Probabilistic Inference for Cold Start Knowledge Base Population with Prior World Knowledge

Bonan Min and **Marjorie Freedman** and **Talya Meltzer** *
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 02138, USA
{bonan.min, marjorie.freedman}@raytheon.com

Abstract

Building knowledge bases (KB) automatically from text corpora is crucial for many applications such as question answering and web search. The problem is very challenging and has been divided into sub-problems such as mention and named entity recognition, entity linking and relation extraction. However, combining these components has shown to be under-constrained and often produces KBs with supersize entities and common-sense errors in relations (a person has multiple birthdates). The errors are difficult to resolve solely with IE tools but become obvious with world knowledge at the corpus level. By analyzing Freebase and a large text collection, we found that per-relation cardinality and the popularity of entities follow the power-law distribution favoring flat long tails with low-frequency instances. We present a probabilistic joint inference algorithm to incorporate this world knowledge during KB construction. Our approach yields state-of-the-art performance on the TAC Cold Start task, and 42% and 19.4% relative improvements in F1 over our baseline on Cold Start hop-1 and all-hop queries respectively.

1 Introduction

Automatically transforming a large corpus into a structured knowledge base (KB) has long been a goal of information extraction (IE) research. KB population incorporates many IE tasks including named entity recognition, entity linking and relation extraction, each of which rely on deeper linguistic analysis, e.g., syntactic parsing or anaphora resolution. Since 2012, NIST [1] has run an open shared task in KB population (KBP) under TAC [2]. Most participating systems (Mayfield et al., 2014; Min et al., 2015; Roth et al., 2015; Angeli et al., 2014; Nguyen et al., 2014; Monahan and Carpenter, 2012) combine many independent components to perform the full task.

As will be familiar to most IE researchers, the individual components are not perfect. When combined into a pipeline, errors compound. We found that a KB produced with a simple combination of state-of-the-art IE components (Ramshaw et al., 2011) is very sensitive to component-level errors (Grishman, 2013).

Table 1 illustrates a real entity coreference mistake. *Barack Obama* and *Ehud Barak* were incorrectly linked because of ambiguous context and high lexical overlap. The mistake leads to errorful facts about employment, familial relations, etc. We see additional mistakes when we review the names in those entities with the most mentions: the *U.S.* entity contains more than 20,000 mentions. 85% are correct (e.g., *United States, U.S.*), but there is a long tail of incorrect yet infrequent(each accounting for $< 1\%$) mentions linked to the entity e.g., *North America, Latin American*. We also see counter-intuitive errors in relation extraction: 5% of person entities have multiple birthdates; the KB asserts 8 spouses for an infrequently

*The third author is currently with The Affinity project. This work was done while she was at BBN.

This work was supported by DARPA/I2O Contract No. FA8750-13-C-0008 under the DEFT program. The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

[1] U.S. National Institute of Standards and Technology

[2] Text Analysis Conference: www.nist.gov/tac/

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 601–612,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

mentioned entity. Similar errors have been reported in (McNamee et al., 2013) and (Singh et al., 2013b).

Named mentions of PER:***Barack Obama***: *Barack*, ***Barack Obama***, ***Ehud Barak***, *Barak*
Text: ***Barak*** *endorses* ***Barack***, *... Defense Minister* ***Ehud Barak*** *said* ***Barack Obama*** *has been the most supportive president on Israeli security*

Table 1: Example of entity linking errors.

Analyzing these errors suggests a limitation of performing KB population solely with IE tools. These mistakes only become obvious in the context of external world knowledge with the full set of facts extracted from many documents, e.g. when applying our expectations about the cardinality of a relation. With just a single document, resolving these mistakes requires challenging inference (Ji et al., 2005).

In this paper, we present a probabilistic framework to incorporate real-world knowledge into Cold Start KB population. Our contributions include:

- Identifying from real world datasets that entity popularity and each relation's cardinality follow the power-law distribution with long tails of low-frequency instances.
- Defining a corpus-level joint objective for KBP that incorporates multiple IE components and prior world knowledge on entity popularity and per-relation cardinalities, and showing the prior knowledge helps to reduce errors.
- Outperforming the top-ranked entry in Cold Start 2015.

The paper is organized as follows: we first introduce the Cold Start KBP task, then present the joint probabilistic framework, followed by analysis of the world knowledge and how to incorporate it. We then describe our inference algorithm. Lastly, we present experimental results, related work, and conclude with suggestions for future research.

2 Cold Start KBP

The schema consists of 3 entity types (person, organization, and GPE) and 42 slots (relation classes) [3]. Systems start with an empty KB (cold start) and populate it according to the schema with information extracted from the corpus. All facts in the KB must be grounded with justifying text from the corpus.

A KB entity is defined as a cluster of mentions that refer to the same real-world entity, e.g., *Smith*, *John Smith*, and *John H Smith* are 3 mentions for the entity *John_H_Smith*. Every named mention of an entity is recorded. A relation is a triplet *(subject, slot, object)*, where *subject* and *object* are entities, [4] and *slot* is the relation between them. For example, *(Bart Simpson, per:siblings, Lisa Simpson)* is the relation *Bart Simpson is a sibling of Lisa Simpson*. A relation's provenance points to up to 4 snippets in the corpus that justify the relation. The evaluation process is described in the Experiments Section.

3 A Probabilistic Framework

Following most Cold Start KBP systems (Mayfield et al., 2014; Min et al., 2015; Roth et al., 2015; Angeli et al., 2014; Nguyen et al., 2014; Monahan and Carpenter, 2012), our baseline uses a cascade of NLP components from document-level analysis to corpus-level aggregation. We run BBN's SERIF (Ramshaw et al., 2011) for mention, value and name tagging, coreference resolution, sentence-level relation extraction, alongside other analysis such as syntactic parses. Then we aggregate entities with entity discovery and linking and relations with relation extraction.

