Biography-Dependent Collaborative Entity Archiving for Slot Filling

Yu Hong[†]
Soochow University
& Rensselaer Polytechnic Institute
tianxianer@gmail.com

Xiaobin Wang[†]**, Yadong Chen, Jian Wang**
Soochow University
Suzhou, CHN (215006)
czwangxiaobin@gmail.com

Tongtao Zhang, Heng Ji[†]
Rensselaer Polytechnic Institute
NY, USA (12180)
jih@rpi.edu

Abstract

Knowledge Base Population (KBP) tasks, such as slot filling, show the particular importance of entity-oriented automatic relevant document acquisition. Rich, diverse and reliable relevant documents satisfy the fundamental requirement that a KBP system explores the nature of an entity. Towards the bottleneck problem between comprehensiveness and definiteness of acquisition, we propose a collaborative archiving method. In particular we introduce topic modeling methodologies into entity biography profiling, so as to build a bridge between fuzzy and exact matching. On one side, we employ the topics in a small-scale high-quality relevant documents (i.e., exact matching results) to summarize the life slices of a target entity (i.e., biography), and on the other side, we use the biography as a reliable reference material to detect new truly relevant documents from a large-scale partially complete pseudo-feedback (i.e., fuzzy matching results). We leverage the archiving method to enhance slot filling systems. Experiments on KBP corpus show significant improvement over state-of-the-art.

1 Introduction

Entity archiving is an entity-oriented document retrieval task. Towards a target entity of a specific type, such as the ones discussed in this paper, a person or an organization, the goal of entity archiving is to search and collect all relevant documents from large-scale data sets under limited prior knowledge of the entity. We limit our study to the regular English entity archiving, in which the prior knowledge contains the commonly used full name (formatted by English entity naming criteria) along with a gold-standard reference document, such as a news story on President "*George W. Bush*".

Entity archiving plays a fundamental role in KBP tasks. It narrows down the range of the data source for knowledge discovery to small-scale closely related documents. Such documents, on one hand, contain informative content on a target entity, which is extremely favorable for background knowledge extraction. On the other hand, the documents provide definitive evidence for verifying the claimed identity of the entity.

As for KBP slot filling and verification tasks (Surdeanu and Ji, 2014), the archived relevant documents to an entity provide sufficient contexts (provenances) of the concrete instances (fillers) of the entity attributes (slots). See Figure 1, in which both the *provenances* support filler extraction, meanwhile the *provenance 1* additionally provides the evidence to verify the fillers (e.g., is *episcopalism* the true religion of *Bush*?).

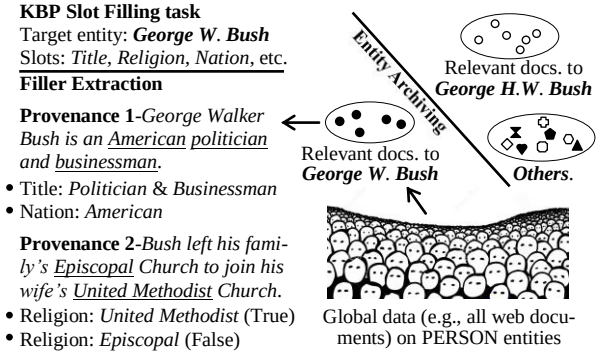

Figure 1**:** Use of entity archiving in slot filling

The main challenges of entity archiving are as

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 665–675,
Lisbon, Portugal, 17-21 September 2015.

follows: 1) it is difficult to retrieve all relevant documents through exact matching at the level of entity name, because an entity can be mentioned in various forms, such as alternate names and abbreviations; 2) in contrast, fuzzy matching introduces a large amount of noise into retrieval results (see the examples in Figure 1), although it is capable of recalling an overwhelming majority of relevant documents; 3) inadequate prior knowledge makes it difficult to generate a full profile of an entity; 4) although pseudo-feedback is helpful to enrich the prior knowledge, traditional entity profiling (e.g., bag-of-words) methods establish vague boundaries among different life slices of an entity. For example, they are incapable of distinguishing the slice of the "*Church Scientologist in Sea Organization*" of *Mark Fisher*[1] from the freelance career as the "*Corporate liaison to Miscavige*". As a result it is difficult to enhance the independent effects of different slices on the entity-document relevance determination.

To solve these problems, we propose a collaborative entity archiving (CEA) method. It employs the exact-matching based document retrieval to obtain a few high-quality reference documents, and leverages fuzzy matching for high-speed acquisition of adequate candidate documents (section 3). In addition, CEA uses the reference documents as prior knowledge to model the topic-level biography of an entity, and identifies the truly relevant documents from the candidates based on biography-document relevance (section 4). Experiments show that CEA has substantial advantages over traditional retrieval methods (section 5.1). We apply CEA to state-of-the-art slot filling systems. Experimental results show that CEA provides consistent gains (section 5.2).

2 Related Work

- **Entity Search**

One research topic similar to entity archiving is entity search. Entity search aims to seek, collect, and rank entities associated with specific information needs (Balog et al, 2011). The TREC Enterprise track featured an *expert finding* task (Balog et al, 2008a): given a topic, return a ranked list of experts on the topic. In response to this problem, there have been considerable efforts on content based retrieval models, as well as feature selection, such as proximity (Petkova and Croft, 2007), document priors (Hu et al, 2006; Zhu et al, 2010), expert-document associations (Balog and De-Rijke, 2008) and external evidence (Serdyukov and Hiemstra, 2008).

Since INEX was launched in 2002, which is an entity ranking task specific to structured data and multimedia (Demartini et al, 2010), structured features have been widely used in entity search, such as the most recent studies on Wikipedia links and categories (Vercoustre et al, 2008; Zhu et al, 2008; Jiang et al, 2009; Weerkamp et al, 2009; Kaptein and Kamps, 2009; Balog et al, 2011) and web link structure (Balog et al, 2008b; You et al, 2011; Blanco et al, 2011; Neumayer et al, 2012; Bron et al, 2013).

