Joint Inference for Knowledge Base Population

Liwei Chen[1], **Yansong Feng** [1*], **Jinghui Mo**[1], **Songfang Huang**[2], and **Dongyan Zhao**[1]
[1]ICST, Peking University, Beijing, China
[2]IBM China Research Lab, Beijing, China
{chenliwei,fengyansong,mojinghui,zhaodongyan}@pku.edu.cn
huangsf@cn.ibm.com

Abstract

Populating Knowledge Base (KB) with new knowledge facts from reliable text resources usually consists of linking name mentions to KB entities and identifying relationship between entity pairs. However, the task often suffers from errors propagating from upstream entity linkers to downstream relation extractors. In this paper, we propose a novel joint inference framework to allow interactions between the two subtasks and find an optimal assignment by addressing the coherence among preliminary local predictions: whether the types of entities meet the expectations of relations explicitly or implicitly, and whether the local predictions are globally compatible. We further measure the confidence of the extracted triples by looking at the details of the complete extraction process. Experiments show that the proposed framework can significantly reduce the error propagations thus obtain more reliable facts, and outperforms competitive baselines with state-of-the-art relation extraction models.

1 Introduction

Recent advances in natural language processing have made it possible to construct structured KBs from online encyclopedia resources, at an unprecedented scale and much more efficiently than traditional manual edit. However, in those KBs, entities which are popular to the community usually contain more knowledge facts, e.g., the basketball player *LeBron James*, the actor Nicholas Cage, etc., while most other entities often have fewer facts. On the other hand, knowledge facts should be updated as the development of entities, such as changes in the cabinet, a marriage event, or an acquisition between two companies, etc.

In order to address the above issues, we could consult populating existing KBs from reliable text resources, e.g., newswire, which usually involves enriching KBs with new entities and populating KBs with new knowledge facts, in the form of *<Entity, Relation, Entity>* triple. In this paper, we will focus on the latter, identifying relationship between two existing KB entities. This task can be intuitively considered in a pipeline paradigm, that is, name mentions in the texts are first linked to entities in the KB (entity linking, EL), and then the relationship between them are identified (relation extraction, RE). It is worth mentioning that the first task EL is different from the task of named entity recognition (NER) in traditional information extraction (IE) tasks, where NER recognizes and classifies the entity mentions (to several predefined types) in the texts, but EL focuses on linking the mentions to their corresponding entities in the KB. Such pipeline systems often suffer from errors propagating from upstream to downstream, since only the local best results are selected to the next step. One idea to solve the problem is to allow interactions among the local predictions of both subtasks and jointly select an optimal assignment to eliminate possible errors in the pipeline.

Let us first look at an example. Suppose we are extracting knowledge facts from two sentences in Figure 1: in sentence [1], if we are more confident to extract the relation *fb:org.headquarters*[1], we will be then prompted to select *Bryant University*, which indeed favors the RE prediction that requires an organization to be its subject. On the other side, if we are sure to link to *Kobe Bryant* in sentence [2], we will probably select *fb:pro_athlete.teams*, whose subject position expects an athlete, e.g., an NBA player. It is not difficult to see that the argument type expectations of relations can encourage the two subtasks interact with each other and select coherent predictions for

[1]The prefix *fb* means the relations are defined in Freebase.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1912–1923,
October 25-29, 2014, Doha, Qatar.

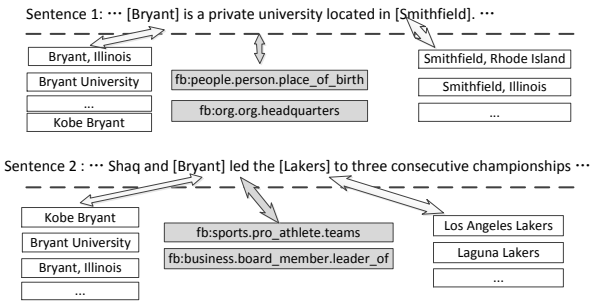

Figure 1: Two example sentences from which we can harvest knowledge facts.

both of them. In KBs with well-defined schemas, such as Freebase, type requirements can be collected and utilized explicitly (Yao et al., 2010). However, in other KBs with less reliable or even no schemas, it is more appropriate to implicitly capture the type expectations for a given relation (Riedel et al., 2013).