Given a set of pre-trained NLP components, the process is essentially an inference task. We introduce the following notation:

- M is the list of mentions
- E is the set of entities to populate the KB
- R is the set of relation types.
- x_i is the observed text for mention i, $x_i \in M$
- u_i is entity ID from the KB assigned to mention $i, i \in \{1, 2, ..., |M|\}, u_i \in \{1, 2, ..., |E|\}$
- $z_{i,j} = r$ indicates the relation r between a pair of mentions $x_i, x_j \in M$ and $r \in R \bigcup \{Other\}$
- $y^r_{i,j}$ is an indicator variable: $y^r_{i,j} = 1$ if a relation $r \in R$ exist between entity pair $< e_i, e_j >, i, j \in \{1, 2, ..., |E|\}$, and 0 otherwise.

The key steps in the pipeline are the following:

Mention extraction: We use a structured perceptron model (Ramshaw et al., 2011) to extract named mentions M.

[3] We will use *slot* and *relation* interchangeably in this paper.

[4] A small number of the relations take values not entities. We do not differentiate in this work.

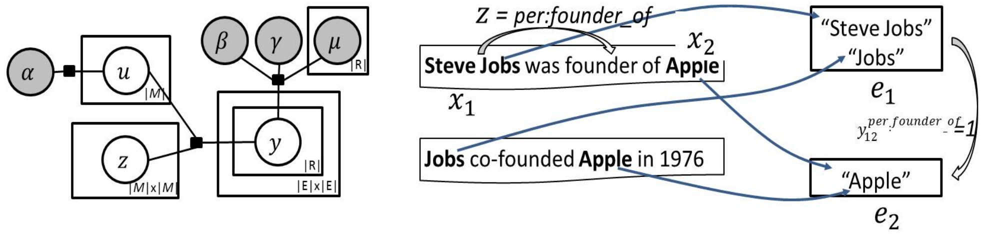

Figure 1: A simplified plate model of the probabilistic model(left), and an example (right) illustrates the KB construction process with aligned variables. The plate model only shows RE and EDL factors, and factors incorporating world knowledge. The example(right) compensates by showing the process without priors.

Entity Discovery & Linking (EDL): This step creates a candidate entity set E for the KB and infers which document entity (represented by its named mentions) is associated with which KB entity , i.e. assigning values for $\{u_i\}$. We use a sieve-like (Raghunathan et al., 2010) algorithm for in-document coreference. For simplicity, we only model EDL of names [5].

We define potential functions over variables $\{u_i\}$ for each pair of mention x_i and the jth entity e_j:

$$\Psi_i^{EDL}(u_i = j|x_i) = exp(\Sigma_k \theta_k \phi_k(u_i = j|x_i))$$

The baseline system solves the EDL problem by inferring $u_i^* = \arg\max \Psi_i^{EDL}(u_i)$. It uses a name database collected from Freebase (Bollacker et al., 2008) and GeoNames [6]. First, it clusters novel names to create new candidate entities in addition to entries in the name database. A novel name is defined as a name that could not be resolved to a database entry. It then rescans the corpus and links each document-level entity to a corpus-level entity (an entry in the name database or a novel name). The EDL model Ψ^{EDL} uses features such as string edit distance and indicators representing whether appearing in the same name variant set. $\{\theta_k\}$ and $\{\phi_k\}$ are the weights and feature functions respectively.

Mention-level Relation Extraction (MRE): This step infers which relation $z_{i,j} = r(r \in R \bigcup \{Other\})$ exists between each pair of mentions $< x_i, x_j >$, we define potential functions:

$$\Psi_{i,j}^{MRE}(z_{i,j} = r|x_i, x_j) = exp(\Sigma_k \theta'_k \phi'_k(z_{i,j} = r|x_i, x_j))$$

We run several relation finding algorithms (Min et al., 2015) , including statistical models trained from ACE [7] relation annotation and distant supervision (Mintz et al., 2009) in which we align Freebase pairs into Gigaword (Parker et al., 2011) to generate training examples, and a pattern matcher with a few hand-written patterns that capture local contexts.

To train a log-linear model Ψ^{MRE} for combining these algorithms and to tune the confidences of their extractions, we follow (Viswanathan et al., 2015) and train a stacked classifier using output and confidences of the extractors. We use assessment datasets from TAC Cold Start KBP 2013 and 2014, and Slot Filling evaluations in 2013 and 2014. The features we used are: source algorithm name, slot, confidence score (if exists), argument mention level (*pronoun, name, or nominal*), lexical sequence between pair of arguments, propositional path between the pair of arguments. $\{\theta'_k\}$ and $\{\phi'_k\}$ are the weights and feature functions respectively.

Relation Extraction (RE): This aggregation step infers which relations exist between each pair of entities at the KB level, i.e. assigning values for $\{y_{i,j}^r\}$. We define the potential functions over the indicator variables $\{y_{i,j}^r\}$, by looking at all pairs of mentions $x_m, x_n \in M$, their potential to have a relation r and likelihood to be associated with entities $e_i, e_j \in E$:

$$\Psi_{i,j,r}^{RE}(y_{i,j}^r = 1|x) = \max_{<m,n>}(\Psi_m^{EDL}(u_m = i|x_m)\Psi_n^{EDL}(u_n = j|x_n) \Psi_{m,n}^{MRE}(z_{m,n} = r|x_m, x_n))$$

and $\Psi_{i,j,r}^{RE}(y_{i,j}^r = 0|x) = \Psi_0^{RE}$ where Ψ_0^{RE} is a parameter learned from previously seen data. The aggregation from a set of $z_{m,n}$ to a set of $y_{i,j}$ is similar to noisy-or relation aggregation (Hoffmann et al., 2011; Riedel et al., 2010; Sur-

[5] Decisions made for named mentions will be applied to the corresponding document-level entities.

[6] www.geonames.org

[7] itl.nist.gov/iad/mig/tests/ace/2005/

deanu et al., 2012) and supports overlapping relations (Hoffmann et al., 2011; Surdeanu et al., 2012).

