Structured Approaches: For many NLP tasks including this new TSF task, one main challenge lies in capturing long contexts. Semantic analysis such as dependency parsing can make unstructured data more structured by *compressing* long contexts and thus reduce ambiguities. However, current core NLP annotation tools such as dependency parsing and coreference resolution are not yet ideal for real applications. The deeper the representation is, the more risk we have to introduce annotation errors. Therefore TSF systems can benefit from a more conservative approach combining benefits from both flat approach (local context, short dependency path, etc.) and structured approach (e.g. dependency path kernel).

8.3.2 Remaining Challenges

This is the first year for the temporal slot filling task, and we have observed reasonable success. However, a lot of challenges remain. In the following we summarize the significant challenges and suggest some research directions for the next years.

Implicit and Wide Contexts: In many cases the temporal information is implicitly represented and thus a system is required to capture deep semantic knowledge.

Coreference Resolution: Like in other IE tasks, coreference resolution is another bottleneck. TSF performance suffers from coreference errors of all types of entity mentions (names, nominals and pronouns).

Temporal Reasoning: As in regular slot filling (Ji and Grishman, 2011), inferences are required to extract temporal information for the remaining difficult cases. We can roughly categorize them into the following types. Some previous work in TempEval (Yoshikawa et al., 2009; Tatu and Srikanth, 2008; Ling and Weld, 2010) or ACE temporal information extraction (Gupta and Ji, 2009) conducted temporal reasoning; but all of them focused on single-document extraction. Some TSF systems such as CUNY did shallow reasoning such as propagation of temporal information from cities_of_residence to stateorprovinces_of_residence. The TSF task in KBP2011 was designed as a top-down question answering task, by sending one entity query and one slot fill each time. However, various entities (both queries and non-queries) and their attributes are often inter-dependent and thus their temporal boundaries can be used to infer from each other and ensure consistency.

Distant Supervision: Distant supervision methods have achieved some reasonable success for this new TSF task, but some significant challenges remain. In some cases when the following assumptions are invalid temporal reasoning or centroid entity detection are needed: "One sense per query", "One query per context" and "One sentence per query".

"Long Tail" Problem: The final challenge lies in the long-tailed distribution of temporal context patterns - a high percentage that match a few instances, plus a few other patterns that match many instances. Dependency parsing can filter out some irrelevant contexts but it may require deeper understanding to generalize the diverse lexical contexts. For example, the starting date of an employment relation can be expressed by many long-tail patterns such as "would join", "would be appointed", "will start at", "went to work", "was

transferred to", "was recruited by", "took over as", "succeeded PERSON", "began to teach piano", etc.

9 Something Old and Something New

For this year's KBP evaluation, we had four major tasks, representing a mix of evaluation continuity and ambitious pilot efforts.

Mono-lingual Entity Linking The entity linking task of the last two years was extended to include NIL clustering. System performance on the basic task has continued to improve, and the best systems are approaching human performance. Approaches to linking are observed to be converging. NIL clustering was also successful, although most cases in this year's evaluation could be handled by string matching alone. In the future, more challenging queries might be desired for NIL clustering. Another promising direction is to extend the task to more informal genres.

Cross-lingual Entity Linking This represented a first effort at introducing a cross-lingual component into TAC. Overall performance for the cross-lingual task proved only slightly lower than for the mono-lingual task, although linking person names proved particularly challenging, as did NIL clustering. As was the case for the monolingual task, we can expect cross-lingual systems to mature if the task is repeated next year. To assist the further development of this task, in the future we may want to provide more resources for Person name translation, and more training data for NIL clustering. It will be also interesting to extend this task to new genres and new foreign languages.