- **Slot Filling**

The goal of slot filling is to seek and extract the concrete instances (fillers) specific to multiple entity attributes (slots) from a large-scale textual data set (Ji et al., 2010 and 2011; Surdeanu, 2013; Surdeanu and Ji, 2014). The quality of the fillers largely depends on the performance of entity archiving and information extraction.

Related studies on archiving mainly employed traditional retrieval techniques, including query expansion and string matching (Ji and Grishman, 2011). A few studies involved document ranking and prioritizing by using probability model (Byrne and Dunnion, 2010; Roth et al, 2014) and statistical language model (Chrupala et al, 2010).

For filler extraction, great efforts were made to generate effective patterns and structure perceptrons by supervised learning and reasoning (Chen et al, 2010; Grishman and Min, 2010; Gao et al, 2010; Surdeanu et al, 2011; Louis et al, 2011; Kisiel et al, 2013). And effective feature selection and distant-supervision based classifiers have been explored (Lehmann et al, 2010; Artiles et al, 2011; Sun et al, 2011; Roth and Klakow, 2013; Roth et al, 2014). Recently active learning (Angeli et al, 2014), truth-finding (Yu et al, 2014) as well, scanning (Yu et al., 2015) and ensemble learning (Viswanathan et al., 2015) were introduced to this field.

- **Brief Summary**

In all, entity search concentrates on the analysis of a single specific aspect of an entity, which is of interest or related to a domain. In the *expert finding* task, only *academic career*s of person entities (potential experts) are of concern in entity-document relevance determination. By contrast, for the sake of comprehensive understanding of an entity, entity archiving necessarily

1 Mark Fisher: a PERSON query name in Slot Filing evaluation of 2014, ID=SF14_ENG_031

takes multiple and diverse aspects into account, such as a person's career, family, religion, sociality, academics, etc. Due to the difference in goals, entity search techniques cannot be used directly to solve entity archiving problems.

The performance of conventional retrieval techniques was generally limited due to the lack of precise modeling of the characteristics of an entity. Sparse prior knowledge and absence of effective profiling methods cause difficulties in characterizing the entity. The rest of the paper will be about knowledge acquisition and partition, as well as the collaborative method, along with a topic-level biographical profiling method.

3 Prior Knowledge Acquisition

We use string matching based retrieval methods to acquire relevant documents. It is worth considering that the acquired documents are not straightforwardly defined as the final entity archiving results. As we will show in this section, some of them are reliable, while others are full of noise. Instead, we regard them as the prerequisite knowledge for a coarse-to-fine processing.

In the retrieval phase, a query *Q* is formulated as the full name of the target entity, while a document *D* is represented as a string of words. *D* is determined to be relevant only if it contains some words that match *Q*. Accordingly we name such words as entity mentions. Both *Q* and *D* are preprocessed by tokenization and stop-word filtering. Other commonly used preprocessing steps (stemming and lemmatization) are disabled because they may cause confusion between entity mentions and common words. Table 1 shows the examples where the underlined words in <1> denotes an entity mention but <2> does not.

Mark Fisher (PER) <ID: SF14_ENG_031 >
<1> *Mark Fisher*, Sea Org member After stemming: ***mark fish***
<2>How to *mark fishing* landmarks? After stemming: ***mark fish***
3rd Guard Division (ORG) <ID: SF14_ENG_085 >
<1>The 3rd *Guard Division* of People's Liberation Army of China. After lemmatization: ***guard divide***
<2>The 24 *guards divide* up into 2 groups. After lemmatization: ***guard divide***

Table 1: Inappropriate use of preprocessing

We employ two matching methods for the relevance determination: exact and fuzzy matching.

Exact matching (EM) requires that a sequence of successive words in *D* exactly matches *Q*. By EM, entity archiving regards a full entity name as an indivisible word-order-fixed unit. Accordingly it only acquires the documents which contain the entity mentions in the form of completely-preserved full name.

Fuzzy matching (FM) relaxes the conditions to a large extent, allowing *Q* to be split into nonadjacent words. In particular, it supports the change in word order as well as partial match. By FM, entity archiving is able to retrieve documents that contain the entity mentions in the form of separated, pruned and/or reordered names. Table 2 shows some examples of using these matching methods, where the mark "•" denotes the available methods for a certain form of entity mention.

Mark Fisher (PER) <ID: SF14_ENG_031 >
Mark Fisher, Sea Org member. (**exact**) Availability: EM (•) FM (•)
Mark, husband of Julie *Fisher*. (**separation**) *Fisher* had been Miscavige's aide for 7 years.(**pruning**) *Fisher*'s first name, *Mark*, is impressive due to his inconceivable career change. (**reordering**) Availability: EM ( ) FM (•)

Table 2: Examples of string matching results

EM and FM have substantially different advantages and disadvantages in entity archiving. Table 3 shows the performance of EM and FM based entity archiving on KBP corpus (Surdeanu, 2013). We will introduce the corpus in details in Section 5. EM yields precise archiving results because the constraint conditions are helpful to reduce uncertainty in string matching. In contrast FM-based archiving is able to match entity name mentions with various forms, and thus it achieves higher recall.

	Precision	Recall	F-measure
EM	**72.5**	52.6	61.0
FM	10.8	**86.8**	19.2

Table 3: Effects of EM and FM on archiving

However FM generally introduces much noise, namely those mistakenly retrieved irrelevant documents. The documents are recalled because some pseudo entity mentions they contain can easily satisfy the constraints of fuzzy matching. See examples of the pseudo mentions in Table 4. As a result, FM yields a very low precision score.

Mark Fisher (PER) <ID: SF14_ENG_031 >
PlantWeb is a *mark* of the *Fisher*-Rosemount group of companies. (**separation**)
Deutsche *Mark* was the currency in Germany (**pruning**)
Iconic *Fisher*-Price *mark*. (**reordering**)

Table 4: Examples of pseudo entity mentions obtained by using FM

Undoubtedly it is helpful for global optimization of entity archiving to make full use of the advantages of EM and FM. In view of the above-mentioned investigation, we partition the string matching results into two parts, exact and fuzzy ones, which are used as reliable prior knowledge (named *reference source*) and unrefined prior knowledge (*candidate source*) respectively. Most documents in the *reference* are truly related to the target entity but the scale is not big (see Recall of EM in Table 3), while the *candidate* is full of both true answers and noise (see Precision and Recall of FM in Table 3), respectively. As we will show in the next section, the final archiving results are generated by synthesizing the sources in a collaborative coarse-to-fine way.