The joint distribution defined over the full set of variables u, y is:

$$Pr(u,y|x) \propto \prod_{m} \Psi_{m}^{EDL}(u_m|x_m) \cdot \prod_{i,j,r} \Psi_{i,j,r}^{RE}(y_{ij}^{r}|x)$$

The Cold Start KBP problem can be seen as finding the maximum a posteriori (MAP) configuration:

$$(u^*, y^*) = \arg\max_{(u,y)} Pr(u,y|x)$$

The baseline system approximates the solution by solving in 2 separate stages: solve EDL by fixing $u_m^* = \arg\max_{1 \le i \le |E|} \Psi_m^{EDL}(u_m = i|x_m)$ for all x_m, then solve RE by fixing $y_{i,j}^{*r} = \arg\max_{\delta \in \{0,1\}} \Psi_{i,j,r}^{RE}(y_{i,j}^{r} = \delta|x, u^*)$ for all i, j, r. The potential $\Psi_{i,j,r}^{RE}(y_{i,j}^{r} = 1|x, u^*)$ is a relaxed form for $\Psi_{i,j,r}^{RE}(y_{i,j}^{r} = 1|x)$:

$$\begin{aligned} &\Psi_{i,j,r}^{RE}(y_{i,j}^{r} = 1|x, u^*) = \\ &\max_{\substack{m,n: \\ (u_m^*, u_n^*)=(i,j)}} (\Psi_m^{EDL}(u_m^* = i|x_m)\Psi_n^{EDL}(u_n^* = j|x_n) \\ &\qquad \Psi_{m,n}^{MRE}(z_{m,n} = r|x_m, x_n)) \end{aligned}$$

In the relaxed form, the optimization problem required for estimating the potential of $y_{i,j}^{*r}$ is limited to search only over pairs of mentions x_m, x_n which are now associated with the entities e_i, e_j, i.e. $u_m = i$ and $u_n = j$, instead of enumerating over all pairs of mentions. This can be done efficiently.

4 Incorporating World Knowledge

We incorporate world knowledge related to *entity popularity* and a set of *per-relation cardinalities* as additional factors in the objective. To learn these factors' form, we analyze real-world datasets and find that both factors follow the power-law distribution with long tails of low-frequency instances.

4.1 Entity Popularity

We define *entity popularity* (EP) as the number of mentions of an entity in a corpus. Entities vary in *popularity*– famous people (e.g. politicians, athletes), countries, and large organizations will be mentioned frequently in news, while other entities– a small city, the local valedictorian may only be mentioned a few times. Ideally, we would model the EP distribution with counts from a large corpus annotated for names and cross-document entity coreference. As we are not aware of any such resource, we look at two approximations:

Name variants (NV): We collect name variants (e.g., *UN and United Nations*) for PER and ORG from Freebase and for GPE from GeoNames.

Name Mentions (NM): From a 50K document sample of Gigaword, for each entity, we search for exact matches to its name variants, and count these matches to estimate the number of entity mentions.

Figure 2 (left) plots the per-entity relationship between the count of NM and NV with rank. Both follow the power-law distribution (i.e. the plots are close to straight lines in a log-log scale). In other words, most entities have only a small number of variants and are mentioned only a few times. A handful of popular entities are mentioned frequently and have many variants [8]. The size of entities in the Kripke (Finin et al., 2014) system follows power-law distribution, further supporting our findings.

Formally, we define for ith entity the popularity variable $q_i(u) = \Sigma_m I(u_m = i)$ and the potential for EP factor $\Psi_i^{EP}(u)$ as follows:

$$\Psi_i^{EP}(u) \propto exp(\theta^{EP}(q_i(u)))$$

in which $\theta^{EP}(q_i(u)) = \alpha \ln(q_i)$.

The parameter $\alpha > 0$ is initially fit from Freebase entities and then finetuned with TAC 2014 dataset to reflect real-world distributions. The EP term favors EDL solutions u with a popularity distribution that follows a power-law with a long tail of low-frequency entities.

4.2 Relation Cardinality

We define relation r's *cardinality* regarding e_i as the number of entities or values associated with e_i through r. For example, if *John Smith* has 3 children, the cardinality of r=*per:children* regarding e_i=*John Smith* is 3. Formally, we notate the set of variables $\{y_{ij}^{r}\}$ as y_i^r, and the cardinality of a relation r of entity e_i as $d_i^r = \Sigma_j y_{ij}^r$.

Per-relation cardinalities (RC) often reflect real world constraints– people have at most one birthdate and typically no more than 5 siblings. To understand the cardinality constraints for the Cold Start relations, we use Freebase, a large, manually

[8] We confirm that a few entities have many variants e.g. for *Elizabeth II*, 364 variants including *Elizabeth II of the UK, Her Majesty the Queen, Queen of Australia, etc.*

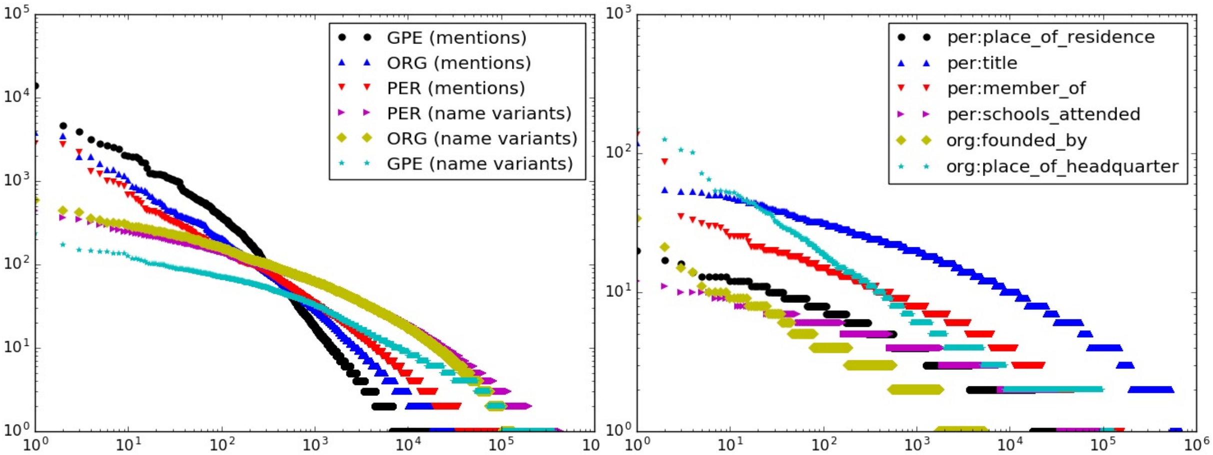

Figure 2: Real-world entity popularity (left) and per-relation cardinality (right) both follow the power-law distribution. x-axis shows ranks and y-axis shows counts (both are in log scale). The Left figure plots both numbers of name variants and numbers of mentions to ranks for PER, ORG and GPE.

