Overview of the English Slot Filling Track at the TAC2014 Knowledge Base Population Evaluation

Mihai Surdeanu[1], **Heng Ji**[2]

[1] School of Information: Science, Technology, and Arts
University of Arizona, Tucson, AZ, USA
msurdeanu@email.arizona.edu

[2] Computer Science Department
Rensselaer Polytechnic Institute, NY, USA
jih@rpi.edu

Abstract

We overview the English Slot Filling (SF) track of the TAC2014 Knowledge Base Population (KBP) evaluation. The goal of this KBP track is to promote research in the extraction of binary relations between named and numeric entities from free text. The main changes this year include: (a) the inclusion of ambiguous queries, i.e., queries that point to multiple real-life entities with the same name; (b) accepting outputs created through inference; and (c) a simplification of the task and of the input format by removing references to the knowledge base for the entities included in queries. The SF track attracted 31 registered teams, out of which 18 teams submitted at least one run. The highest score this year was 36.72 F1, with a median of 19.80 F1.

1 Introduction

The Knowledge Base Population (KBP) track at TAC 2014 aims to promote research on automated systems that discover information about named entities and incorporate this information in a knowledge source, or database. This effort can be seen as a natural continuation of previous conferences and evaluations, such as the Message Understanding Conference (MUC) (Grishman and Sundheim, 1996) and the Automatic Content Extraction (ACE) evaluations[1].

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

Within this larger effort, the slot filling (SF) subtask must extract the values of specified attributes (or *slots*) for a given entity from large collections of natural language texts. Examples of slots include age, birthplace, and spouse for a person or founder, top members, and website for organizations. This document focuses only on the English Slot Filling (SF) task. For the other tasks part of KBP 2014, please visit the KBP web page: http://www.nist.gov/tac/2014/KBP/.

This is the sixth year a SF evaluation takes place. This year, 31 teams registered, and 18 teams submitted results. These numbers are similar to 2013. More importantly, the approximately 50% retention rate highlights that many groups continue to find this task difficult.

The slot filling task at TAC-KBP 2014 follows closely the 2013 definition (Surdeanu, 2013). There are, however, three important changes that were implemented this year:

1. This year, a percentage of the queries contained entity names that are ambiguous across the document collection. For example, "Michael Jordan" may refer to the basketball player or the Berkeley machine learning professor. The goal of this exercise is to encourage participants to combine multiple KBP tasks, in this particular case, entity linking and slot filling.

2. This year we accepted outputs created through inference, provided it is justified in the KBP document collection. For example, a system could correctly infer the filler "per:country_of_birth=France" from

two texts (potentially appearing in two different documents): "He was born in Paris" and "Paris is the capital of France". To accommodate this change, the output format for SF changes this year. See Section 2 for details.

3. This year the input format of the evaluation queries was simplified: the queries no longer include links to the reference KB, or specify slots to ignore.

2 Task Definition

The goal of the SF is to collect information on certain attributes (or *slots*) of entities, which may be either persons or organizations. Guidelines for each of the slots are available at: `http://surdeanu.info/kbp2014/def.php`. Table 1 lists the slots for this year's SF evaluation, which are carried over from 2013. Note that for list-valued slots, fillers returned for the same entity and slot must refer to distinct individuals. It is not sufficient that the strings be distinct; for example, if a system finds both "William Jefferson Clinton" and "Bill Clinton" as fillers for the same entity and slot, it should return only one of those fillers (the other would be considered redundant and reduce system precision).

2.1 Input Format

This years input query format is close to the 2013 format, with two changes:

1. We removed the <nodeid> field, which links the input entity to the reference knowledge base. The reasoning behind this decision is to align the input formats between Slot Filling and Cold Start. Note that the entity can still be disambiguated using the provided docid and beginning/end offsets.

2. Because the link between the entity and the KB is no longer provided, the <ignore> field, which listed slots to be ignored during extraction because they were already populated in the KB, was also removed.

Thus, 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, and the start and end offsets of the name as it appears in the document (to disambiguate the query in case there are multiple entities with the same name). An example query is:

```
<query id="SF_002">
    <name>PhillyInquirer</name>
    <docid>eng-NG-31-141808-9966244</docid>
    <beg>757</beg>
    <end>770</end>
    <enttype>ORG</enttype>
</query>
```

2.2 Output Format

The new SF output format is driven by two observations:

1. It is designed to allow justifications that aggregate information from multiple different documents. This was not supported by the 2013 SF output format. However, the LDCs assessment guidelines did not change, other than accepting justifications coming from multiple documents: a slot filler is considered correct only if the justification unambiguously supports the extraction.

2. During the 2013 assessment process, LDC derived no benefit from having the entity and filler provenances (Columns 6 and 7 in the 2013 format). Thus, we simplified the requirements for provenance. We will still require the provenance for the relation itself (formerly Column 8 in the 2013 format) and a simplified form of filler provenance (see below).

Similar to 2013, the 2014 format requires that system output files be in UTF-8 and contain at least one response for each query-id/slot combination. A response consists of a single line, with a separate line for each slot value. Lines should have the seven tab-separated columns summarized in Table 2. For each query, the output file should contain exactly one line for each single-valued slot. For list-valued slots, the output file should contain a separate line for each list member. When no information is believed to be learnable for a slot, Column 4 should be NIL and Columns 5 through 7 should be left empty.