4 Collaborative Entity Archiving (CEA)

We propose a Collaborative Entity Archiving approach (CEA for short). CEA synthesizes the reference source and candidate source in a collaborative manner (section 4.1) through a biography-document relevance determination method (section 4.2 and 4.3). In addition, CEA involves mention disambiguation and query expansion in pre-processing to optimize the quality of both sources (section 4.4)

4.1 Overall Framework of CEA

CEA models the biography of an entity by using the topics in the reference source, in which, each topic serves as the description of a slice of life of the target entity (*life slice* for short), as shown in Figure 2. The *Life slice* means an episode in the whole story of the entity, which may represent an event, state or scenario at a certain moment, such as a person's *birth* or an organization's *establishment*.

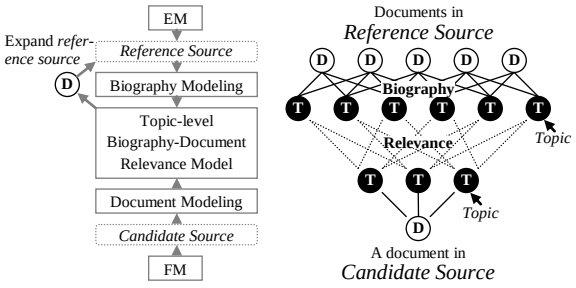

Figure 2**:** Framework of collaborative archiving

CEA pulls out a document from the candidate source, one by one, and measures the biography-document relevance at the topic level. By using a relevance threshold as the discrimination factor, CEA either preserves the document if it is relevant, or filters otherwise. Meanwhile, CEA adds the newly found relevant documents to the reference source, and updates the biography by reshaping life slices (i.e., topics). CEA iteratively goes through the process of biography formulation, biography-document relevance measurement and determination until a condition is satisfied. Finally CEA selects all the preserved documents in the reference source as the final output. Figure 2 shows the framework.

4.2 Biography-Document Relevance Models

We design a generative approach to estimate the biography-document relevance *r*, which calculates the conditional probability that a candidate document *D* generates the biography *B*:

$$r = P(B|D) \quad (1)$$

In total we leverage three probabilistic models for modeling *B* and *D*, including relevance model, topic model and context-level topic model. Then we introduce Hellinger Distance (Lindsay, 1994) into relevance measurement.

- *Relevance Model* **(*RM*)**

Generally, *Relevance Model* (RM) (Huang and Croft, 2009) refers to the probability distribution over all words conditioned on their occurrences in a set of previously-known relevant documents (or high-quality pseudo-relevant documents), i.e., $\forall w \in V, P(w|R)$, where *V* is the vocabulary, *R* is the document set, and *P*(*w*|*R*) can be estimated by TF-IDF. RM is often used in combination with *Document Model* (DM). Similar to RM, DM refers to the probability distribution over words in a particular document, i.e., $\forall w \in V, P(w|D)$. The relevance between *R* and *D* is normally determined by the agreement of RM and DM. The agreement can be estimated with Hellinger Distance between the models:

$$H(RM|DM) = \sum_{w \in V} (\sqrt{P(w|R)} - \sqrt{P(w|D)})^2 \quad (2)$$

RM is a widely-used probabilistic model for information retrieval. It determines the relevance of a document to an object in accordance with homogeneousness in content between the document and the relevant documents of the object.

For an entity, in our case, we generate RM on the reference source, and regard it as the probabilistic model of a macro-level all-embracing biography *B* over the prior knowledge *R*. For a candidate document *D*, the biography-document relevance *r* is measured with Hellinger Distance between RM and DM: *P*(*B*|*D*)=*H*(*RM*|*DM*). We

will demonstrate the effect of RM heavily relies on the quality of reference source in experiments.

- *Topic Model* **(*TM*)**

Empirically, RM is coarse-grained. It mixes up different, separate and incoherent life slices of an entity. A more serious problem is that RM assigns uneven weights to life slices, giving excessive weights to the words about the popular slices, but low or even zeroth weights to the unpopular ones. A popular slice is defined as the slice of greater concern, which is normally frequently mentioned in the reference source, such as the slice of "*the career of George W. Bush* as *the President*" (high weight) versus "*his childhood*" (low weight). As a result, the RM based biography-document relevance is only helpful to identify and recall the documents relevant to the popular slices but not to the unpopular ones.

As a modification, we employ *Topic Model* (TM) for biography modeling. We define a topic in the reference source as an independent fine-grained representation for a microscopic life slice. Accordingly we treat the biography as a bucket of topics. We leverage Latent Dirichlet Allocation (LDA) (Blei et al, 2003; Wei and Croft, 2006) for topic discovery and modeling in the reference source. A topic is modeled as a probability distribution over all words in lexicon conditioned on the association of the words with the topic, denoted as $\forall w \in V$, $P(w|t_R)$, in which t_R refers to a topic in the reference source, representing a life slice *s*. Table 5 shows partial topic models (slices) in the reference source of *Mark Fisher*, where the highlighted probability values by a box reveal the words that well characterize a topic (slice). In the same way, we survey the topics t_C in the *candidate source*, modeled as $\forall w \in V$, $P(w|t_C)$. It is worth noting that those topics (t_C) may represent anything, namely the slices of the target entity or namesakes, related or unrelated events, etc. It means they are full of noise.