curated KB. We align Freebase to the TAC schema following (Chen et al., 2010) and then generate a cardinality for each relation for all entities. The relationship between RC and RC-rank for the the 6 most frequent relations is plotted in Figure 2 (right). For these relations, RC closely resembles a power-law distribution. To favor both power-law and a soft size-limit on cardinality, we define the following potentials for RC factors for each relation-entity pair:

$$\Psi^{RC}_{i,r}(y^r) \propto exp(\theta^{RC}(d^r_i))$$

in which $\theta^{RC}(d^r_i)) = \beta \ln(d^r_i) - \gamma(\max(d^r_i - \mu^r, 0))^2$ with parameters $\beta, \gamma > 0$, and μ^r as the mean of the cardinalities of a relation r (estimated from Freebase). The first term in the potential has the power-law assumption, while the second term penalizes large cardinalities for going beyond the mean μ^r.

4.3 Incorporating Prior World Knowledge

Incorporating the EP and per-relation RC terms into the joint distribution, we obtain the joint objective:

$$Pr^*(u, y|x) \propto Pr(u, y|x) \cdot \prod_{1 \le i \le |E|} \Psi^{EP}_i(u) \cdot \prod_{\substack{1 \le i \le |E| \\ r \in R}} \Psi^{RC}_{i,r}(y^r_i)$$

with $Pr(u, y|x)$ as the baseline objective. A simplified plate diagram is shown in Figure 1.

Learning constraints for real-world corpora: As we're not aware of any large corpus annotated exhaustively with entities and relations, we fit the parameters of the constraints initially from Freebase entities and relations, and then fine-tune them using empirical utility maximization (Jansche, 2005; Ye et al., 2012) for TAC Cold Start all-hop F1 with grid search in the parameter space, using previous years' TAC assessment. Freebase is used in initialization because of its scale while finetuning with TAC assessment ensures the factors to more appropriately represent the underlying distribution of entity popularity and relation cardinality in a real-world corpus.

5 Jointly Inferring Entities and Relations

The problem of Cold Start KBP becomes finding a MAP assignment of u and y for $Pr^*(u, y|x)$. Finding the exact solution is hard, as many terms in the objective involve large groups of variables. We propose Algorithm 1 as an approximate heuristic. Line 1 generates an initial KB by approximating a solution for the baseline objective $Pr(u, y|x)$ (Section 3), but tends to overlink entities and over-aggregate relations. Lines 2-8 iteratively refine the KB by searching over the (u, y)-space using operation $o \in$ {SplitE, PruneR}. At t-th iteration, it performs the operation o with the highest potential gain $\Delta \ln Pr^*(o(u^t, y^t)|x)$. The process is repeated until the gain is smaller than a very small value ϵ.

SplitE: splits an entity e_i into two entities. Since there are an exponential number of possible SplitE actions, we uses the following two heuristics: 1) cluster name mentions by their string forms, and find an "outlier" cluster of mentions, 2) rank e_i's mentions $\{x_m : u^*_m = i\}$ by their local EDL potential $\Psi^{EDL}_m(u^*_m|x_m)$ and find the lowest-ranked mention as an "outlier". After find-

```
Input  : $x, \alpha, \beta, \gamma, \mu^r$ for $r \in R$
Output: $u, y$
$(u^0, y^0) = \arg\max_{(u,y)} Pr(u, y|x)$
$t \leftarrow 0$
repeat
    $o = \arg\max_o \Delta \ln Pr^*(o(u^t, y^t)|x)$
    if $o! = null$ then
        $(u^{t+1}, y^{t+1}) \leftarrow Execute(o(u^t, y^t))$
        $t \leftarrow t + 1$
until $\Delta \ln Pr^*(o(u^t, y^t)|x) < \epsilon$;
```

Algorithm 1: The MAP inference algorithm

ing an "outlier" mention cluster of e_i, we divide it into two entities: e_g with the "core" mentions, and e_h with the "outlier" mention cluster. We repeat the process to find all outlier entities and separate them from the entity. Relation arguments will be reattached to the new entities accordingly. We only consider a short list of most popular entities and split each using the heuristics described above.

PruneR: removes a batch of relations ($Y = \{y^r_{i,j}\}$) by setting 0: $y^r_{i,j} \leftarrow 0$ for each $y^r_{i,j} \in Y$. The batch is generated with the following steps: first select a set of entity-relation pairs (e_i, r) with the highest cardinality $d^r_i = \sum_j y^r_{i,j}$, then repeatedly select the associated relation with the lowest potential $j^* = \arg\min_{j:y^r_{i,j}=1} \Psi^{RE}(y^r_{i,j}|x)$. Each $y^r_{i,j}$ will be added into the batch until its size reaches 50.