Relation Provenance

The provenance stored in Column 4 must contain text that justifies the extracted relation. That is, it must include some mention of the subject and object entities and some text supporting the slot/predicate that connects them. For example,

Person Slots			Organization Slots		
Name	**Type**	**List?**	**Name**	**Type**	**List?**
per:alternate_names	Name	Yes	org:alternate_names	Name	Yes
per:date_of_birth	Value		org:political_religious_affiliation	Name	Yes
per:age	Value		org:top_members_employees	Name	Yes
per:country_of_birth	Name		org:number_of_employees_members	Value	
per:stateorprovince_of_birth	Name		org:members	Name	Yes
per:city_of_birth	Name		org:member_of	Name	Yes
per:origin	Name	Yes	org:subsidiaries	Name	Yes
per:date_of_death	Value		org:parents	Name	Yes
per:country_of_death	Name		org:founded_by	Name	Yes
per:stateorprovince_of_death	Name		org:date_founded	Value	
per:city_of_death	Name		org:date_dissolved	Value	
per:cause_of_death	String		org:country_of_headquarters	Name	
per:countries_of_residence	Name	Yes	org:stateorprovince_of_headquarters	Name	
per:statesorprovinces_of_residence	Name	Yes	org:city_of_headquarters	Name	
per:cities_of_residence	Name	Yes	org:shareholders	Name	Yes
per:schools_attended	Name	Yes	org:website	String	
per:title	String	Yes			
per:employee_or_member_of	Name	Yes			
per:religion	String	Yes			
per:spouse	Name	Yes			
per:children	Name	Yes			
per:parents	Name	Yes			
per:siblings	Name	Yes			
per:other_family	Name	Yes			
per:charges	String	Yes			

Table 1: List of slots for TAC KBP 2014 slot filling. The slot types can be: Name, i.e., named entities such as person, organizations, or locations; Value, i.e., numeric entities such as dates or other numbers; and String, which do not fall in any of the previous two categories. The list column indicates if the slot accepts multiple values for a given entity.

consider the query "per:country_of_birth" for the entity "Michele Obama" and the texts:

> *Michelle Obama started her career as a corporate lawyer specializing in marketing and intellectual property. She was born in Chicago.*
>
> ...
>
> *Chicago is the third most populous city in the United States, after New York City and Los Angeles.*

Using this information, a system can correctly extract the filler "per:country_of_birth=United States" for the above query. The provenance for this filler must include elements of the last two sentences, at least: "She was born in Chicago" and "Chicago is the third most populous city in the United States" (which were necessary to perform the inference that generated this slot filler). Importantly, the provenance no longer has to include text that disambiguates ambiguous mentions of entity and filler (although systems will not be penalized if they do). In this particular example, the entity mention is ambiguous in the above provenance ("She"). LDC assessors will manually disambiguate such mentions by reading a few sentences surrounding the provided provenance (this was proved sufficient in the previous evaluations). The human assessor will judge the correctness of the (possibly normalized) slot filler string, and correctness of the provenance offsets. We will report two different scores for this task: (a) ignoring the provenance offsets, and (b) scoring the provenance offsets, i.e., a slot filler will be considered correct only if both its value and its justification are correct. All in all, assuming the first block of text starts at offset 100 in document D1, and the second starts at offset 200 in document D2, a valid encoding for this provenance would be (without the quotes): "D1:209-232,D2:200-260".

Filler Values

Column 5 (if present) contains the canonical string representing the slot filler; the string should be extracted from the filler provenance in Column 6, except that any embedded tabs or newline characters should be converted to a space character and dates must be normalized. Systems have to normalize document text strings to standardized

Column 1	Query id (same as 2013)
Column 2	Slot name (same as 2013)
Column 3	A unique run id for the submission (same as 2013)
Column 4	NIL, if the system believes that no information is learnable for this slot, in which case Columns 5 through 7 are empty; or provenance for the relation between the query entity and slot filler, consisting of up to **4 triples** in the format: docid:startoffset-endoffset separated by comma. Each of these individual spans may be **at most 150 UTF-8 characters**. Similar to 2013, each document is represented as a UTF-8 character array and begins with the "<DOC>" tag, where the "<" character has index 0 for the document. Note that the beginning <DOC> tag varies slightly across the different document genres included in the source corpus: it can be spelled both with upper case and lower case letters, and it may include additional attributes such as "id" (e.g., <doc id="doc_id_string"> is a valid document start tag). Thus, offsets are counted before XML tags are removed. In general, start offsets in these columns must be the index of the first character in the corresponding string, and end offsets must be the index of the last character of the string (therefore, the length of the corresponding mention string is endoffset startoffset + 1).
Column 5	A slot filler (possibly normalized, e.g., for dates) (same as 2013)
Column 6	Provenance for the slot filler string. This is either a single span (docid:startoffset-endoffset) from the document where the canonical slot filler string was extracted, or (in the case when the slot filler string in Column 5 has been normalized) a set of up to two docid:startoffset-endoffset spans for the base strings that were used to generate the normalized slot filler string. Same as Column 4, multiple spans must be separated by commas. The documents used for the slot filler string provenance must be a subset of the documents in Column 4. LDC will judge Correct vs. Inexact with respect to the document(s) provided in the slot filler string provenance.
Column 7	Confidence score (same as Column 9 in 2013)

Table 2: Description of SF output format.

month, day, and/or year values, following the TIMEX2 format of yyyy-mm-dd (e.g., document text "New Years Day 1985" would be normalized as "1985-01-01"). If a full date cannot be inferred using document text and metadata, partial date normalizations are allowed using X for the missing information. For example:

- "May 4th" would be normalized as "XXXX-05-04";
- "1985" would be normalized as "1985-XX-XX";
- "the early 1900s" would be normalized as "19XX-XX-XX" (note that there is no aspect of the normalization that captures the "early" part of the filler).

See the assessment guidelines document[2] for more details on the normalization requirements.

Filler Provenance

As mentioned in Table 2, the filler provenance must point to a canonical mention, rather than an arbitrary mention. For example, if the provenance document for the above per:country_of_birth example contains both "United States" and "US", the filler and the corresponding provenance must point to "United States".

Confidence Scores

To promote research into probabilistic knowledge bases and confidence estimation, each non-NIL response must have an associated confidence score. Confidence scores will not be used for any official TAC 2014 measure. However, the scoring system may produce additional measures based on confidence scores. For these measures, confidence scores will be used to induce a total order over the responses being evaluated; when two scores are equal, the response appearing earlier in the submission file will be considered to have a higher confidence score for the purposes of ranking. A confidence score must be a positive real number between 0.0 (representing the lowest confidence) and 1.0 (inclusive, representing the highest confidence), and must include a decimal point (no commas, please) to clearly distinguish it from a document offset. In 2014, confidence scores may not be used to qualify two incompatible fills for a single slot; submitter

[2] http://surdeanu.info/kbp2014/def.php

systems must decide amongst such possibilities and submit only one. For example, if the system believes that Barts only sibling is Lisa with confidence 0.7 and Milhouse with confidence 0.3, it should submit only one of these possibilities. If both are submitted, it will be interpreted as Bart having two siblings.