Mark Fisher (Slice Modeling) <ID: SF14_ENG_031>			
Slice 1(***s1***), topic t_{s1}: ***political*** *career*			
$P(w\|t_{s1})=0.003$ "*Parliament*"	$P(w\|t_{s1})=8E-5$ "*screenwriter*"	$P(w\|t_{s1})=8E-5$ "*film*"	$P(w\|t_{s1})=4E-5$ "*Bear Fisher*" …
Slice 2(***s2***), topic t_{s2}: ***artistic*** *career*			
$P(w\|t_{s2})=6E-5$ "*Parliament*"	$P(w\|t_{s2})=0.003$ "*screenwriter*"	$P(w\|t_{s2})=0.001$ "*film*"	$P(w\|t_{s2})=3E-5$ "*Bear Fisher*" …
Slice 3(***s3***), topic t_{s3}: ***family***			
$P(w\|t_{s3})=9E-5$ "*Parliament*"	$P(w\|t_{s3})=1E-5$ "*screenwriter*"	$P(w\|t_{s3})=5E-6$ "*film*"	$P(w\|t_{s3})=0.001$ "*Bear Fisher*" …

Table 5: Example of life slice modeling by TM

In practice, given a target entity, its reference source (exact matching results) is a subset of the candidate source (fuzzy matching results). We picked the *reference source* out of the candidate to parse topics independently, forming the set of t_R. Meanwhile, we parse topics in the candidate source to form the set of t_C. Benefitting from the separate treatment, some of the truly related topics (t_R) to the entity (correct slices) can be collected along with less noise. Using the topics as references, we detect the relevant documents in the *candidate source* in terms of the topic-level biography-document relevance $P(B|D)$.

Given a document D in the candidate source, we transform $P(B|D)$ into the combination of topic-document relevance of all topics in the *reference source*. We measure the topic-document relevance with the conditional probability $P(t_R|D)$ that the topic t_R occurs in the document D. Accordingly, $P(B|D)$ is estimated by:

$$r = P(B \mid D) = \prod_{t_R} P(t_R \mid D)$$
$$\log(r) = \sum_{t_R} \log P(t_R \mid D) \quad (3)$$

where, we incorporate the log likelihood into the numerical calculation for the sake of nonzero joint probability.

Due to the separate topic modeling procedures for the reference and candidate sources, the probability $P(t_R|D)$ − a topic t_R in the *reference source* occurs in a candidate document D − cannot be obtained directly. To solve the problem, we introduce the joint probability of topic-topic relevance between topics (t_R) in reference and topics (t_C) in candidate (see the mode in Figure 2) into the probability calculation:

$$\forall t_R \in T_R,\ P(t_R \mid D) = \prod_{t_C \in T_C} P(t_R \mid t_C) P(t_C \mid D) \quad (4)$$

where T_R is the set of all topics in the reference source while T_C is the candidate. The topic-topic relevance $P(t_R|t_C)$ is approximated by Hellinger distance estimation between the topic models of t_R and t_C. As a whole, we measure the biography-document relevance as:

$$\begin{aligned} \log(r) &= \log \prod_{t_R \in T_R} \prod_{t_C \in T_C} P(t_R \mid t_C) P(t_C \mid D) \\ &= \sum_{t_R \in T_R} \sum_{t_C \in T_C} \log P(t_R \mid t_C) P(t_C \mid D) \\ &= \sum_{t_R \in T_R} \sum_{t_C \in T_C} \log H(t_R \mid t_C) P(t_C \mid D) \\ &= \sum_{t_R \in T_R} \sum_{t_C \in T_C} \sum_{w \in V} \log(\sqrt{P(w \mid t_R)} - \sqrt{P(w \mid t_C)})^2 P(t_C \mid D) \end{aligned} \quad (5)$$

We employ the toolkit GibbsLDA++[2] in topic modeling, which is an implementation of LDA using Gibbs sampling (Porteous et al, 2008). GibbsLDA++ makes it easy to parse the topics in a document set as well as estimate topic models $P(w|t)$. Besides, GibbsLDA++ offers the probability over topics in generating a specific document, facilitating the estimation of $P(t_C|D)$ in equation (5). Table 6 shows the operating parameters what we set in experiments, where the ones $\{\alpha, \beta\}$ were set as the default values while the iterative number *num* is an empirical value.

α= 1	*num*= 200
β= 0.1	N_t←HDP

Table 6: Operating parameters of GibbsLDA++

The necessary precondition for GibbsLDA++ in topic partition is to define the number N_t of potential topics in a set of documents. We execute the Hierarchical Dirichlet Processes (Teh et al, 2005), abbr., HDP, to predict N_t. HDP is similar to current hierarchical information organization methods, such as the hierarchical clustering (Kummamuru et al, 2004), unsupervised and coarse-to-fine grained. Hence HDP is useful in exploring the basic rules of topic partition in an automatic way, such as number and granularity. We employ HDP to estimate the number (N_t) of all possible topics in *reference source* and *candidate* separately, acquiring two N_t for each target entity, one per source.

- *Context-level Topic Model* **(*CTM*)**

In consideration of the reliability of contexts in representing closely-related life slices to the entity, we use the contexts around entity mentions to improve the slice-oriented topic modeling.

SEN: A sentence where an entity mention occurs **NEB**: Left and right neighbor sentences of SEN **DEP**: Words dependent on entity mention **SYN**: Words in maximum syntactic subtree in SEN
Context 1: SEN **Context 2**: DEP + SYN **Context 3**: SEN + left NEB + Right NEB

Table 7: Instructions of various types of contexts

A context consists of the words co-occurring with an entity mention in a radius-fixed text span or syntactic or dependent structure (see instructions in Table 7). Given a target entity, the entity mention in the *reference source* is its full name. The union of all contexts in the source defines the vocabulary V_R that most probably represent the slices of the entity. In the *candidate sourc*e, on the contrary, the entity mention can be a reordered, separated or pruned entity name, as well as abbreviation or alias, such as *GWB* (abbr.) and *Dubya* (alias) of *George W. Bush*. Different from the cases in *reference*, the vocabulary V_C obtained from the contexts in *candidate* are closely related to diverse entities or other objects with the same name (see Table 2&4).

CTM measures the biography-document relevance in the same way with *TM*, estimating the topic-level $P(B|D)$ by equation (5). The only difference lies in the available words in topic model $P(w|t)$. For the ones not included in V_R and V_C, CTM assigns a weight zero in topic model no matter what GibbsLDA++ does.