We define the gain for SplitE and PruneR as:

$$\begin{aligned}
&\Delta \ln Pr^*(SplitE(e_g, e_h \leftarrow e_i)|x)) = \\
&\quad \Sigma_m(\ln \Psi^{EDL}_m(g) + \ln \Psi^{EDL}_m(h) - \ln \Psi^{EDL}_m(i)) \\
&\quad + \Sigma_{r\in R}(\ln \Psi^{RE}_g + \ln \Psi^{RE}_h - \ln \Psi^{RE}_i \\
&\quad + \ln \Psi^{RC}_r(y'^r) - \ln \Psi^{RC}_r(y^r)) \\
&\quad + \ln \Psi^{EP}_g(u') + \ln \Psi^{EP}_h(u') - \ln \Psi^{EP}_i(u)
\end{aligned}$$

$$\begin{aligned}
&\Delta \ln Pr^*(PruneR(Y)|x) = \\
&\quad \Sigma_{r\in R}(\ln \Psi^{RC}_r(y'^r) - \ln \Psi^{RC}_r(y^r)) \\
&\quad + \Sigma_{y^r_{i,j}\in Y}(\ln \Psi^{RE}_0 - \ln(\Psi^{RE}_{i,j,r}(y^r_{ij} = 1|x)))
\end{aligned}$$

in which $\Psi^{EDL}_m(i)$ is short for $\Psi^{EDL}_m(u_m = i|x_m)$ with m ranges over IDs of mentions in e_i, and Ψ^{RE}_i is the sum of the RE factors which are related to entity e_i. y', u' are the assignment to y, u if a SplitE or a PruneR operation is executed. We also use the short form $\Psi^{RC}_r(y'^r)$ as the sum of the RC factors which have changed because of a SplitE or a PruneR operation.

Since the gain is only computed for the shortlisted entities and relations, and we only calculate the subset of factors (EDL, RE, RC, and EP) related to the operation, $\Delta \ln Pr^*(SplitE|x)$ and $\Delta \ln Pr^*(PruneR|x)$ can be calculated efficiently.

6 Experiments

We evaluate our system with resources provided to TAC 2015 participants, including 1) a source corpus of 50,000 documents from newswire and discussion forums, 2) a query set consisting of 317 hop-0 entities (expanded to 1,148 hop-0 entry-point mentions and 8,191 hop-1 queries), 3) LDC [9] assessment of participant responses from automatic submissions [10] and a manually created submission [11], and 4) software that retrieves answers from a KB and measure performance with the assessment. Additionally, we use TAC 2013 and 2014 datasets for tuning parameters and training stacked classifiers. $\alpha = 10, \beta = 5, \gamma = 0.1$ are set empirically following Section 4.3. We run each experiment 20 times and average the scores.

6.1 Queries, Assessment, and Scoring

We briefly describe the evaluation process and scoring metrics. More details appear in (Mayfield, 2014). The Cold Start evaluation measures KB-quality by probing the KB with two types of queries. The queries are either at **hop-0** (e.g., *which organization(s) is(are) founded by Bill Gates?*) or **hop-1** (e.g., *in which city(-ies) the organization(s) founded by Bill Gates is(are) headquartered?*). More formally, the evaluation software tries to find an entity e_0 in the submitted KB that covers the entry-point mention of a hop-0 query q_0, then finds all relational triples matching $(e_0, r_1, ?)$. X, the set of entities matching the open variable, is reviewed by annotators for: (a) assessment of correctness and (b) the identification of non-redundant subset X'. The software generates an hop-1 query $q_1 = x'$ for each $x' \in X'$, finds the entity e_1 that aligns with q_1, and then finds triples matching $(e_1, r_2, ?)$. This results in response set Y, the set of entities matching the second open variable. Set Y is assessed by LDC in the same manner as Set X. The process is performed over all submitted KBs [12]. The answers in X (hop-0)

[9] https://www.ldc.upenn.edu/

[10] 20 teams participants in the task with >50 KBs submitted.

[11] created by time-limited LDC annotators

[12] LDC generates the time-limited manual run directly from text queries, but treats the responses identically for assessment

Systems	CS-SF								CS-LDC-MAX							
	Hop-0				Hop-1				Hop-0				Hop-1			
	P	R	F1	Ign	P	R	F1	Ign	P	R	F1	Ign	P	R	F1	Ign
TAC rank1	48	30	37	0	31	17	22	0	50	35	40	0	28	20	23	0
offset-based	50	31	38	64	31	18	23	23	52	35	42	19	30	21	24	6
string-match	49	31	38	13	32	19	24	11	52	36	42	6	29	21	25	3
assess	**50**	**31**	**39**	0	**32**	**19**	23	0	**53**	**37**	**43**	0	**29**	**21**	**24**	0

Table 2: CS-SF and CS-LDC-Max micro-averaged precision, recall and F1 of hop-0 and hop-1 queries for the TAC 2015 top-ranked KB submission (Min et al., 2015) and our KBC+E+R system (3 bottom rows) using various post-hoc scoring techniques (offset-based, string-match and assess). *Ign* is the number of unassessed answers.

are correct when when sufficiently justified in the source corpus. The answers in Y (hop-1) are correct only if both the element of X that generated the response is correct and the response in Y is justified in the text.

NIST reports two metrics, **CS-SF** and **CS-LDC-MAX**, which differ in the treatment of multiple entry-point mentions for a single real-world entity. CS-SF treats each distinct mention as an independent query. CS-LDC-MAX takes only the entry-point mention which maximizes system performance for a given query-entity (i.e. either the responses for *Bill Gates* or *William Gates*). For both metrics, NIST calculates micro-averaged precision, recall, and F1 over all queries. As mentioned above, the official evaluation is a human post-hoc assessment of KB output. A system developed outside of the evaluation window, e.g., our proposed algorithm, will likely include responses for which truth is not known, which are ignored by the scoring software. Table 2 compares the TAC top-ranked system to our full configuration using three post-hoc scoring strategies: strict **offset-based** match, **string-match** match, and **assess** in which we apply the offset-based metric using additional internally performed assessments. For the ablation study in Table 3, we use the official scorer's **string-match** mode. A small number of responses are ignored (**Ign**) even in string-match mode. We further account for these responses by re-estimating precision for hop-0 and hop-1 assuming that the precision of the ignored responses at hop-1 is the same as the hop-0 precision [13]. When this optimistic estimate differs from reported precision, we report it in parentheses.