Furthermore, previous RE approaches usually process each triple individually, which ignores whether those local predictions are compatible with each other. For example, suppose the local predictions of the two sentences above are *<Kobe Bryant, fb:org.headquarters, Smithfield, Rhode Island>* and *<Kobe Bryant, fb:pro_athlete.teams, Los Angeles Lakers>*, respectively, which, in fact, disagree with each other with respect to the KB, since, in most cases, these two relations cannot share subjects. Now we can see that either the relation predictions or the EL results for "*Bryant*" are incorrect. Those disagreements provide us an effective way to remove the possible incorrect predictions that cause the incompatibilities.

On the other hand, the automatically extracted knowledge facts inevitably contain errors, especially for those triples collected from open domain. Extractions with confidence scores will be more than useful for users to make proper decisions according to their requirements, such as trading recall for precision, or supporting approximate queries.

In this paper, we propose a joint framework to populate an existing KB with new knowledge facts extracted from reliable text resources. The joint framework is designed to address the error propagation issue in a pipeline system, where subtasks are optimized in isolation and locally. We find an optimal configuration from top k results of both subtasks, which maximizes the scores of each step, fulfills the argument type expectations of relations, which can be captured explicitly or implicitly, in the KB, and avoids globally incoherent predictions. We formulate this optimization problem in an Integer Linear Program (ILP) framework, and further adopt a logistic regression model to measure the reliability of the whole process, and assign confidences to all extracted triples to facilitate further applications. The experiments on a real-world case study show that our framework can eliminate error propagations in the pipeline systems by taking relations' argument type expectations and global compatibilities into account, thus outperforms the pipeline approaches based on state-of-the-art relation extractors by a large margin. Furthermore, we investigate both explicit and implicit type clues for relations, and provide suggestions about which to choose according to the characteristics of existing KBs. Additionally, our proposed confidence estimations can help to achieve a precision of over 85% for a considerable amount of high quality extractions.

In the rest of the paper, we first review related work and then define the knowledge base population task that we will address in this paper. Next we detail the proposed framework and present our experiments and results. Finally, we conclude this paper with future directions.

2 Related Work

Knowledge base population (KBP), the task of extending existing KBs with entities and relations, has been studied in the TAC-KBP evaluations (Ji et al., 2011), containing three tasks. The entity linking task links entity mentions to existing KB nodes and creates new nodes for the entities absent in the current KBs, which can be considered as a kind of entity population (Dredze et al., 2010; Tamang et al., 2012; Cassidy et al., 2011). The slot-filling task populates new relations to the KB (Tamang et al., 2012; Roth et al., 2012; Liu and Zhao, 2012), but the relations are limited to a predefined sets of attributes according to the types of entities. In contrast, our RE models only require minimal supervision and do not need well-annotated training data. Our framework is therefore easy to adapt to new scenarios and suits real-world applications. The cold-start task aims at constructing a KB from scratch in a slot-filling style (Sun et al., 2012; Monahan and Carpenter, 2012).

Entity linking is a crucial part in many KB re-

lated tasks. Many EL models explore local contexts of entity mentions to measure the similarity between mentions and candidate entities (Han et al., 2011; Han and Sun, 2011; Ratinov et al., 2011; Cheng and Roth, 2013). Some methods further exploit global coherence among candidate entities in the same document by assuming that these entities should be closely related (Han et al., 2011; Ratinov et al., 2011; Sen, 2012; Cheng and Roth, 2013). There are also some approaches regarding entity linking as a ranking task (Zhou et al., 2010; Chen and Ji, 2011). Lin et al. (2012) propose an approach to detect and type entities that are currently not in the KB.

Note that the EL task in KBP is different from the name entity mention extraction task, mainly in the ACE task style, which mainly identifies the boundaries and types of entity mentions and does not explicitly link entity mentions into a KB (ACE, 2004; Florian et al., 2006; Florian et al., 2010; Li and Ji, 2014), thus are different from our work.

Meanwhile, relation extraction has also been studied extensively in recent years, ranging from supervised learning methods (ACE, 2004; Zhao and Grishman, 2005; Li and Ji, 2014) to unsupervised open extractions (Fader et al., 2011; Carlson et al., 2010). There are also models, with distant supervision (DS), utilizing reliable texts resources and existing KBs to predict relations for a large amount of texts (Mintz et al., 2009; Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012). These distantly supervised models can extract relations from texts in open domain, and do not need much human involvement. Hence, DS is more suitable for our task compared to other traditional RE approaches.