3 Scoring

The scoring procedure adapts the 2013 procedure to account for the multi-document provenance introduced this year. Same as before, we will pool the responses from all the systems and have human assessors judge the responses. To increase the chance of including answers that may be particularly difficult for a computer to find, LDC will prepare a manual key, which will be included in the pooled responses. The slot filler (Column 5) in each non-Nil response is assessed as Correct, ineXact, Redundant, or Wrong, as follows:

1. A response that contains more than four provenance triples (Column 4) will be assessed as Wrong.

2. Otherwise, if the text spans defined by the offsets in Column 4 (+/- a few sentences on either side of each span) do not contain sufficient information to justify that the slot filler is correct, then the slot filler will also be assessed as Wrong.

3. Otherwise, if the text spans justify the slot filler but the slot filler in Column 5 either includes only part of the correct answer or includes the correct answer plus extraneous material, the slot filler will be assessed as ineXact. No credit is given for ineXact slot fillers, but the assessor will provide a diagnostic assessment of the correctness of the justification offsets for the response. Note: correct filler strings will be assessed using the information provided in Column 6 of the output format (see Table 2).

4. Otherwise, if the text spans justify the slot filler and the slot filler string in Column 5 is exact, the slot filler will be judged as Correct (if it is not in the live Wikipedia at the date of query development) or Redundant (if it exists in the live Wikipedia). The assessor will also provide a diagnostic assessment of the correctness of the justification offsets for the response.

Two types of redundant slot fillers are flagged for list-valued slots. First, two or more system responses for the same query entity and slot may have equivalent slot fillers; in this case, the system is given credit for only one response, and is penalized for all additional equivalent slot fillers. (This is implemented by assigning each correct response to an equivalence class, and giving credit for only one member of each class.) Second, a system response will be assessed as Redundant with the live Wikipedia; in KBP 2014, these Redundant responses are counted as Correct, but NIST will also report an additional score in which such Redundant responses are neither rewarded nor penalized (i.e., they do not contribute to the total counts of Correct, System, and Reference below).

Given these judgments, we count:

- Correct = total number of correct equivalence classes in system responses;
- System = total number of non-NIL system responses; and
- Reference = number of single-valued slots with a correct non-NIL response + number of equivalence classes for all list-valued slots.

The official evaluation scoring metrics are:

- Recall (R) = Correct / Reference
- Precision (P) = Correct / System
- $F1 = \frac{2PR}{P+R}$

The above F1 score is the primary metric for system evaluation.

4 Data

The 2014 SF task will use the same knowledge base and source document collection as 2013. We detail these resources below.

4.1 Knowledge Base and Source Document Collection

The reference knowledge base includes nodes for 818,741 entities based on articles from an October 2008 dump of English Wikipedia. Each entity in the KB will include the following:

Type	Source	Person Count	Organization Count
Training	2009 Evaluation	17	31
	2010 Participants	25	25
	2010 Training	25	25
	2010 Training (Surprise SF task)	24	8
	2010 Evaluation	50	50
	2010 Evaluation (Surprise SF task)	30	10
	2011 Evaluation	50	50
	2012 Evaluation	40	40
	2013 Evaluation	50	50
Evaluation	2014 Evaluation	50	50

Table 3: English Monolingual Slot Filling Data.

- A name string
- An assigned entity type of PER, ORG, GPE, or UKN (unknown)
- A KB node ID (a unique identifier, such as "E101")
- A set of 'raw' (Wikipedia) slot names and values
- Some disambiguating text (i.e., text from the Wikipedia page)

The 'raw' slot names and the values in the reference KB are based on an October 2008 Wikipedia snapshot. To facilitate use of the reference KB, a partial mapping from raw Wikipedia infobox slot names to generic slots is provided in training corpora. Note that this year the reference KB is used solely as a potential training resource. As discussed above, the assessment of Redundant filler is performed against the live Wikipedia.

The source documents for the KBP 2014 English Slot Filling tasks will be identical to 2013, and will include approximately one million newswire documents from a subset of Gigaword (5th edition), approximately one million web documents, and approximately 100 thousand documents from discussion fora. This collection will be distributed by LDC to KBP participants as a single corpus, entitled "TAC 2014 KBP English Source Corpus", with Catalog ID LDC2014E13. In addition of the source documents, this corpus contains the output of BBNs SERIF NLP pipeline on these documents, in the hope that this simplifies data management and system development for participants.

4.2 Training and Evaluation Corpus

Table 3 summarizes the KBP 2014 training and evaluation data provided to participants.

5 Participants Overview

Table 4 summarizes the participants that submitted at least one SF run. A larger number of teams (31) registered, but only 18 teams submitted results. Out of these 18 teams, six are new participants in 2014. Table 5 compares the number of participants and submissions with previous years. The last table shows that the numbers of submissions increased this year, but the number of participants remained flat.

6 Results and Discussion

6.1 Overall Results

Table 6 lists the results of the best run for each participating team. These scores are comparable with last year's scores as the scorer largely follows the same measures as last year (with minor implementation adjustments to account for the modified relation provenance).

Similar to last year, we also report diagnostic scores, which ignore fillers redundant with the knowledge base[3]. Generally, diagnostic scores are marginally lower than the corresponding official scores. Our conjecture is that fillers that exist in the KB are the somewhat easier, e.g., with higher redundancy in the dataset, which increases the likelihood that systems extract them, and that Wikipedia page authors include them in the infoboxes.

In general, the official scores show a higher median than last year (19.8 vs. 15.7 F1 points) but a lower maximum score (36.72 vs. 37.28 F1). However, the top two groups from 2013 did not participate this year. If we remove them from this analysis, the maximum score of systems that participated in both years increased from 33.89

[3] As discussed, this year we checked redundancy against the live Wikipedia, whereas last year's assessments used the reference KB, which is a 2008 Wikipedia snapshot.