4.3 Unsupervised Threshold Estimation

For each target entity, CEA measures the biography-document relevance for all documents in the candidate source. In the light of the relevance scores, CEA ranks the documents and sets a clear threshold θ to cut off the long tail in the ranking list, in other words, filtering the documents that have a relevance score lower than θ. The preserved documents will be added to the reference source for both biography reformulation and archiving result output.

We estimate the threshold by learning density distribution of documents over relevance scores (Arampatzis et al, 2009). *Density* means the number of documents that have similar relevance scores. The *distribution* is produced by densities within all interval ranges of relevance score. Our empirical findings show that the density distribution fits a mixture of two Gaussians, where the highly relevant documents and the irrelevant ones distribute in two separate Gaussian peaks respectively. Accordingly we define the threshold as the range of relevance score at the extreme point between the peaks, as shown in Figure 3.

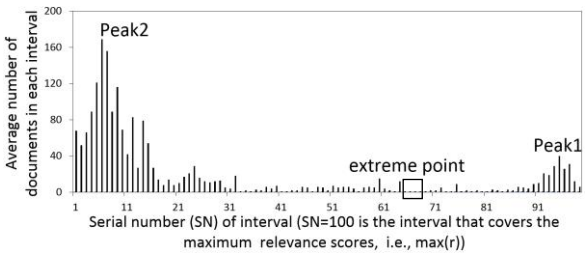

Figure 3: Extremum detection for threshold selection. (Note: Y-axis indicates the density in a specific interval range of relevance score)

In order to detect the extreme point, we firstly use a cubic polynomial function to approximate

2 http://gibbslda.sourceforge.net/

the density distribution. Second, we go through the integral solution of the function in every fine-grained interval range of relevance score (interval is set as max(r)/100). We finally detect the extremum between peaks. The threshold is initialized during runtime exclusively for each target entity, without training. It is re-estimated every time when the biography is reformulated.

4.4 Termination Criterion for Iteration

CEA identifies relevant documents in *candidate source* and moves them to *reference*. Then CEA reshapes statistical models (RM, TM or CTM) over the updated sources. In terms of the reformed models, CEA starts a new round of relevance determination, data movement, and statistical modeling. CEA keeps it going until meeting any of the following termination criterions:

- T1: No more new topic occurs in the *reference source* (N_t doesn't change).
- T2: The number of the documents in Peak1 (Figure 3) begins to increase continuously.

T2 is triggered if T1 loses its efficacy. The invalidation happens when some general slices (i.e., general topics) are mistakenly introduced into the *reference source*, causing large-scale irrelevant document to be recalled and moved to *reference*. It will dramatically increase the number (N_t) of topics in a long term in the iterative procedure, driving CEA to capture more irrelevant documents. Thus Peak1 will be enlarged continuously. However, if as expected, Peak1 should be narrowed with increasing the iteration times because:

- Fewer new related slices appear.
- The number of documents related to the slices is less than that in previous iterations.

4.5 Optimization of EM and FM

In the preprocessing phase, we improve the precision of EM because higher-quality EM results can offer more reliable reference documents for biography modeling. In addition, we expand queries for FM to recall a larger number of relevant documents. It is helpful to minimize the loss of relevant documents before proceeding to CEA.

To improve EM, we focus on identifying the common words that completely match the full name of the target entity. The words normally are elusive and easily treated as a correct entity name, called deceptive name, see that in (1).

(1) *Countrywide Financial* <ORG>
True: *Countrywide Financial Corporation*.
Deceptive: *Bank of America purchased the failing countrywide financial for $4.1 billion*.

To reduce EM errors caused by deceptive names, we use name tagging (Miller et al, 2004) to distinguish deceptive names and true names. Further, we filter the documents that are mistakenly retrieved based on the match between a deceptive name and the full-entity-name based query Q.

We leverage an *Alternate Name Table* (ANT) for query expansion. ANT is a mapping table between entity name and alternate name. An alternate name is either generated according to the naming conventions (Burman et al, 2011), such as abbreviation, suffixation and revivification. Some alternative names were extracted from knowledge base through redirect links (Nia et al, 2014), such as nicknames in Wikipedia dumps. For an entity, we reformulate query Q by the combination of the pre-assigned full entity name and all possible alternate names in ANT, see (2).

(2) **Initial** Q: *Countrywide Financial Corporation.*
Expanded: *Countrywide Financial+Corporation +Corp. +Company +Co. +Ltd. +Co Ltd. +CFC.*

We use the expanded query for FM to increase the number of relevant documents in the candidate source, regardless of whether or not it will introduce a larger scale of new noises.

5 Experiments

We evaluate our methods on KBP 2013 corpus. The corpus contains 2.1M texts collected from web pages, newswires and discussion forums. From this corpus, a slot filling system is required to find fillers for 41 types of slots that represent the attributes of the target entities. There are 25 slot types of person and 16 slot types of organization, such as a Person's *birth date* and an Organization's *founder* (Ji et al., 2010 and 2011).

KBP 2013 includes 100 target entities and ground-truth fillers and provenances, where the ground-truth data was obtained by manual verification and annotation on the pool of system outputs. The provenances contain the IDs of documents relevant to target entities and fine-grained text spans which illustrate the eligibility of fillers.

In total there are 1,851 gold standard relevant documents available for the evaluation of entity archiving. However the data is far from enough because it only covers a small portion of all relevant documents in the pool. Most are excluded since KBP annotators ignore relevant documents in which there isn't any filler for the assigned slots or, although exists, the fillers were inexactly identified by Slot Filling systems. Therefore, we manually went over the pool and extracted 4,405 relevant documents as our ground-truth.