Slot Filling

This year's slot filling evaluation represented an effort at continuity; the task guidelines were clarified but otherwise the task remained unchanged. A number of new players got introduced to the slot filling task, but there was no progress in overall system performance. It remains difficult to achieve F-measure higher than 30%. Reaching competitive performance on this task requires a fairly mature NLP system, such as high-quality name tagging, coreference resolution and syntactic analysis. Compared to previous IE evaluation campaigns, KBP slot filling also requires certain degrees of cross-document cross-entity cross-slot inference techniques which are usually not available in existing IE systems. Such requirements make it harder to evaluate more exotic approaches. Therefore the entry cost for this task remains high.

Several participants made use of distant supervision, offering the potential for interesting comparisons of strategies for addressing some of the problems of distant supervision.

It is not clear what would be required to achieve significant progress in performance. Error analysis indicates that the failures are scattered across a variety of analysis modules in the IE pipeline, thus requiring module-by-module improvement, possibly complemented by (more) joint inference – potentially a large system development effort. Encouraging research focus on fewer productive slots and providing richer annotation of training data might help.

Temporal Slot Filling Temporal slot filling was the most ambitious effort this year, layering significant new difficulties on top of an already difficult task (regular slot filling). The effort alone to prepare training and evaluation data for this task – selecting suitable documents and annotating them – stretched the limits of TAC and LDC capabilities and will need to be revisited after the workshop. So will the scoring metric. Although participating systems had difficulty beating relatively simple baselines in the diagnostic task due to the unrepresentative data collection, the participation yielded a much better understanding of the problems involved, such as documents with multiple and nested reference times, added burdens on coreference, and the need for particular types of temporal reasoning. In the future the approach to select representative queries and documents for the diagnostic task needs to be significantly improved. More efficient ways of manual annotations need to be investigated in order to reduce burden of assessment and evaluation.

References

James F. Allen. 1983. Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, November 1983, Volume 26, Number 11, pp. 832-843.

Ivo Anastacio, Bruno Martins and Pavel Calado. 2011. Supervised Learning for Linking Named Entities

to Knowledge Base Entries. *Proc. Text Analysis Conference (TAC2011)*.

Javier Artiles, Qi Li, Taylor Cassidy, Suzanne Tamang and Heng Ji. 2011. CUNY_BLENDER TAC-KBP2011 Temporal Slot Filling System Description. *Proc. Text Analysis Conference (TAC2011)*.

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

Chitta Baral, Gregory Gelfond, Michael Gelfond and Richard B. Scherl. 2005. Textual Inference by Combining Multiple Logic Programming Paradigms. *Proc. AAAI 2005 Workshop on Inference for Textual Question Answering*.

L. Bentivogli, P. Clark, I. Dagan, H.T. Dang and D. Giampiccolo. 2010. The Sixth PASCAL Recognizing Textual Entailment Challenge. *Proc. Text Analysis Conference (TAC2010)* .

Taylor Cassidy, Zheng Chen, Javier Artiles, Heng Ji, Hongbo Deng, Lev-Arie Ratinov, Jing Zheng, Jiawei Han and Dan Roth. 2010. CUNY-UIUC-SRI TAC-KBP2011 Entity Linking System Description. *Proc. Text Analysis Conference (TAC2011)* .

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

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber and Tim Finin. 2010. Entity Disambiguation for Knowledge Base Population. *Proc. NAACL2010*.

Noemie Elhadad, Regina Barzilay and Kathleen McKeown. 2002. Inferring Strategies for Sentence Ordering in Multidocument Summarization. *JAIR, 17:35-55*.

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

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

Swapna Gottipati and Jing Jiang. 2011. Linking Entities to a Knowledge Base with Query Expansion. *Proc. EMNLP2011*.

Prashant Gupta and Heng Ji. 2009. Predicting Unknown Time Arguments based on Cross-event Propagation. *Proc. ACL-IJCNLP 2009*.

Xianpei Han and Le Sun. 2011. A Generative Entity-Mention Model for Linking Entities with Knowledge Base. *Proc. ACL2011*.

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

Ludovic Jean-Louis, Romaric Resancon, Olivier Ferret and Wei Wang. 2011. Using a Weakly Supervised Approach and Lexical Patterns for the KBP Slot Filling Task. *Proc. TAC2011*.