Joint inference over multiple local models has been applied to many NLP tasks. Our task is different from the traditional joint IE works based in the ACE framework (Singh et al., 2013; Li and Ji, 2014; Kate and Mooney, 2010), which jointly extract and/or classify named entity mentions to several predefined types in a sentence and identify in a sentence level which relation this specific sentence describes (between a pair of entity mentions in this sentence). Li and Ji (2014) follow the ACE task definitions and present a neat incremental joint framework to simultaneously extract entity mentions and relations by structure perceptron. In contrast, we link entity mentions from a text corpus to their corresponding entities in an existing KB and identify the relations between pairs of entities based on that text corpus. Choi et al. (2006) jointly extracts the expressions and sources of opinion as well as the linking relations (i.e., a source entity expresses an opinion expression) between them, while we focus on jointly modeling EL and RE in open domain, which is a different and challenging task.

Since the automatically extracted knowledge facts inevitably contain errors, many approaches manage to assign confidences for those extracted facts (Fader et al., 2011; Wick et al., 2013). Wick et al. (2013) also point out that confidence estimation should be a crucial part in the automated KB constructions and will play a key role for the wide applications of automatically built KBs. We thus propose to model the reliability of the complete extraction process and take the argument type expectations of the relation, coherence with other predictions and the triples in the existing KB into account for each populated triple.

3 Task definition

We formalize our task as follows. Given a set of entities sampled from an existing KB, $E = \{e_1, e_2, ..., e_{|E|}\}$, a set of canonicalized relations from the same KB, $R = \{r_1, r_2, ..., r_{|R|}\}$, a set of sentences extracted from news corpus, $SN = \{sn_1, sn_2, ..., sn_{|SN|}\}$, each contains two mentions m_1 and m_2 whose candidate entities belong to E, a set of text fragments $\mathcal{T} = \{t_1, t_2, ..., t_{|\mathcal{T}|}\}$, where t_i contains its corresponding target sentence sn_i and acts as its context. Our task is to link those mentions to entities in the given KB, identify the relationship between entity pairs and populate new knowledge facts into the KB.

4 The Framework

We propose to perform joint inference over subtasks involved. For each sentence with two entity mentions, we first employ a preliminary EL model and RE model to obtain entity candidates and possible relation candidates between the two mentions, respectively. Our joint inference framework will then find an optimal assignment by taking the preliminary prediction scores, the argument type expectations of relations and the global compatibilities among the predictions into account. In the task of KBP, an entity pair may appear in multiple sentences as different relation instances, and the crucial point is whether we can identify all the cor-

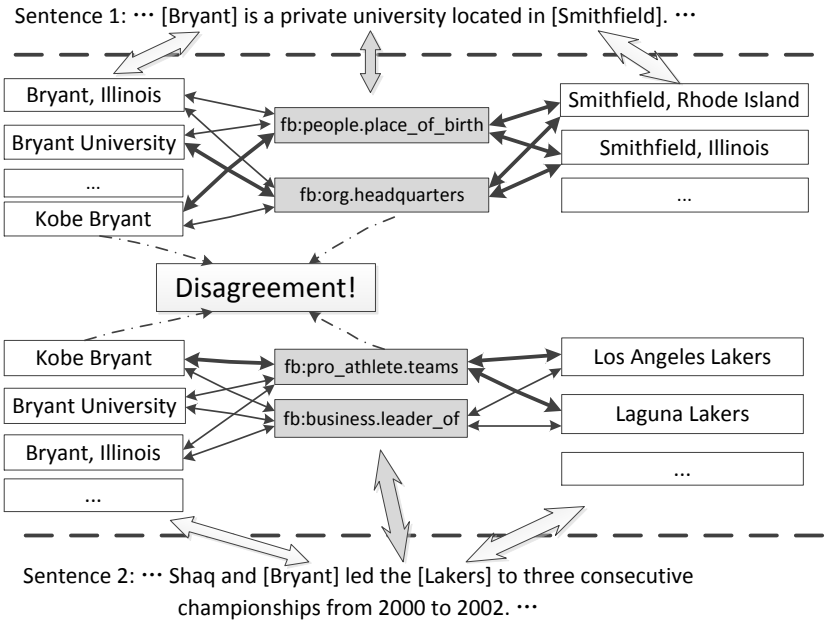

Figure 2: An example of our joint inference framework. The top and bottom are two example sentences with entity mentions in the square brackets, candidate entities in the white boxes, candidate relations in the grey boxes, and the solid lines with arrows between relations and entities represent their preference scores, with thickness indicating the preferences' value.

rect relations for an entity pair. Thus, after finding an optimal sentence-level assignment, we aggregate those local predictions by ORing them into the entity pair level. Finally, we employ a regression model to capture the reliability of the complete extraction process.