6.2 Results and Discussion

Table 2 compares our full system (KBC+E+R) to the top performing system in TAC Cold Start 2015 using three different approaches to post-hoc scoring. Without manual effort, our joint modeling approach exceeds the performance of the top-ranked system, which uses a cascade of manually-specified rules (Min et al., 2015). Our system obtains 5.4% and 4.8% relative improvement in hop-0 and hop-1 CS-SF F1 over the top-ranked system. Improvement is observed in both hop-0 and hop-1 and with both CS-SF and CS-LDC-MAX showing that the improvement is robust. A sign test shows that the improvements are significant with $p < 0.01$.

Systems	Hop-0			Hop-1		
	P	R	F1	P	R	F1
KBC	45	**33**	37.9	14(16)	**21**	17
KBC+E	45	32	37.9	18(21)	21	19
KBC+R	49	32	38.2	27	19	23
KBC+E+R	**49**	31	38.4	**32**	19	**24**

Table 3: CS-SF scores (string-match) for different priors: KBC (baseline: no world knowledge), KBC+E (only entity-based factors), KBC+R(only relation-based factors), KBC+E+R(both sets of factors). Numbers in parentheses indicate the optimistic estimate when it differs from the number reported by the scoring software.

Table 3 ablates each type of world knowledge to show the impact of entity and relation-based factors independently when compared to a version of our system without world knowledge. As expected, the impact of world knowledge is seen in improvements in precision at minor costs to recall. Both types of world knowledge have higher impact on hop-1 than hop-0 as hop-1 measures the formation of the KB with multiple hops in relations. Adding the relation factors has a larger

[13] This overestimates the hop-1 precision which is lower than hop-0 precision because of error compounding.

impact than adding the entity factors because our splitting of entities is conservative (only affects $< 0.1\%$ entities) while relations' factors removes 7.3% relations. The two classes of factors appear to have largely independent impacts– combining them yields a large improvement. In total, adding prior world knowledge yields relative improvements of 9% in hop-0 precision, 131% on hop-1 precision, 42% on hop-1 F1, and 19.4% on all-hop F1 over the baseline. A sign test shows the improvements are significant with $p < 0.01$.

Reduction of errors: With relation factors added (KBC+R and KBC+E+R), 7.3% relations (out of 243K) are removed by **PruneR** with minimal recall loss. The median number of fillers for relations for the top 1% entities drops, e.g. *per:title* from 7 to 5, *per:employee_or_member_of* from 5 to 2, and *per:city_of_birth* from 3 to 1. Inspection shows that our approach addresses many obvious mistakes: *U.K.* is removed as a response to *(Securities and Exchange Commission(SEC), org:country_of_headquarters, ?)* while *U.S.* remains. The error, caused by *UK's SEC* which means *UK*'s analog to the *SEC* of *US*, is very hard to resolve without world knowledge. With cardinality constraints that favor only one *country_of_headquarters* for an ORG and *U.K* has a lower confidence than *U.S.* as a filler, the model identifies *U.K.* as an incorrect answer.

With entity factors added (KBC+E+R and KBC+E), the model favors a larger amount of smaller but more precise entities. It generates 4% new entities (out of 212K) by splitting the largest $< 0.1\%$ entities with the **SplitE** heuristics described in section 5. For example, the entity *Australia* is splitted into 3 entities, *Australia* and two outliers *West Aussie* and *Australian Capital Territory*. It also singles out entities such as *South America*, *Idaho*, *Colorado* from the giant *U.S.* entity with $> 20,000$ mentions. When querying the KB facts related to *U.S.*, erroneous answers that would otherwise be reported through relations associated with *South America* or the *U.S.* states will be removed.

7 Related Work

Cold Start KBP The TAC Cold Start KBP workshop has attracted many text-based KBP systems (McNamee et al., 2012; McNamee et al., 2013; Mayfield et al., 2014; Min et al., 2015; Roth et al., 2015; Angeli et al., 2014; Nguyen et al., 2014; Monahan and Carpenter, 2012). KELVIN (Mayfield et al., 2014) and BBN system (Min et al., 2015) both use hand-crafted rules to limit the number of fillers, e.g., remove less precise relations if a *person* has more than 8 (current and ex-) spouses. (Wolfe et al., 2015) and (He and Grishman, 2015) proposed interactive tools for KB construction with human guidance.

Knowledge Base Completion With the recent popularity of structured KBs such as Freebase (Bollacker et al., 2008), YAGO (Suchanek et al., 2007) and above-mentioned KBP techniques, there is a growing interest in completing a partially-complete KB with tensor decomposition (Chang et al., 2014), matrix factorization (Riedel et al., 2013), graph random walk (Lao et al., 2011; Lao et al., 2012; Gardner et al., 2014), neural networks (Socher et al., 2013; Neelakantan et al., 2015; Dong et al., 2014) and others (Guu et al., 2015; Gardner et al., 2013; Das et al., 2016). Knowledge Vault (Dong et al., 2014) pushes it further by combining many extraction components while estimating the confidence of their extractions and scales it to the Web. Model combination (Viswanathan et al., 2015) and confidence estimation (Wick et al., 2013; Li and Grishman, 2013) is related to our model for combining extraction components. The work described here differs from KB completion tasks in its requirement that the initial KB is empty and that all information in the KB be grounded in a text corpus.

Joint Modeling and Inference for IE To address the problem of compounding errors with multiple NLP components for IE, several papers (Finkel and Manning, 2009; Mccallum and Jensen, 2003; Finkel et al., 2006; Yao et al., 2010; Singh et al., 2009; Poon and Domingos, 2007; Wellner et al., 2004; Poon and Vanderwende, 2010; Riedel and McCallum, 2011; Chen et al., 2014; Kate and Mooney, 2010; Miwa and Sasaki, 2014) propose joint modeling and inference for IE. (Roth and Yih, 2007) use the ILP framework to enforce manually-specified constraints between entity and relation identification, while (Yu and Lam, 2010) models these two tasks in encyclopedia articles using a discriminative probabilistic model. (Li and Ji, 2014) jointly extracts entity mentions and relations with a structured perception with beam search. (Singh et al., 2013a) performs joint inference for entity, relation and coreference with an extension of the belief propagation algorithm. The work described here differs in its use of world knowledge. The joint modeling and inference for IE is not comparable but complementary to our method, therefore can be incorporated into our system for further gain.