Team Id	Organization(s)	New Participant?	# of Runs Submitted
BUPT_PRIS	Beijing University of Posts and Telecommunications		5
CIS	University of Munich	Y	5
CMUML	Carnegie Mellon University		5
Compreno	ABBYY		1
GEOL	GEOLSemantics	Y	1
ICTAS_OKN	Chinese Academy of Sciences – Institute of Computing Technology	Y	4
IIRG	University College Dublin		4
IRTSX	IRT SystemX	Y	1
NYU	New York University		4
RPI_BLENDER	Rensselaer Polytechnic Institute		5
SAFT_ISI	University of Southern California – Information Sciences Institute		1
SYDNEY	University of Sydney		5
StaRAI2014	Indiana University – School of Informatics and Computing	Y	5
Stanford	Stanford University		5
UGENT_IBCN	Ghent University – iMinds	Y	4
UMass_IESL	University of Massachusetts Amherst – Information Extraction and Synthesis Lab		3
UWashington	University of Washington – Department of Computer Science and Engineering		3
utaustin	University of Texas at Austin – AI Lab		4

Table 4: Overview of the SF participants at KBP 2014.

	Teams	Submissions
2009	8	16
2010	15	31
2011	14	31
2012	11	27
2013	18	53
2014	18	67

Table 5: Number of participants and submissions in the past six years of KBP SF.

F1 in 2013 to 36.72 F1 this year. This analysis, combined with the observation that the queries this year were harder (see Section 6.4 for a detailed discussion), indicates that considerable progress was achieved in one year. This highlights the merit of repeating evaluations with minimal changes for several years, to allow technology to reach maturity. However, these results also indicate that it takes considerable time to reach this maturity: out of the nine systems ranked over the median, only two are new participants. The seven others participated in several previous SF evaluations.

Similar to last year, the top system this year is at approximately 52% of human performance (i.e., of the LDC annotators), and the median score is at only 28% of human performance. This is much lower than other NLP tasks, such as part-of-speech tagging or named entity recognition, where machines approach human performance. Given that we continue to see progress from year to year, this suggests that the SF evaluation should continue, to motivate information extraction (IE) technology to improve.

With respect to technology, several observations can be made:

- Similar to previous years, there are some clear trends: (a) most approaches use distant supervision (DS) (only four out of the 18 participating teams did not use DS); (b) many teams combined DS with rule-based approaches; and (c) most successful approaches used query expansion (QE). The top three systems all shared this architecture. Notably, the highest score this year was obtained using DeepDive[4] (Niu et al., 2012), an IE framework based on DS that is a first participant in the SF evaluation. Most systems combine multiple approaches (e.g., DS and patterns) by simply concatenating the outputs produced by the individual components. NYU is an exception to this rule: they used the pattern-based guidance mechanism of (Pershina et al.,

[4] http://deepdive.stanford.edu/

	Diagnostic Scores			Official Scores		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
Stanford	**0.2776**	**0.5443**	**0.3677**	**0.2771**	**0.5461**	**0.3677**
RPI_BLENDER	0.2706	0.4424	0.3358	0.2751	0.4487	0.3411
ICTCAS_OKN	0.2122	0.5184	0.3012	0.2153	0.5242	0.3053
utaustin	0.2334	0.3879	0.2914	0.2333	0.39	0.2919
NYU	0.2565	0.3373	0.2914	0.2552	0.3381	0.2909
UMass_IESL	0.2042	0.4092	0.2724	0.2053	0.4128	0.2743
SAFT_ISI	0.1680	0.3205	0.2204	0.1684	0.3231	0.2214
BUPT_PRIS	0.2213	0.1959	0.2078	0.2233	0.19875	0.2103
SYDNEY	0.1619	0.2728	0.2032	0.1625	0.2753	0.2043
UGENT_IBCN	0.1609	0.2413	0.1931	0.1595	0.2413	0.1920
Compreno	0.1348	0.1668	0.1491	0.1385	0.1720	0.1535
UWashington	0.0754	0.5	0.1311	0.0767	0.5065	0.1333
CMUML	0.0663	0.2704	0.1066	0.0677	0.2764	0.1088
StaRAI2014	0.0704	0.0852	0.0771	0.0717	0.0874	0.0788
CIS	0.0402	0.0487	0.0440	0.0428	0.0521	0.0470
IIRG	0.0291	0.0417	0.0343	0.0309	0.0445	0.0364
IRTSX	0.0331	0.0265	0.0295	0.0329	0.0265	0.0294
GEOL	0.0070	0.2692	0.0137	0.0069	0.2692	0.0136
LDC	0.5895	0.8746	0.7043	0.5882	0.8753	0.7036

Table 6: Overall results for SF, for the 100 entities in the evaluation dataset. The diagnostic score ignores fillers that are redundant with the reference KB (similar to previous years). The official score considers these redundant fillers as correct during scoring (similar to 2013). If multiple runs were submitted, we report the best run for each group. Results are listed in descending order of the official F1 score. The LDC score corresponds to the output created by the LDC experts.

2014) to relabel relation instances during the training of the DS model.

- Active learning was used this year by the top two systems to select more informative training data for relations (Stanford) (Angeli et al., 2014), or to optimize the acquisition of relevant training documents (RPI_BLENDER). This suggests that active learning is a successful direction to mitigate the noise introduced in the training process by DS.

- With respect to pattern-based approaches, BUPT_PRIS used a bootstrapping approach to acquire extraction patterns based on dependency tree paths. They performed above the median but not in the top three, where DS dominates. This indicates that in the debate of what is a better strategy for IE: DS (many noisy examples) or bootstrapping (few, high quality examples), DS appears to be winning. Two groups used patterns on top of Open IE (UWashington, CMUML) but they did not perform above the median. SYDNEY, which performed just above the median, relied solely on manually developed patterns. This suggests that machine-learning-based approaches perform better for SF than models crafted by domain experts.