4.1 Preliminary Models

Entity Linking The preliminary EL model can be any approach which outputs a score for each entity candidate. Note that a recall-oriented model will be more than welcome, since we expect to introduce more potentially correct local predictions into the inference step. In this paper, we adopt an unsupervised approach in (Han et al., 2011) to avoid preparing training data. Note the challenging NIL problem, i.e., identifying which entity mentions do not have corresponding entities in the KB (labeled as NIL) and clustering those mentions, will be our future work. For each mention we retain the entities with top p scores for the succeeding inference step.

Relation Extraction The choice of RE model is also broad. Any sentence level extractor whose results are easy to be aggregated to entity pair level can be utilized here (again, a recall-oriented version will be welcome), such as Mintz++ mentioned in (Surdeanu et al., 2012), which we adapt into a Maximum Entropy version. We also include a special label, *NA*, to represent the case where there is no predefined relationship between an entity pair. For each sentence, we retain the relations with top q scores for the inference step, and we also call that this sentence **supports** those candidate relations. As for the features of RE models, we use the same features (lexical features and syntactic features) with the previous works (Chen et al., 2014; Mintz et al., 2009; Riedel et al., 2010; Hoffmann et al., 2011).

4.2 Relations' Expectations for Argument Types

In most KBs' schemas, canonicalized relations are designed to expect specific types of entities to be their arguments. For example, in Figure 2, it is more likely that an entity *Kobe Bryant* takes the subject position of a relation *fb:pro_athlete.teams*, but it is unlikely for this entity to take the subject position of a relation *fb:org.headquarters*. Making use of these type requirements can encourage the framework to select relation and entity candidates which are coherent with each other, and discard incoherent choices.

In order to obtain the preference scores between

the entities in E and the relations in R, we generate two matrices with $|E|$ rows and $|R|$ columns, whose elements sp_{ij} indicates the preference score of entity i and relation j. The matrix S_{subj} is for relations and their subjects, and the matrix S_{obj} is for relations and their objects. We initialize the two matrices using the KB as follows: for entity i and relation j, if relation j takes entity i as its subject/object in the KB, the element at the position (i, j) of the corresponding matrix will be 1, otherwise it will be 0. Note that in our experiments, we do not count the triples that are evaluated in the testing data, to build the matrices. Now the problem is how we can obtain the unknown elements in the matrices.

Explicit Type Information Intuitively, we should examine whether the explicit types of the entities fulfill the expectations of relations in the KB. For each unknown element $S_{subj}(i, j)$, we first obtain the type of entity i, which is collected from the lowest level of the KB's type hierarchy, and examine whether there is another entity with the same type taking the subject position of relation j in the initial matrix. If such an entity exists, $S_{subj}(i, j) = 1$, otherwise 0. For example, for the subject *Jay Fletcher Vincent* and the relation *fb:pro_athlete.teams*, we first obtain the subject's type *basketball_player*, and then we go through the initial matrix and find another entity *Kobe Bryant* with the same type taking the subject position of *fb:pro_athlete.teams*, indicating that *Jay Fletcher Vincent* may take the relation *fb:pro_athlete.teams*. The matrix S_{obj} is processed in the same way.

Implicit Type Expectations In practice, few KBs have well-defined schemas. In order to make our framework more flexible, we need to come up with an approach to implicitly capture the relations' type expectations, which will also be represented as preference scores.

Inspired by Riedel et al. (2013) who use a matrix factorization approach to capture the association between textual patterns, relations and entities based on large text corpora, we adopt a collaborative filtering (CF) method to compute the preference scores between entities and relations based on the statistics obtained from an existing KB.