Sources / Archiving	**Before** Optimization of EM & FM (%)						**After** Optimization of EM & FM (%)					
	Micro-Average			Macro-Average			Micro-Average			Macro-Average		
	P	R	F	P	R	F	P	R	F	P	R	F
EM	65.4	40.1	49.7	72.5	52.6	61.0	78.8	24.6	37.5	81.1	46.9	59.4
FM	9.4	74.6	16.7	10.8	86.8	19.2	6.3	92.9	11.8	6.9	94.2	12.9
Baseline	29.1	50.6	36.9	36.1	62.9	45.9	22.6	33.3	26.9	25.1	52.6	34.0
CEA(RM)	49.7	86.5	63.1	62.3	85.8	72.2	59.5	84.3	69.7	64.9	87.4	74.5
CEA(TM)	52.6	87.7	65.8	63.2	86.1	72.9	60.6	82.9	70.0	65.7	85.8	74.4
CEA(CTM1)	63.9	81.7	71.7	69.7	82.7	75.7	65.1	75.6	70.0	69.8	84.4	76.4
CEA(CTM2)	63.9	81.7	71.7	69.7	82.7	75.7	65.1	75.6	70.0	69.8	84.4	76.4
CEA(CTM3)	61.8	84.3	71.4	68.1	83.8	75.1	66.2	71.2	68.6	70.6	77.7	74.0

Table 8: Archiving results (CTM1, 2 &3 are CTM using different types of contexts, context1, 2 &3 in Table 7)

5.1 Archiving Results and Analysis

We evaluate the entity archiving methods by micro and macro Precision (P), Recall (R) and F metrics. Table 8 shows the main results.

- *Overall Archiving Results*

Overall, the proposed CEA methods perform much better than the string matching based entity archiving methods (i.e., EM and FM).

In addition the methods outperform a random-sampling based CEA (baseline), which randomly selects a certain number of documents (candidates) from the candidate source to combine with reference source straightforwardly (for final archiving results generation). The sampling number is set to be the same as the number of candidates eventually archived by RM-based CEA.

Random	RM	TM	CTM1	CTM2	CTM3
5,901	5,901	5,578	3,866	3.866	4,243
5,158	5,158	4,943	4,032	4,032	3,655

Table 9: The number of candidates added to reference source for archiving results generation (the second row indicates the number before optimizing EM and FM, while the third row indicates after optimization)

Table 9 shows the number of candidates archived from the candidate source by all kinds of CEA methods. It demonstrates that the biography-based CEAs yield higher precision (Table 8) after introducing the same or smaller number of candidates in the reference source, revealing the positive effect of biography modeling on entity-oriented document relevance determination.

- *Reliability versus Comprehensiveness*

CEA achieves higher precision by using the optimized EM results as reference source. It demonstrates the importance of reliable prior knowledge for entity understanding as well as detecting relevant documents. However, the reference source causes lower recall scores of all CEA methods. The reason lies in reduction of prior knowledge. As shown in Table 8, the refined reference source (i.e., optimized EM results) covers only 24.6% of all relevant documents, which is far less than the coverage before optimization (nearly 40%).

The reduced prior knowledge provides fewer available life slices of an entity for constructing an informative biography, inevitably resulting in missing some relevant documents. In order to confirm this, we regard the 41 KBP slot types as some readily-made visible life slices, and use the manual annotations of the slot fillers to verify whether a life slice appears in a relevant document. For example, the filler "*Corporate liaison* of the slot *Title* reveals the slice of *freelance career* of *Mark Fisher*. Then we figure out the coverage rate of life slices for both the original reference source and the refined.

Figure 4 exhibits the coverage rates for 5 most frequently occurred life slices. The coverage rate is calculated by the number of reference sources that contain a specific life slice versus 100, i.e., the number of reference sources for the 100 KBP entities (one per entity). It can be found that the refined reference sources miss lots of life slices.

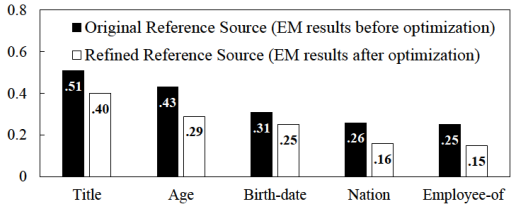

Figure 4: Coverage rates of life slices (for top 5)

- *Comparison among Biography Models*

RM is biased towards the popular life slices in biography modeling. The reasons are as following: 1) RM gives greater weights to the high-frequency words, and 2) popular slices are of much greater public interest and hence frequently mentioned in relevant documents. However, some entities not only share similar names but similar popular slices, such as the *religious vocation* of different *church scientologists*. Therefore

RM is extremely likely to acquire the documents related to the namesakes if they have similar popular background as the target entity, causing a greater loss of precision.

Table 10 shows the top highly-weighted words in RM for the target *Mark Fisher* (a *church scientologist*), along with 2 namesakes who occur most frequently in the incorrect archiving results.

Keywords in RM	The most similar entities
church *committee* *religious* *Sea Org* *policy*	***Miles Mark Fisher*** *Church historian, Educator, Baptist minister and writer* ***Mark Fisher*** *Senior Pastor*

Table 10: Entities of similar background

By contrast, TM independently represents different life slices and combines the effects of the slices on biography-document relevance determination, evenly and exhaustively. Comprehensive and unbiased measurement of every known life slices is helpful in disambiguating entities that have similar backgrounds (definitely not the same in all). As a result, TM improves the precision. And the context-based TM goes further.

5.2 Slot Filling Results and Analysis

We apply our entity archiving methods to two top-ranked slot filling systems in the evaluation of KBP 2013, including LSV (Roth and Klakow, 2013) and Blender (Yu et al 2013).

The LSV incorporates a string matching based entity archiving and a SVM classifier based filler extraction. LSV's archiving model expands queries by using suffixes and Wikipedia anchor texts, and uses mutual information based relevance measure in document ranking and filtering.

Blender employs a hybrid retrieval model for archiving relevant documents. It combines Boolean and VSM models and expands query by an alternate name table similar to ours. For filler extraction, Blender implements truth finding over conflicting claims from multiple rule-based extraction systems.

Methods	P	R	F
LSV's archiving	53.0	88.2	66.2
Blender's archiving	54.6	71.7	62.0

Table 11: Archiving performance of LSV &BLD

Table 11 shows entity archiving performances of LSV and Blender (Macro-Average P, R and F). All CEA methods perform better than the both. With the aim to optimize provenances of fillers, we modify the slot filling systems by substituting their archiving methods with ours. Table 12 exhibits the performance gains after replacement.