- For the SF problem, inference is hard. For example, utaustin augmented relations that are explicitly stated in the text, which were extracted by the system of (Roth et al., 2014), with ones that are inferred from the stated relations using probabilistic rules that encode commonsense world knowledge. These probabilistic first-order logic rules were learned using Bayesian Logic Programs (BLP) (Raghavan et al., 2012). However, the inference rules degraded the performance of the original system (the performance in Table 6 is the system without inference).

- Other notable approaches used unsupervised learning. UMass_IESL's universal schema

	Official Score with `ignoreoffsets`			Official Score with `anydoc`			
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**	**F1 Increase**
Stanford	**0.2814**	**0.5540**	**0.3732**	**0.2977**	**0.5854**	**0.3947**	+2.70
RPI_BLENDER	0.2764	0.4504	0.3426	0.2937	0.4780	0.3638	+2.27
ICTCAS_OKN	0.2175	0.5291	0.3083	0.2427	0.5898	0.3439	+3.86
utaustin	0.2355	0.3933	0.2946	0.2567	0.4283	0.3210	+2.91
NYU	0.2564	0.3394	0.2922	0.2797	0.3698	0.3185	+2.76
UMass_IESL	0.2065	0.4148	0.2758	0.2207	0.4428	0.2946	+2.03
BUPT_PRIS	0.2265	0.2014	0.2132	0.2527	0.2244	0.2377	+2.74
SAFT_ISI	0.1696	0.3250	0.2229	0.1788	0.3422	0.2349	+1.35
UGENT_IBCN	0.1606	0.2428	0.1933	0.1858	0.2805	0.2235	+3.15
SYDNEY	0.1656	0.2804	0.2082	0.1758	0.2972	0.2209	+1.66
Compreno	0.1417	0.1757	0.1569	0.1498	0.1856	0.1658	+1.23
UWashington	0.0778	0.5131	0.1351	0.0859	0.5657	0.1491	+1.58
CMUML	0.0698	0.2845	0.1121	0.0779	0.3170	0.1251	+1.63
IIRG	0.0848	0.1221	0.1001	0.1048	0.1508	0.1237	**+8.73**
StaRAI2014	0.0718	0.0874	0.0789	0.0749	0.0911	0.0822	+0.34
CIS	0.0459	0.0558	0.0503	0.0599	0.0728	0.0657	+1.87
IRTSX	0.0339	0.0273	0.0303	0.0419	0.0338	0.0374	+0.80
GEOL	0.0079	0.3076	0.0155	0.0099	0.3846	0.0194	+ 0.58
LDC	0.5898	0.8768	0.7052	0.5954	0.8842	0.7116	+0.80

Table 7: Results for SF ignoring justification. In the `ignoreoffsets` configuration justifications are considered correct if the correct document is reported (similar to past years' evaluations). In the `anydoc` configuration justifications are completely ignored, and fillers are marked as correct solely based on string matching with gold fillers. For comparison purposes, we used the same runs for each participant as in Table 6. Results are listed in descending order of the F1 score with `anydoc`. The LDC score corresponds to the output created by the LDC experts.

model combines observed and unlabeled data by performing a joint optimization over the train and test data together to factorize a matrix consisting of observed relations between entities (Riedel et al., 2012). Although this year's UMASS_IESL is conceptually similar to last year's system, its performance doubled this year. This a further argument for repeated evaluations, which allow technology to mature.

6.2 Results without Justification

Table 7 lists system results when we relax the constraints on the justification. The left block of the table includes results when the scorer has the parameter `ignoreoffsets` set to true, which means that the justification is considered correct when the reported document id is correct (i.e., all offsets are ignored). The right block in the table shows results when the scorer has the parameter `anydoc` set to true, in which case the entire justification is ignored and fillers are considered correct if they match a gold filler. Note that these lenient scoring strategies have an important side effect: they collapse per:title fillers with the same value but applied to different organizations (e.g., "**CEO** of Apple" is different than "**CEO** of Next") because, without document ids and in-document offsets, we can no longer differentiate between them. Empirically, we observed that this collapsing of per:title fillers impacts mostly the `anydoc` configuration. For this reason, these lenient scores are not immediately comparable with the official scores in Table 6.

Despite the above limitation, several observations can be made based on the results in Table 7:

- In the `ignoreoffsets` configuration, scores are generally only approximately 1 F1 point higher. This indicates that the requirement to provide in-document offsets for justification does not impact the overall

	Official Scores			
	Recall	**Precision**	**F1**	**F1 Difference**
Stanford	0.2463	0.4214	**0.3109**	-0.0568
UMass_IESL	0.2367	0.4454	0.3091	+0.0348
ICTCAS_OKN	0.1787	**0.5362**	0.2681	-0.0372
NYU	0.2173	0.30	0.2521	-0.0388
utaustin	0.1884	0.2977	0.2307	- 0.0612
SAFT_ISI	0.1932	0.2797	0.2285	+0.0071
RPI_BLENDER	0.1932	0.2758	0.2272	-0.1139
SYDNEY	0.1884	0.2689	0.2215	+0.0172
UGENT_IBCN	0.1739	0.2432	0.2028	+0.0108
BUPT_PRIS	**0.2512**	0.1507	0.1884	-0.0219
UWashington	0.1014	0.5833	0.1728	**+0.0395**
CMUML	0.0821	0.2741	0.1263	+0.0175
Compreno	0.0966	0.1459	0.1162	-0.0373
StaRAI2014	0.0483	0.0917	0.0632	-0.0156
CIS	0.0338	0.0555	0.0420	-0.0050
IRTSX	0.0386	0.0346	0.0365	+0.0071
IIRG	0.0144	0.0158	0.0151	-0.0213
GEOL	0.0048	0.20	0.0094	-0.0042
LDC	0.5362	0.8161	0.6472	-0.0564

Table 8: Results for the 15 confusable entities in the evaluation dataset: SF14_ENG_012, 014, 016, 019, 022, 027, 031, 061, 062, 066, 076, 079, 096, 097, and 099. For comparison purposes, we used the same runs for each participant as in Table 6; the "F1 Difference" column indicates the difference between the F1 scores in this table and the official score F1 values in Table 6. Results are listed in descending order of the official F1 score.