In CF, the preferences between customers and items are calculated via matrix factorization over the initial customer-item matrix. In our framework, we compute the preference scores between entities and relations via the same approach over the two initialized matrices S_{subj} and S_{obj}, resulting in two entity-relation matrices with estimated preference values. We use ALS-WR (Zhou et al., 2008) to process the matrices and compute the preference of a relation taking an entity as its subject and object, respectively. We normalize the preference scores of each entity using their means μ and standard deviations σ.

4.3 Compatibilities among Predicted Triples

The second aspect we investigate is whether the extracted triples are compatible with respect to all other knowledge facts. For example, according to the KB, the two relations *fb:org.headquarters* and *fb:pro_athlete.teams* in Figure 2 cannot share the same entity as their subjects. So if such sharing happens, that will indicate either the predictions of the relations or the entities are incorrect. The clues can be roughly grouped into three categories, namely whether two relations can share the same subjects, whether two relations can share the same objects, and whether one relation's subject can be the other relation's object.

Global compatibilities among local predictions have been investigated by several joint models (Li et al., 2011; Li and Ji, 2014; Chen et al., 2014) to eliminate the errors propagating in a pipeline system. Specifically, Chen et al. (2014) utilized the clues with respect to the compatibilities of relations in the task of relation extraction. Following (Li et al., 2011; Chen et al., 2014), we extend the idea of global compatibilities to the entity and relation predictions during knowledge base population. We examine the pointwise mutual information (PMI) between the argument sets of two relations to collect such clues. For example, if we want to learn whether two relations can share the same subject, we first collect the subject sets of both relations from the KB, and then compute the PMI value between them. If the value is lower than a certain threshold (set to -3 in this paper), the clue that *the two relations cannot share the same subject* is added. These clues can be easily integrated into an optimization framework in the form of constraints.

4.4 Integer Linear Program Formulation

Now we describe how we aggregate the above components, and formulate the joint inference problem into an ILP framework. For each candi-

date entity e of mention m in text fragment t, we define a boolean decision variable $d_t^{m,e}$, which denotes whether this entity is selected into the final configuration or not. Similarly, for each candidate relation r of fragment t, we define a boolean decision variable d_t^r. In order to introduce the preference scores into the model, we also need a decision variable $d_t^{r,m,e}$, which denotes whether both relation r and candidate entity e of mention m are selected in t.

We use $s_{el}^{t,m,e}$ to represent the score of mention m in t disambiguated to entity e, which is output by the EL model, $s_{re}^{t,r}$ representing the score of relation r assigned to t, which is output by the RE model, $s_p^{r,e}$ the explicit/implicit preference score between relation r and entity e.

Our goal is to find the best assignment to the variables d_t^r and $d_t^{m,e}$, such that it maximizes the overall scores of the two subtasks and the coherence among the preliminary predictions, while satisfying the constraints between the predicted triples as well. Our objective function can be written as:

$$\max el \times conf^{ent} + re \times conf^{rel} + sp \times coh^{e-r} \tag{1}$$

where el, re and sp are three weighting parameters tuned on development set. $conf^{ent}$ is the overall score of entity linking:

$$conf^{ent} = \sum_t \sum_{m \in M(t)} \sum_{e \in C_e(m)} s_{el}^{t,m,e} d_t^{m,e} \tag{2}$$

where $M(t)$ is the set of mentions in t, $C_e(m)$ is the candidate entity set of the mention m. $conf^{rel}$ represents the overall score of relation extraction:

$$conf^{rel} = \sum_t \sum_{r \in C_r(t)} s_{re}^{t,r} d_t^r \tag{3}$$

where $C_r(t)$ is the set of candidate relations in t. coh^{e-r} is the coherence between the candidate relations and entities in the framework:

$$coh^{e-r} = \sum_t \sum_{r \in C_r(t)} \sum_{m \in M(t)} \sum_{e \in C_e(m)} s_p^{r,e} d_t^{r,m,e} \tag{4}$$

Now we describe the constraints used in our ILP problem. The first kind of constraints is introduced to ensure that each mention should be disambiguated to only one entity:

$$\forall t, \forall m \in M(t), \quad \sum_{e \in C_e(m)} d_t^{m,e} \leq 1 \tag{5}$$