8 Conclusion and Future Work

We present a joint probabilistic framework for end-to-end Cold Start KBP with prior world knowledge. Experiments show it surpassing the best-performing system at the NIST TAC 2015 Cold Start evaluation. We plan to investigate additional world knowledge in the near future.

Acknowledgments

The authors would like to thank Scott Miller, Ralph Weischedel, and the anonymous reviewers for their very helpful comments on improving the paper.

References

Gabor Angeli, Sonal Gupta, Melvin Johnson Premkumar, Christopher D Manning, Re Christopher, Julie Tibshirani, Jean Y Wu, Sen Wu, and Ce Zhang. 2014. Stanford's Distantly Supervised Slot Filling Systems for KBP 2014. In *Text Analysis Conference*, Gaithersburg, Maryland.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD '08*, page 1247, New York, New York, USA. ACM Press.

Kai-Wei Chang, Wen-tau Yih, Bishan Yang, and Christopher Meek. 2014. Typed Tensor Decomposition of Knowledge Bases for Relation Extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1568–1579, Doha, Qatar, oct. Association for Computational Linguistics.

Zheng Chen, Suzanne Tamang, Adam Lee, Xiang Li, Wen-Pin Lin, Javier Artiles, Matthew Snover, Marissa Passantino, and Heng Ji. 2010. CUNY-BLENDER TAC-KBP2010 Entity Linking and Slot Filling System Description. In *Text Analysis Conference*, Gaithersburg, Maryland.

Liwei Chen, Yansong Feng, Jinghui Mo, Songfang Huang, and Dongyan Zhao. 2014. Joint Inference for Knowledge Base Population. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1912–1923, Doha, Qatar, oct. Association for Computational Linguistics.

Rajarshi Das, Arvind Neelakantan, David Belanger, and Andrew McCallum. 2016. Incorporating Selectional Preferences in Multi-hop Relation Extraction. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 18–23, San Diego, CA, jun. Association for Computational Linguistics.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pages 601–610, New York, New York, USA. ACM Press.

Tim Finin, Paul McNamee, Dawn Lawrie, James Mayfield, and Craig Harman. 2014. Hot stuff at cold start: HLTCOE participation at TAC 2014. In *Text Analysis Conference*, Gaithersburg, Maryland.

Jenny Rose Finkel and Christopher D Manning. 2009. Joint Parsing and Named Entity Recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 326–334, Boulder, Colorado, jun. Association for Computational Linguistics.

Jenny Rose Finkel, Christopher D Manning, and Andrew Y Ng. 2006. Solving the Problem of Cascading Errors: Approximate Bayesian Inference for Linguistic Annotation Pipelines. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 618–626, Sydney, Australia, jul. Association for Computational Linguistics.

Matt Gardner, Partha Pratim Talukdar, Bryan Kisiel, and Tom Mitchell. 2013. Improving Learning and Inference in a Large Knowledge-Base using Latent Syntactic Cues. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 833–838, Seattle, Washington, USA, oct. Association for Computational Linguistics.

Matt Gardner, Partha Talukdar, Jayant Krishnamurthy, and Tom Mitchell. 2014. Incorporating Vector Space Similarity in Random Walk Inference over Knowledge Bases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 397–406, Doha, Qatar, oct. Association for Computational Linguistics.

Ralph Grishman. 2013. Off to a Cold Start: New York University's 2013 Knowledge Base Population Systems. In *Text Analysis Conference*, Gaithersburg, Maryland.

Kelvin Guu, John Miller, and Percy Liang. 2015. Traversing Knowledge Graphs in Vector Space. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages

318–327, Lisbon, Portugal, sep. Association for Computational Linguistics.

Yifan He and Ralph Grishman. 2015. ICE: Rapid Information Extraction Customization for NLP Novices. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 31–35, Denver, Colorado, jun. Association for Computational Linguistics.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA, jun. Association for Computational Linguistics.

Martin Jansche. 2005. Maximum Expected F-Measure Training of Logistic Regression Models. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 692–699, Vancouver, British Columbia, Canada, oct. Association for Computational Linguistics.

Heng Ji, David Westbrook, and Ralph Grishman. 2005. Using Semantic Relations to Refine Coreference Decisions. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 17–24, Vancouver, British Columbia, Canada, oct. Association for Computational Linguistics.

Rohit J Kate and Raymond Mooney. 2010. Joint Entity and Relation Extraction Using Card-Pyramid Parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 203–212, Uppsala, Sweden, jul. Association for Computational Linguistics.

Ni Lao, Tom Mitchell, and William W Cohen. 2011. Random Walk Inference and Learning in A Large Scale Knowledge Base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 529–539, Edinburgh, Scotland, UK., jul. Association for Computational Linguistics.

Ni Lao, Amarnag Subramanya, Fernando Pereira, and William W Cohen. 2012. Reading The Web with Learned Syntactic-Semantic Inference Rules. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1017–1026, Jeju Island, Korea, jul. Association for Computational Linguistics.

Xiang Li and Ralph Grishman. 2013. Confidence Estimation for Knowledge Base Population. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 396–401, Hissar, Bulgaria, sep. INCOMA Ltd. Shoumen, BULGARIA.

Qi Li and Heng Ji. 2014. Incremental Joint Extraction of Entity Mentions and Relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, Baltimore, Maryland, jun. Association for Computational Linguistics.

James Mayfield, Paul Mcnamee, Craig Harman, Tim Finin, and Dawn Lawrie. 2014. KELVIN: Extracting Knowledge from Large Text Collections. In *AAAI Fall Symposium on Natural Language Access to Big Data*, Arlington, Virginia.

James Mayfield. 2014. Cold Start Knowledge Base Population at TAC 2014. In *Text Analysis Conference*, Gaithersburg, Maryland.