	Official Score 2014			Official Score 2013		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
Stanford	0.2622	0.2915	0.2761	0.2841	0.3586	0.3170
UWashington	0.0598	0.7228	0.1104	0.1029	0.6345	0.1770

Table 9: Comparison of identical systems for the 2014 vs. 2013 test queries. Note that the 2014 runs in this table were not the best runs for each group.

score in a considerable way. This suggests that, as long as systems manage to retrieve a correct supporting document, they generally extract justifications and provenances that are considered correct by LDC evaluators. The same behavior was observed last year.

- One exception to the rule is the IIRG system, whose performance increased approximately 7 F1 points in the `ignoreoffsets` configuration, and over 8 points in the `anydoc` configuration. This suggests a bug in offset generation.

- On the other hand, identifying a valid supporting document for the extracted relation remains a challenge for some systems. Note that the `anydoc` scores are further removed from the official scores because ignoring the document id causes more collapsing for the per:title slots than the `ignoreoffsets` option. For example, because of this, the LDC score, which indicates the performance of the human expert, is boosted by almost 1 F1 point. However, even when accounting for this discrepancy, it is clear that some systems were penalized for not reporting a correct supporting document. For example, the performance of two systems (ICTAS_OKN and UGENT_IBCN) increased by over

3 F1 points in this configuration. Six other systems improved by more two F1 points, suggesting that, surprisingly, the identification of a supporting remains a challenge.

6.3 Performance on Confusable Queries

The test queries this year included 15 entities that were "confusable", i.e., are ambiguous across the document collection. For example, one of these entities, "John Graham" (SF_ENG_019), has 25 Wikipedia pages, corresponding to the various individuals with this name, which range from a 16th century Scottish nobleman to a modern day British journalist. Table 8 lists the performance of the participating systems only on these 15 ambiguous queries.

The table demonstrates that most systems perform worse on these queries. The maximum performance on these queries is 31 F1 (a drop of 5 points from the complete evaluation), and the median is 19.6 F1 (a smaller drop of 0.2 F1 points from the median in Table 6). This result suggests that the idea of disambiguating entities through entity linking *before* slot filling is attempted is valuable, at least for ambiguous queries.

Table 8 also shows that systems are affected differently by these ambiguous queries. While the top ranked system remained the same (Stanford), some systems were more affected than others. The most affected by the ambiguous queries was RPI_BLENDER, whose F1 score drops 11 points. In general, most systems (12 out of 18) see a performance penalty. For the six that are not negatively affected, the increase in performance is generally minimal, with the exception of UWashington and UMass_IESL, whose performance increased by more than 3 F1 points.

6.4 Were the Queries more Difficult this Year?

To understand if the queries were more difficult this year than 2013, we compared the performance of identical systems on the two datasets. Two groups, Stanford and University of Washington, submitted one run this year using a system identical to last year. These results are compared in Table 9.

The results in the table indicate that the test queries this year were indeed more difficult than last year. The performance of the Stanford systems decreases from 31.7 F1 points in 2013 to 27.6 this year. Similarly, the performance of the UWashington system decreases from 17.7 to 11 F1 points. A side effect of this analysis is that, by comparing these runs against their best runs in Table 6, we can understand how much progress the two groups made in one year. For example, for Stanford, this difference is considerable: their performance increases from 27.6 F1 (using the 2013 system) to 36.7 F1, a 33% relative improvement!

The increase in complexity has two probable causes. First, as discussed in the previous section, this year's dataset contained 15 "confusable" queries, which introduced a certain ambiguity in the task. This is the likely cause for the drop in precision in the Stanford system. Second, this year's dataset contains more obscure queries, with less support in the document collection. This is the cause for the drop in recall for the UWashington system. The University of Washington group graciously offered their internal analysis, which supports the latter observation. Their analysis shows that in 2013 only 9 query entities participated in less than 10 Open IE tuples, which indicates minimal support in the collection. This year 30 query entities had this property. Similarly, their entity linker could link 36 query entities to their Wikipedia page; this year the linker succeeded for only 9 entities.

6.5 System Error Analysis

6.5.1 Error Distribution

In this section we will investigate whether the main causes of the remaining errors have changed across years. The previous work (Ji et al., 2010; Min and Grishman, 2012; Pink et al., 2014) provided some nice categorizations of errors from KBP2010 slot filling results. Figure 1 is from (Min and Grishman, 2012). Following a similar categorization scheme, we analyzed the spurious errors from *all* KBP2914 slot filling teams, and the answers missed by *all* systems. Figure 2 presents the distribution of various failure cases. We can see that the outstanding challenges still come from two major natural language understanding problems: ambiguity (coreference) and variety (inference and implicit relations). One new cause comes from document retrieval because 10% queries in 2014 were intentionally selected to be highly ambiguous entities. Sentence-level IE

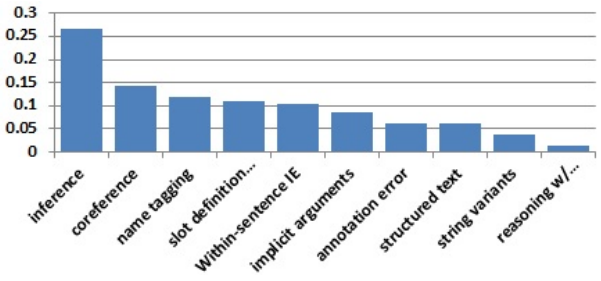

Figure 1: Error distribution in the KBP2010 Slot Filling track

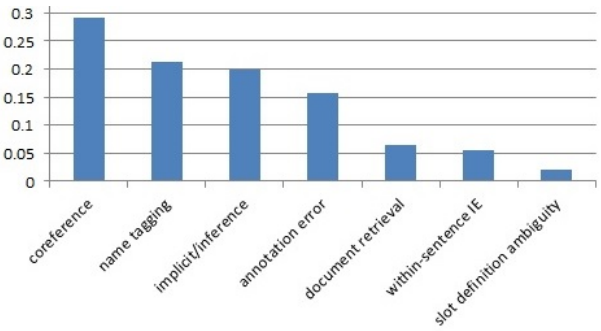

Figure 2: Error distribution in the KBP2014 Slot Filling track

caused a smaller percentage of errors. We present some detailed examples for each category next.