The second type of constraints ensure that each entity mention pair in one sentence can only take one relation label:

$$\forall t, \quad \sum_{r \in C_r(t)} d_r^t \leq 1 \tag{6}$$

The third is introduced to ensure the decision variable $d_t^{r,m,e}$ equals 1 if and only if both the corresponding variables d_t^r and $d_t^{m,e}$ equal 1.

$$\forall t, \forall r \in C_r(t), \forall m \in M(t), \forall e \in C_e(m)$$

$$d_t^{r,m,e} \leq d_t^r \tag{7}$$

$$d_t^{r,m,e} \leq d_t^{m,e} \tag{8}$$

$$d_t^r + d_t^{m,e} \leq d_t^{r,m,e} + 1 \tag{9}$$

As for the compatibility constraints, we need to introduce another type of boolean decision variables. If a mention m_1 in t_1 and another mention m_2 in t_2 share an entity candidate e, we add a variable y for this mention pair, which equals 1 if and only if both $d_{t_1}^{m_1,e}$ and $d_{t_2}^{m_2,e}$ equal 1. So we add the following constraints for each mention pair m_1 and m_2 satisfies the previous condition:

$$y \leq d_{t_1}^{m_1,e} \tag{10}$$

$$y \leq d_{t_2}^{m_2,e} \tag{11}$$

$$d_{t_1}^{m_1,e} + d_{t_2}^{m_2,e} \leq y + 1 \tag{12}$$

Then we further add the following constraints for each mention pair to avoid incompatible predictions:

$$\forall r_1 \in C_r(t_1), r_2 \in C_r(t_2)$$

$$\text{If } (r_1, r_2) \in \mathcal{C}^{sr}, p(m_1) = subj, p(m_2) = subj$$

$$d_{t_1}^{r_1} + d_{t_2}^{r_2} + y \leq 2 \tag{13}$$

$$\text{If } (r_1, r_2) \in \mathcal{C}^{ro}, p(m_1) = obj, p(m_2) = obj$$

$$d_{t_1}^{r_1} + d_{t_2}^{r_2} + y \leq 2 \tag{14}$$

$$\text{If } (r_1, r_2) \in \mathcal{C}^{sro}, p(m_1) = obj, p(m_2) = subj$$

$$d_{t_1}^{r_1} + d_{t_2}^{r_2} + y \leq 2 \tag{15}$$

where $p(m)$ returns the position of mention m, either $subj$ (subject) or obj (object). $\mathcal{C}^{sr}$ is the pairs of relations which cannot share the same subject, $\mathcal{C}^{ro}$ is the pairs of relations which cannot share the same object, $\mathcal{C}^{sro}$ is the pairs of relations in which one relation's subject cannot be the other one's object.

We use IBM ILOG Cplex[2] to solve the above ILP problem.

[2]http://www.cplex.com

Table 1: The features used to calculate the confidence scores.

Type	Feature
Real	The RE score of the relation.
Real	The EL score of the subject.
Real	The EL score of the object.
Real	The preference score between the relation and the subject.
Real	The preference score between the relation and the object.
Real	The ratio of the highest and the second highest relation score in this entity pair.
Real	The ratio of the current relation score and the maximum relation score in this entity pair.
Real	The ratio of the number of sentences supporting the current relation and the total number of sentences in this entity pair.
Real	Whether the extracted triple is coherent with the KB according to the constraints in Section 4.3.

4.5 Confidence Estimation for Extracted Triples

The automatically extracted triples inevitably contain errors and are often considered as with high recall but low precision. Since our aim is to populate the extracted triples into an existing KB, which requires highly reliable knowledge facts, we need a measure of confidence for those extracted triples, so that others can properly utilize them.

Here, we use a logistic regression model to measure the reliability of the process, how the entities are disambiguated, how the relationships are identified, and whether those predictions are compatible. The features we used are listed in Table 1, which are all efficiently computable and independent from specific relations or entities. We manually annotate 1000 triples as correct or incorrect to prepare the training data.

5 Experiments

We evaluate the proposed framework in a real-world scenario: given a set of news texts with entity mentions and a KB, a model should find more and accurate new knowledge facts between pairs of those entities.