Slot Filling / Archiving	LSV (%)			Blender (%)		
	P	R	F	P	R	F
Original system	40.8	30.0	34.6	34.1	22.1	26.8
Mod. (RM)^{CEA}	42.1	30.0	35.0	35.6	23.9	28.6
Mod. (TM)^{CEA}	**42.2**	**30.6**	**35.5**	35.6	23.9	28.6
Mod. (CTM3)^{CEA}	42.7	29.3	34.7	**36.0**	**23.8**	**28.7**

Table 12: Slot filling performance gains

Both LSV and Blender achieve significant gains. The most interesting finding is on the different performance gains. It should reveal the fact that the well-supervised classification based filler extraction of LSV has a better capability of noise resistance, while by contrast, the truth-finding in Blender is capable of identifying valid fillers if the quality of archiving results is high, otherwise easily makes mistake.

6 Conclusion

We doubt that it is easy to maintain the stability of current entity-oriented knowledge acquisition methods, including ours, in dealing with ordinary entities. Most target entities now in use for the evaluation are made to stand as "out of the ordinary", such as well-known enterprises, celebrities or domain experts. As a result, a corpus contains abundant relevant documents of the entities but less about the little-known namesakes. It greatly reduces the interference of namesakes and thus the difficulty of the task.

In future work, we will make the task critical for success by employing the little known namesakes as targets. In addition to verifying the robustness of the CEA method, we will work on the relationship among entities (ACE entity relation types, Doddington et al, 2004) and related events (e.g., causal, temporal and sub-event relations), by which to build graph-based biography.

7 Acknowledgment

This work was supported by the U.S. DARPA DEFT Program No. FA8750-13-2-0041, ARL NS-CTA No. W911NF-09-2-0053, NSF CAREER Award IIS-1523198, AFRL DREAM project, gift awards from IBM, Google, Disney and Bosch. It was also supported by Natural Science Foundation of China (NSFC) No. K111818713, K111818612. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. and CHN Governments. The U.S. and CHN Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Gabor Angeli, Sonal Gupta, Melvin Jose, Christopher D. Manning, Christopher Ré, Julie Tibshirani, Jean Y. Wu, Sen Wu, and Ce Zhang. 2014. Stanford's 2014 slot filling systems. In Proceedings of *the 7th TAC*.

Avi Arampatzis, Jaap Kamps, and Stephen Robertson. 2009. Where to stop reading a ranked list: threshold optimization using truncated score distributions. In Proceedings of *the 32nd SIGIR*, pages 524-531, Boston, Massachusetts, USA, July.

Javier Artiles, Qi Li, Taylor Cassidy, Suzanne Tamang, and Heng Ji. 2011. CUNY BLENDER TAC-KBP2011 temporal slot filling system description. In Proceedings of *the 4th TAC*.

Bruce G. Lindsay. 1994. Efficiency versus robustness: the case for minimum hellinger distance and related methods. *Annals of Statistics*, 22(2): 1081-1114.

Krisztian Balog, Lan Soboroff, Paul Thomas, Peter Bailey, Nick Craswell, and Arjen P. De Vries. 2008a. Overview of the TREC 2008 enterprise track. In Proceedings of *the 17th TREC*, NIST.

Krisztian Balog, Wouter Weerkamp, and Maarten De Rijke. 2008b. A few examples go a long way: constructing query models from elaborate query formulations. In Proceedings of *the 31st SIGIR*, pages 371-378, Singapore, July.

Krisztian Balog and Maarten De Rijke. 2008. Associating people and documents. Advances in Information Retrieval: 296-308

Krisztian Balog, Marc Bron and Maarten De Rijke. 2011. Query modeling for entity search based on terms, categories, and examples. *ACM Transaction on Information System*, 29(4): 22:1-22:31.

Krishna Kummamuru, Rohit Lotlikar, Shourya Roy, Karan Singal, and Raghu Krishnapuram. 2004. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In proceedings of *the 13rd WWW*, pages 658-665.

Roi Blanco, Peter Mika, Sebastiano Vigna. 2011. Effective and efficient entity search in RDF data. *The Semantic Web-ISWC*: 83-97.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993-1022.

Marc Bron, Krisztian Balog, and Maarten De Rijke. 2013. Example based entity search in the web of data. *Advances in Information Retrieval*: 392-403.

Amev Burman, Arun Jayapal, Sathish Kannan, Madhu Kavilikatta, Ayman Alhelbawy, Leon Derczynski, and Robert Gaizauskas. 2011. USFD at KBP 2011: entity linking, slot filling and temporal bounding. In Proceedings of *the 4th TAC*.

Lorna Byrne and John Dunnion. 2010. UCD IIRG at TAT 2010 KBP slot filling task. In Proceedings of *the 3rd TAC*, Maryland, USA, November.

Zheng Chen, Suzanne Tamang, Adam Lee, Xiang Li, Marissa Passantino, and Heng Ji. 2010. Top-down and bottom-up: a combined approach to slot filling. In Proceedings of *the 6th AIRS*, pages 300-309, Taipei, Taiwan, December.

Grzegorz Chrupala, Saeedeh Momtazi, Michael Wiegand, and Stefan Kazalski. 2010. Saaland university spoken language systems at the slot filling task of TAC KBP 2010. In Proceedings of *the 3rd TAC*.

Gianluca Demartini, Tereza Iofciu, and Arjen P. De Vries. 2010. Overview of the INEX 2009 entity ranking track. In Proceedings of *the 8th International workshop of the Initiative for the Evaluation of XML Retrieval*, pages 254-264.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. *The ACE program-task, data, and evaluation*. LREC.

Sanyuan Gao, Yichao Cai, Si Li, Zongyu Zhang, Jingyi Guan, Yan Li, Hao Zhang, Weiran Xu, and Jun Guo. 2010. PRIS at TAC2010 KBP track. In Proceedings of *the 3rd TAC*.

Ralph Grishman and Bonan Min. 2010. New York University KBP 2010 slot filling system. In Proceedings of *the 3rd TAC*.

Guoping Hu, Jingjiang Liu, Hang Li, Yunbo Cao, Jian-Yun Nie, and Jianfeng Gao. 2006. A supervised learning approach to entity search. In Proceedings of *the 3rd AIRS*, pages 54-66, Singapore, October.