6.5.2 Document Retrieval

Since Entity Linking and Slot Filling were divorced in 2010, the Information Retrieval/Entity Search problem has been largely neglected in the Slot Filling community. Almost all teams used a standard pipeline of query reformulation/expansion and document retrieval based on Lucene or Indri. The general strategies adopted by most systems for person queries are name structure analysis and exact matching (e.g., so that "*Benjamin Chertof*" and "*Michael Chertoff*" won't match; and "*Ahmed Rashid*" and "*Rashid Ghazi*" won't match). For organization queries which tend to have more alternative names, most systems used more relaxed matching to enhance recall, which unfortunately introduced many irrelevant documents. For example, documents about "*China-Taiwan Orchestra*" were mistakenly retrieved as relevant for "*National Taiwan Symphony Orchestra*".

More importantly, most slot filling systems did not use entity linking during search. For example, in order to decide whether the following document "*...As to her identification as a sister of Sir John de Graham, see J. Ravilious, Addition: Agnes Graham, wife of Sir John Douglas (d. ca. 1350)...*" is relevant to the query "*John Graham*", we need to compare the profiles of the query and the entity mention "*John de Graham*". In some cases document-level profile comparison is not enough. For example, the following document "*...Chen Tao, one of the 13 surviving comfort women in Taiwan,...*" is irrelevant to the query "*Chen Tao*" who is a politician in Mainland of China. In order to solve these problems, RPI's KBP2014 system (Hong et al., 2014) proposed a temporality-based clustering model to extract the biography of a query, and then apply temporal distribution as additional constraints to refine the search results. They demonstrated that even for queries which were not intentionally selected to contain a large degree of ambiguity and variety, it's essential to design more effective entity search methods to retrieve relevant documents and evidence sentences.

6.5.3 Name Entity Recognition

For the KBP data sets, name entity recognition (NER) is not a solved problem. The best NER F-measure on this data is only around 75%. For a detailed analysis please refer to the KBP2014 Entity Discovery and Linking overview paper (Ji et al., 2014).

6.5.4 Coreference Resolution

As we can see from Figure 2, coreference errors increased from 15% in 2010 to 30% in 2014. Many errors involve nominal anaphors and non-identity coreference. For example, it's very difficult for the current statistical coreference resolvers to link "*Muslim*" and "*role model*" in the following sentence "*A convert to Islam stands an election victory away from becoming the second Muslim elected to Congress and a role model for a faith community seeking to make its mark in national politics.*" because it is not clear whether they refer to the same entity or they are two different entities mentioned in a conjunction structure. Many systems mistakenly extracted "*murder*" as the "*per:charges*" slot filler for the query "*Tamaihia Lynae Moore*" from the following sentence: "*No one knows how Tamaihia Lynae Moore died, but the foster mother of the Sacramento toddler has been arrested for her murder.*" due to the incorrect coreference link between "*Tamaihia Lynae Moore*" and "*the foster*

mother". We would need to incorporate the world knowledge that a victim cannot be a murder to fix this error.

This suggests that the KBP community should design a new task specifically to evaluate coreference and develop more external knowledge resources (e.g., can we manually mark up the semantic distance between any two nominal mention heads in WordNet to indicate how likely they are to be coreferential?) instead of adding more manual labels for the end task.

6.5.5 Sentence-level Relation and Event Extraction

We are making consistent progress on sentence-level Information Extraction, but the state-of-the-art performance on system generated entity mentions is still not satisfying: 53% F-score for relation extraction and 48% F-score for event extraction (Li et al., 2014). In particular, deeper analysis (beyond dependency parsing) is necessary to distinguish directed slot types such as parent/subsidiary and members/member_of, and avoid over-generating death related slot fillers from metaphors in discussion forum posts (e.g., "*I didn't want to hurt him . I miss him to death.*" doesn't include any attack or death events even though it includes common trigger words "*hurt*" and "*death*").

6.5.6 Implicit Relations and Inferences

Most relations are represented in many different forms. Below are several examples which will benefit from paraphrase discovery:

- *received a seat*: "In her second term, **she received a seat** on the powerful **Ways and Means Committee**" indicates "she" was a member of "Ways and Means Committee";
- *face*: "**Jennifer Dunn** was the **face** of the **Washington state Republican Party** for more than two decades." indicates "Jennifer Dunn" was a member of "Washington state Republican Party";
- *escaped into*: "**Buchwald** lied about his age and **escaped into** the **Marine Corps**." indicates "Buchwald" joined "Marine Corps".
- *completed a dual review*: "**I** have just **completed a dual review** for **Catholic News Service**, Washington, of this important topic, and share it with you here." indicates "I" is an employee of "Catholic News Service";
- *own*: "**We** decided to visit **Alberta**, our **own** province, in 2007 and now want to share some of that in words and pictures." indicates "We" lived in "Alberta";

Other cases need further inference using world knowledge. For example, "**Buchwald** 's 1952 **wedding** – **Lena Horne arranged** for it to be held in London 's Westminster Cathedral – was attended by Gene Kelly , John Huston , Jose Ferrer , Perle Mesta and Rosemary Clooney , to name a few." indicates "Buchwald" and "Lena Horne" are spouses.

The regular slot filling 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 can be used to infer from each other and ensure consistency. Systems would benefit from specialists which are able to reason about times, locations, family relationships, and employment relationships. For example, in the KBP2014 slot filling guideline, if A lives in B, and B is part of C, then we should infer A lives in C. Therefore from the following sentence " Rarely in **my** 36 Western Canada Liberal years - **Ontario**-born but transplanted permanently here. **Ontario** is **Canada** 's populous and second-largest province." we should infer "Canada" as a filler for "per:countries_of_residence".