5.1 Dataset

We use New York Times dataset from 2005 to 2007 as the text corpus, and Freebase as the KB. We divide the corpus into two equal parts, one for creating training data for the RE models using the distant supervision strategy (we do not need training data for EL), and the other as the testing data. For the convenience of experimentation, we randomly sample a subset of entities for testing. We first collect all sentences containing two mentions which may refer to the sampled entities, and prune them according to: (1)there should be no more than 10 words between the two mentions; (2)the prior probability of the mention referring to the target entity is higher than a threshold (set to 0.1 in this paper), which is set to filter the impossible mappings; (3)the mention pairs should not belong to different clauses. The resulting test set is split into 10 parts and a development set, each with 3500 entity pairs roughly, which leads to averagely 200,000 variables and 900,000 constraints per split and may take 1 hour for Cplex to solve. Note that we do not count the triples that will be evaluated in the testing data when we learn the preferences and the clues from the KB.

5.2 Experimental Setup

We compare our framework with three baselines. The first one, *ME-pl*, is the pipeline system constructed by the entity linker in (Han et al., 2011) and the MaxEnt version of Mintz++ extractor mentioned in (Surdeanu et al., 2012). The second and third baselines are the pipeline systems constructed by the same linker and two state-of-the-art DS approaches, MultiR (Hoffmann et al., 2011) and MIML-RE (Surdeanu et al., 2012), respectively. They are referred to as *MultiR-pl* and *MIML-pl* in the rest of this paper.

We also implement several variants of our framework to investigate the following two components in our framework: whether to use explicit (E) or implicit (I) argument type expectations, whether to take global (G) compatibilities into account, resulting in four variants: *ME-JE*, *ME-JI*, *ME-JEG*, *ME-JIG*.

We tune the parameters in the objective function on the development set to be $re = 1$, $el = 4$, $sp = 1$. The numbers of preliminary results retained to the inference step are set to $p = 2$, $q = 3$. Three metrics used in our experiments include: (1)the precision of extracted triples, which is the ratio of the number of correct triples and the number of total extracted triples; (2)the number of correct triples (NoC); (3)the number of correct triples in the results ranked in top n. The third metric is crucial for KBP, since most users are only interested in the knowledge facts with high confidences. We compare the extracted triples against

Table 2: The results of our joint frameworks and the three baselines.

Approach	Precision	NoC	Top 50	Top 100
ME-pl	28.7 ± 0.8	725 ± 12	38 ± 2	75 ± 4
MultiR-pl	31.0 ± 0.8	647 ± 15	39 ± 2	71 ± 3
MIML-pl	33.2 ± 0.6	608 ± 16	40 ± 3	74 ± 5
ME-JE	32.8 ± 0.7	768 ± 10	46 ± 2	90 ± 3
ME-JEG	34.2 ± 0.5	757 ± 8	46 ± 2	90 ± 3
ME-JI	34.5 ± 1.0	784 ± 9	43 ± 3	88 ± 3
ME-JIG	35.7 ± 1.0	772 ± 8	43 ± 3	88 ± 4

Freebase to compute the precision, which may underestimate the performance since Freebase is incomplete. Since we do not have exact annotations for the EL, it is difficult to calculate the exact recall. We therefore use NoC instead. We evaluate our framework on the 10 subsets of the testing dataset and compute their means and standard deviations.

5.3 Overall Performance

We are interested to find out: **(a)**whether the task benefits from the joint inference i.e., can we collect more and correct facts? Or with a higher precision? **(b)** whether the argument type expectations (explicit and implicit) and global compatibility do their jobs as we expected? And, how do we choose from these components ? **(c)**whether the framework can work with other RE models? **(d)**whether we can find a suitable approach to measure the confidence or uncertainty during the extraction so that users or other applications can better utilize the extracted KB facts?

Let us first look at the performance of the baselines and our framework in Table 2 for an overview. Comparing the three pipeline systems, we can discover that using the same entity linker, *MIML-pl* performs the best in precision with slightly fewer correct triples, while *ME-pl* performs the worst. It is not surprising, *ME-pl*, as a strong and high-recall baseline, outputs the most correct triples. As for the results with high confidences, *MultiR-pl* outputs more correct triples in the top 50 results than *ME-pl*, and *MIML-pl* performs better or comparable than *ME-pl* in top n results.