Xuanjing Huang and W. Bruce Croft. 2009. A unified relevance model for opinion retrieval. In Proceedings of *the 18th CIKM*, pages 947-956, Hong Kong, China, November.

Jiepu Jiang, Wei Lu, Xianqian Rong, and Yangyan Gao. 2009. Adapting language modeling methods for expert search to rank Wikipedia entities. *Lecture Notes in Computer Science*, 5631: 264-272.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC2010 knowledge base population track. In Proceedings of *the 3rd TAC*, Maryland, USA, November.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: successful approaches and challenges. In Proceedings of *the 49th ACL*, pages 1148-1158, Portland, Oregon, USA, June.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the TAC2011 knowledge base population track. In Proceedings of *the 4th TAC*.

Rianne Kaptein and Jaap Kamps. 2009. Finding entities in Wikipedia using links and categories. *Lecture Notes in Computer Science*, 5631: 273-279.

Bryan Kisiel, Justin Betteridge, Matt Gardner, Jayant Krishnamurthy, Ndapa Nakashole, Mehdi Samadi, Partha Talukdar, Drry Wijaya, and Tom Mitchell. 2013. CMUML system for KBP 2013 slot filling. In Proceedings of the *6th TAC*.

John Lehmann, Sean Monahan, Luke Nezda, Arnold Jung, and Ying Shi. 2010. LCC approaches to knowledge base population at TAC 2010. In Proceedings of the *3rd TAC*.

Ludovic Jean-Louis, Romaric Besançon, Olivier Ferret, and Wei Wang. 2011. Using a weakly supervised approach and lexical patterns for the KBP slot filling task. In Proceedings of *the 4th TAC*.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In Proceedings of *HLT-NAACL 2004*, pages 337-342, Boston, Massachusetts, USA, May.

Robert Neumayer, Krisztian Balog, and Kjetil Nørvåg. 2012. On the modeling of entities for ad-hoc entity search in the web of data. *Advances in Information Retrieval*: 133-145.

Morteza Shahriari Nia, Christan Grant, Milenko Petrovic, Yang Peng, and Daisy Zhe Wang. 2014. In proceedings of *the 27th FLAIRS*, pages 467-472, Pensacola Beach, Florida, USA, May.

Desislava Petkova and W. Bruce Croft. 2007. Proximity-based document representation for named entity retrieval. In Proceedings of *the 16th CIKM*, pages 731-740, Lisboa, Portugal, November.

Lan Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In Proceedings of *the 14th SIGKDD*, pages 24-27, Las Vegas, USA, August.

Benjamin Roth and Dietrich Klakow. 2013. Combining generative and discriminative model scores for distant supervision. In Proceedings of *the EMNLP*, pages 24-29, Seattle, Washington, USA, October.

Benjamin Roth, Tassilo Barth, Michael Wiegand, Mittul Singh and Dietrich Klakow. 2014a. Effective slot filling based on shallow distant supervision methods. *Arxiv preprint arxiv*: 1401-1158.

Pavel Serdyukov and Djoerd Hiemstra. 2008. Being omnipresent to be almighty: the importance of the global web evidence for organizational expert finding. In Proceedings of *the 31s SIGIR*, pages 17-24, Singapore, July.

Ang Sun, Ralph Grishman, Wei Xu, and Bonan Min. 2011. New York University 2011 system for KBP slot filling. In Proceedings of *the 4th TAC*.

Mihai Surdeanu, Sonal Gupta, John Bauer, and David McClosky. 2011. Stanford's distantly-supervised slot filling system. In Proceedings of *the 4th TAC*.

Mihai Surdeanu. 2013. Overview of the TAC2013 knowledge base population evaluation: English slot filling and temporal slot filling. In Proceedings of *the 6th TAC*.

Mihai Surdeanu and Heng Ji. 2014. Overview of the English slot filling track at the TAC2014 knowledge base population evaluation. In Proceedings of *the 7th TAC*.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2005. Hierarchival Dirichlet Processes. *Journal of the American Statistical Association* 101(476): 1-29.

Anne-Marie Vercoustre, James, A. Thom, and Jovan Pehcevski. 2008a. Entity ranking in Wikipedia. In Proceedings of *the 23rd SAC*, pages 1101-1106, Fortaleza, Ceara, Brazil, March.

Vidhoon Viswanathan; Nazneen Fatema Rajani; Yinon Bentor; Raymond Mooney. 2015. Stacked Ensembles of Information Extractors for Knowledge-Base Population. In Proceedings of ACL 2015.

Wouter Weerkamp, Krisztian Balog, and Edgar Meij. 2009. A generative language modeling approach for ranking entiies. *Lecture Notes in Computer Science*, 5631: 292-299.

Xing Wei and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In Proceedings of *the 29th SIGIR*, pages 569-577, Seattle, Washington, USA, August.

Gae-won You, Seung-won Hwang, Zaiqing Nie, and Ji-Rong Wen. 2011. Social search: enhancing entity search with social network matching. In Proceedings of *the 14th EDBT*, pages 515-519, Uppsala, Sweden, March.

Dian Yu, Hongzhao Huang, Taylor Cassidy, Heng Ji, Chi Wang, Shi Zhi, Jiawei Han, Clare Voss, and Malik Magdon-Ismail. 2014. The wisdom of minority: unsupervised slot filling validation based on multi-dimensional truth-finding. In Proceedings of *the 25th COLING*, pages 1567-1578, Dublin, Ireland, August.

Dian Yu, Heng Ji, Sujian Li and Chin-Yew Lin. 2015. Why Read if You can Scan: Scoping Strategy for Biographical Fact Extraction. In Proceedings of NAACL-HLT 2015.

Jianhan Zhu, Dawei Song, and Stefan Rűger. 2008. Integrating document features for entity ranking. *Lecture Notes in Computer Science*, 4862: 336-347.

Jianhan Zhu, Xiangji Huang, Dawei Song, and Stefan Rűger. 2010. Integrating multiple document features in language models for expert finding. *Knowledge and Information System*, 23(1):29-54.