Andrew Mccallum and David Jensen. 2003. A Note on the Unification of Information Extraction and Data Mining using Conditional-Probability, Relational Models. In *IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*, Acapulco, Mexico.

Paul McNamee, Veselin Stoyanov, James Mayfield, Tim Finin, Tan Xu, Douglas W. Oard, and Dawn Lawrie. 2012. HLTCOE Participation at TAC 2012: Entity Linking and Cold Start Knowledge Base Construction. In *Text Analysis Conference*, Gaithersburg, Maryland.

Paul McNamee, Tim Finin, Dawn Lawrie, and James Mayfield. 2013. HLTCOE Participation at TAC 2013. In *Text Analysis Conference*, Gaithersburg, Maryland.

Bonan Min, Marjorie Freedman, and Constantine Lignos. 2015. BBN's 2015 System for Cold Start Knowledge Base Population. In *Text Analysis Conference*, Gaithersburg, Maryland.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, aug. Association for Computational Linguistics.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling Joint Entity and Relation Extraction with Table Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869, Doha, Qatar, oct. Association for Computational Linguistics.

Sean Monahan and Dean Carpenter. 2012. Lorify: A Knowledge Base from Scratch. In *Text Analysis Conference*, Gaithersburg, Maryland.

Arvind Neelakantan, Benjamin Roth, and Andrew McCallum. 2015. Compositional Vector Space Models for Knowledge Base Completion. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 156–166, Beijing, China, jul. Association for Computational Linguistics.

Thien Huu Nguyen, Yifan He, Maria Pershina, Xiang Li, and Ralph Grishman. 2014. New york university 2014 knowledge base population systems. In *Text Analysis Conference*, Gaithersburg, Maryland.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. *LDC2011T07*.

Hoifung Poon and Pedro Domingos. 2007. Joint inference in information extraction. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 1*, pages 913–918, Vancouver, British Columbia, Canada. AAAI Press.

Hoifung Poon and Lucy Vanderwende. 2010. Joint Inference for Knowledge Extraction from Biomedical Literature. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 813–821, Los Angeles, California, jun. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nate Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A Multi-Pass Sieve for Coreference Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA, oct. Association for Computational Linguistics.

Lance Ramshaw, Elizabeth Boschee, Marjorie Freedman, Jessica MacBride, Ralph Weischedel, and Alex Zamanian. 2011. *SERIF language processing effective trainable language understanding*. SpringerLink : B{ü}cher. Springer New York.

Sebastian Riedel and Andrew McCallum. 2011. Fast and Robust Joint Models for Biomedical Event Extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1–12, Edinburgh, Scotland, UK., jul. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling Relations and Their Mentions without Labeled Text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Barcelona, Spain. Springer, Berlin, Heidelberg.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation Extraction with Matrix Factorization and Universal Schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84, Atlanta, Georgia, jun. Association for Computational Linguistics.

Dan Roth and Scott Wen-tau Yih, 2007. *Global Inference for Entity and Relation Identification via a Linear Programming Formulation*, pages 553–580. MIT Press, nov.

Benjamin Roth, Nicholas Monath, David Belanger, Emma Strubell, Patrick Verga, and Andrew McCallum. 2015. Building knowledge bases with universal schema: Cold start and slot-filling approaches. In *Text Analysis Conference*, Gaithersburg, Maryland.

Sameer Singh, Karl Schultz, and Andrew McCallum, 2009. *Bi-directional Joint Inference for Entity Resolution and Segmentation Using Imperatively-Defined Factor Graphs*, pages 414–429. Springer Berlin Heidelberg, Berlin, Heidelberg.

Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum. 2013a. Joint inference of entities, relations, and coreference. In *The 3rd Workshop on Automated Knowledge Base Construction*, pages 1–6, New York, New York, USA. ACM Press.

Sameer Singh, Limin Yao, David Belanger, Ari Kobren, Sam Anzaroot, Michael Wick, Alexandre Passos, Harshal Pandya, Jinho Choi, Brian Martin, and Andrew McCallum. 2013b. Universal Schema for Slot Filling and Cold Start: UMass IESL at TACKBP 2013. In *Text Analysis Conference*, Gaithersburg, Maryland.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 926–934. Curran Associates, Inc.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, pages 697–706, New York, New York, USA. ACM Press.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance Multi-label Learning for Relation Extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea, jul. Association for Computational Linguistics.

Vidhoon Viswanathan, Nazneen Fatema Rajani, Yinon Bentor, and Raymond Mooney. 2015. Stacked Ensembles of Information Extractors for Knowledge-Base Population. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 177–187, Beijing, China, jul. Association for Computational Linguistics.

Ben Wellner, Andrew McCallum, Fuchun Peng, and Michael Hay. 2004. An integrated, conditional model of information extraction and coreference with application to citation matching. In *the 20th conference on Uncertainty in artificial intelligence*, pages 593–601, Banff, Canada. AUAI Press.

Michael Wick, Sameer Singh, Ari Kobren, and Andrew McCallum. 2013. Assessing confidence of knowledge base content with an experimental study in entity resolution. In *Proceedings of the 2013 workshop on Automated knowledge base construction - AKBC '13*, pages 13–18, New York, New York, USA. ACM Press.

Travis Wolfe, Mark Dredze, James Mayfield, Paul McNamee, Craig Harman, Tim Finin, and Benjamin Van Durme. 2015. Interactive Knowledge Base Population. *CoRR*, abs/1506.0.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective Cross-Document Relation Extraction Without Labelled Data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1013–1023, Cambridge, MA, oct. Association for Computational Linguistics.

Nan Ye, Kian M Chai, Wee S Lee, and Hai L Chieu. 2012. Optimizing F-measure: A Tale of Two Approaches. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 289–296, New York, NY, USA. ACM.

Xiaofeng Yu and Wai Lam. 2010. Jointly Identifying Entities and Extracting Relations in Encyclopedia Text via A Graphical Model Approach. In *Coling 2010: Posters*, pages 1399–1407, Beijing, China, aug. Coling 2010 Organizing Committee.