6.5.7 Slot Definition and Task Specification

However, it is generally difficult to specify how many steps of inferences are allowed (or necessary) and how much world knowledge should be required for inferences. Some challenging examples are as follows. In the future we aim to explicitly address such situations in the annotation guidelines.

- "**He** has been evacuated to **France** on Wednesday after falling ill and slipping into a coma in Chad, Ambassador Moukhtar Wawa Dahab told The Associated Press. His wife, who accompanied Yoadimnadji to Paris, will **repatriate his body** to Chad, the amba. " – does this sentence indicate he died? and if so did he die in France?

- "Until last week, **Palin** was relatively unknown outside **Alaska**." – does this sentence indicate "Palin" lived in "Alaska"?
- "Police have arrested a Sacramento woman on suspicion of killing her **17-month-old** foster daughter" – should we infer the "age" slot filler as "1" from "17-month-old"?
- "Nine-year-old **Dutch** boy Ruben van Assouw, the sole survivor of a plane crash that killed 103 people" – does "origin" usually indicate "countries_of_residence"?
- "**She** and **Russell Simmons**, 50, **have two daughters**: 8-year-old Ming Lee and 5-year-old Aoki Lee" – does "have two daughters" indicate they are a couple?

6.5.8 Human Assessment Errors

The manual annotation scores are fairly high and stable across years, but not perfect: F-score is around 70%. Furthermore, in addition to the inevitable human annotation errors, we also noticed that human assessors mistakenly judged a few correct system generated answers as wrong. In the future, we could speed up human assessment using the automatic truth-finding methods described in RPI's slot filling validation system (Yu et al., 2014) or Stanford's active learning approach (Angel et al., 2014).

7 Concluding Remarks

With respect to the SF task, this year's evaluation continues the positive trends seen in the past year. First, SF continues to be popular, with 18 teams submitting results in 67 different runs (the largest number of runs to date). SF continues to attract new participants: out of the 18 participating teams, six were first participants this year. Second, this year's results show increased performance. The maximum score of systems that participated in the past two SF evaluations increased from 33.89 F1 points in 2013 to 36.72 F1 this year. Similarly, the median score of all submissions increased from 15.7 F1 points to 19.8. This is despite the fact that the test queries this year were more complex, containing, at the same time, ambiguous entities (i.e., same name, multiple real-world entities), and obscure entities, with minimal support in the document collection.

While this improvement is very positive, it is important to note that SF systems are still far from human performance on this task. The top system this year achieves 52% of human performance, and the median system is at only 28% of human performance. We are still far from solving the SF problem. We believe it is important to continue this evaluation, to allow information extraction technology to advance and mature.

With respect to future work, one immediate change that is necessary is to update the reference knowledge base from the 2008 Wikipedia to a more recent and modern resource, such as DBpedia[5]. This will minimize the disconnect between the SF training data available to participants and the assessment of results, which uses the live Wikipedia. Furthermore, we would like to incorporate (require?) more inference in the SF task (maybe through a closer interaction with Cold Start).

Acknowledgments

We gratefully thank Hoa Dang and LDC, in particular Joe Ellis, for helping with the task organization. They contributed key pieces without which the task could not have taken place, ranging from implementing the evaluation scorer (Hoa), managing registrations and submissions (Hoa), to data generation and assessment (Joe and LDC). We also thank Stephen Soderland, Benjamin Roth, and Gabor Angeli for helping with the data analysis.

[5] http://dbpedia.org/About

References

G. Angel, J. Tibshirani, J. Wu, and C. D. Manning. 2014. Combining distant and partial supervision for relation extraction. In *Proc. The 2014 Conference on Empirical Methods on Natural Language Processing)*.

Gabor Angeli, Julie Tibshirani, Jean Y. Wu, and Christopher D. Manning. 2014. Combining distant and partial supervision for relation extraction. In *Proceedings of EMNLP*.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference – 6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*.

Y. Hong, X. Wang, Y. Chen, J. Wang, T. Zhang, J. Zheng, D. Yu, Q. Li, B. Zhang, H. Wang, X. Pan, and H. Ji. 2014. Rpi-soochow kbp2014 system description. In *Proc. Text Analysis Conference (TAC2014)*.

H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis. 2010. An overview of the tac2010 knowledge base population track. In *Proc. Text Analytics Conf. (TAC'10)*, Gaithersburg, Maryland, Nov.

H. Ji, H. T. Dang, J. Nothman, and B. Hachey. 2014. Overview of tac-kbp2014 entity discovery and linking tasks. In *Proc. Text Analysis Conference (TAC2014)*.

Q. Li, H. Ji, Y. Hong, and S. Li. 2014. Constructing information networks using one single model. In *Proc. the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP2014)*.

B. Min and Grishman. 2012. Challenges in the tac-kbp slot filling task. In *Proceedings of LREC 2012*.

Feng Niu, Ce Zhang, Christopher Re, and Jude W Shavlik. 2012. Deepdive: Web-scale knowledge-base construction using statistical learning and inference. In *Proceedings of VLDS*.

Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. 2014. Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of ACL*.

G. Pink, J. Nothman, and J. R. Curran. 2014. Analysing recall loss in named entity slot filling. In *Proc. The 2014 Conference on Empirical Methods on Natural Language Processing)*.

Sindhu Raghavan, Raymond J. Mooney, and Hyeonseo Ku. 2012. Learning to "read between the lines" using Bayesian Logic Programs. In *Proceedings of 50th Annual Meeting of the Association for Computational Linguistics*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin Marlin. 2012. Relation extraction with matrix factorization and universal schemas. In *Proceedings of North American Chapter of the Association for Computational Linguistics*.

Benjamin Roth, Tassilo Barth, Grzegorz Chrupa, Martin Gropp, and Dietrich Klakow. 2014. Relationfactory: A fast, modular and effective system for knowledge base population. In *Proceedings of Demonstrations at ACL*.

Mihai Surdeanu. 2013. Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *Proceedings of the TAC-KBP 2013 Workshop*.

D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. Magdon-Ismail. 2014. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *Proc. The 25th International Conference on Computational Linguistics (COLING2014)*.