Heng Ji and Ralph Grishman. 2011. Knowledge Base Population: Successful Approaches and Challenges. *Proc. ACL2011*.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt and Joe Ellis. 2010. An Overview of the TAC2010 Knowledge Base Population Track. *Proc. Text Analytics Conference (TAC2010)*.

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*.

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

Xuansong Li, Joe Ellis, Kira Griffitt, Stephanie M. Strassel, Robert Parker and Jonathan Wright 2011. Linguistic Resources for 2011 Knowledge Base Population Evaluation. *Proc. TAC2011*.

Xiao Ling and Daniel S. Weld. 2010. Temporal Information Extraction. *Proceedings of the Twenty Fifth National Conference on Artificial Intelligence*.

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*.

Bonan Min and Ralph Grishman. 2012. Challenges in the TAC-KBP Slot Filling Task. *Proc. 8th International Conf. on Language Resources and Evaluation*.

David Milne and Lan H. Witten. 2008. Learning to Link with Wikipedia. *Proc. CIKM2008*.

Mike Mintz, Steven Bills, Rion Snow and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction without Labeled Data. *Proc. ACL-IJCNLP2009*.

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

Danuta Ploch. 2011. Exploring Entity Relations for Named Entity Disambiguation. *Proc. ACL2011*.

J. Pustejovsky, P.Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B.Sundheim, D. Day and L. Ferro andmar M. Lazo. 2003. The Timebank Corpus. *Corpus Linguistics. pp. 647-656*.

James Pustejovsky and Marc Verhagen. 2009. Semeval-2010 Task 13: Evaluating Events, Time Expressions, and Temporal Relations (Tempeval-2). *Proc. of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.

Will Radford, Ben Hachey, Matthew Honnibal, Joel Nothman and James R. Curran. 2011. Naive but Effective NIL Clustering Baselines - CMCRC at TAC2011. *Proc. TAC2011*.

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

Steven Schockaert, Martine De Cock, David Ahn and Etienne Kerre. 2006. Supporting Temporal Question Answering: Strategies for Offline Data Collection. *Proc. 5th International Workshop on Inference in Computational Semantics (ICoS-5)*.

R. Snodgrass. 1998. Of Duplicates and Septuplets. *Database Programming and Design*.

Harish Srinivasan, John Chen and Rohini Srihari. 2009. Cross Document Person Name Disambiguation Using Entity Profiles. *Proc. Text Analysis Conference (TAC2009)*.

Ang Sun, Ralph Grishman, Bonan Min and Wei Xu. 2011. NYU 2011 System for KBP Slot Filling. *Proc. Text Analysis Conference (TAC2011)*.

Zareen Syed, Tim Finin, and Anupam Joshi. 2008. Wikipedia as an Ontology for Describing Documents. *Proc. the Second International Conference on Weblogs and Social Media*.

Marta Tatu and Munirathnam Srikanth. 2008. Experiments with Reasoning for Temporal Relations between Events. *Proc. COLING2008*.

Marc. Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz and James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval Temporal Relation Identification. *Proc. of ACL 2007 workshop on SemEval*.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly Identifying Temporal Relations with Markov Logic. *Proc. ACL-IJCNLP2009*.

Tao Zhang, Kang Liu and Jun Zhao. 2011. The NLPR_TAC Entity Linking System at TAC2011. *Proc. TAC2011*.

Wei Zhang, Jian Su, Chew Lim Tan and Wen Ting Wang. 2010. Entity Linking Leveraging Automatically Generated Annotation. *Proc. COLING2010*.

Wei Zhang, Jian Su and Chew-Lim Tan. 2011. A Wikipedia-LDA Model for Entity Linking with Batch Size Changing Instance Selection. *Proc. IJCNLP 2011*.

Wei Zhang, Yan Chuan Sim, Jian Su and Chew Lim Tan. 2011. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling. *Proc. IJCAI 2011*.

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