After performing the joint inference, *ME-JE* improves *ME-pl* with 4.1% in precision and 43 more correct triples averagely, and results in better performance in top n results. By taking global compatibilities into consideration, *ME-JEG* further improve the precision to 34.2% in average with slightly fewer correct triples, indicating that

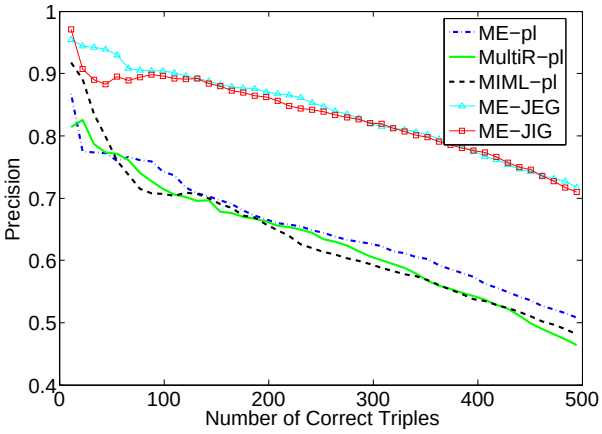

Figure 3: The numbers of correct triples v.s. the precisions for different approaches.

both argument type expectations and global compatibilities are useful in improving the performance: argument type information can help to select the correct and coherent predictions from the candidates EL and RE outputs, while global compatibilities can further prune incorrect triples that cause disagreements, although a few correct ones may be incorrectly eliminated. We can also observe that *ME-JIG* performs even higher than *ME-JEG* in overall precision, but *ME-JEG* collects more correct triples than *ME-JIG* in the top n predictions, showing that explicit type expectations with more accurate type information may perform better in high confidence results.

Furthermore, even though *MultiR-pl* and *MIML-pl* are based on state-of-the-art RE approaches, our model (for example, *ME-JIG*) can still outperform them in terms of all metrics, with 4.7% higher in precision than *MultiR-pl*, 2.5% higher than *MIML-pl*. Our model can extract 125 more correct triples than *MultiR-pl*, 164 more than *MIML-pl*, and perform better in top n results as well.

In previous RE tasks, Precision-Recall curves are mostly used to evaluate the systems' performances. In our task, since it is difficult to calculate the recall exactly, we use the number of correct triples instead, and plot curves of Precision-NoC to show the performance of the competitors and our approaches in more detail. For each value of NoC, the precision is the average of the ten splits of the testing dataset.

As shown in Figure 3, our approaches (*ME-JEG* and *ME-JIG*) obtain higher precisions on each NoC value, and the curves are much smoother than

Table 3: The results of our joint frameworks with MultiR sentence extractor.

Approach	Precision	NoC	Top 50	Top 100
MultiR-pl	31.0 ± 0.8	647 ± 15	39 ± 2	71 ± 3
MultiR-JEG	36.9 ± 0.8	687 ± 15	46 ± 2	88 ± 3
MultiR-JIG	38.5 ± 0.9	700 ± 15	45 ± 2	88 ± 3

the pipeline systems, indicating that our framework is more suitable for harvesting high quality knowledge facts. Comparing the two kinds of type clues, we can see that explicit ones perform better when the confidence control is high and the number of correct triples is small, and then the two are comparable. Since the precision of the triples with high confidences is crucial for the task of KBP, we still suggest choosing the explicit ones when there is a well-defined schema available in the KB, although implicit type expectations can result in higher overall precision.

5.4 Adapting MultiR Sentence Extractor into the Framework

The preliminary relation extractor of our framework is not limited to the MaxEnt[3] extractor. It can be any sentence level recall-oriented relation extractors. To further investigate the generalization of our joint inference framework, we also try to fit other sentence level relation extractors into the framework. Considering that MIML-RE does not output sentence-level results, we only adapt MultiR, with both global compatibilities and explicit/implicit type expectations, named as *MultiR-JEG* and *MultiR-JIG*, respectively. Since the scores output by the original MultiR are unnormalized, which are difficult to directly apply to our framework, we normalize their scores and retune the framework's parameters accordingly. The parameters are set to $re = 1$, $el = 32$, $sp = 16$.

As seen in Table 3, *MultiR-JEG* helps MultiR obtain about 40 more correct triples in average, and achieves 5.9% higher in precision, as well as significant improvements in top n correct predictions. As for *MultiR-JIG*, the improvements are 7.5% in precision and 53 in number of correct triples. In terms of top n results, the explicit and implicit type expectations perform comparable. We also observe that our framework improves MultiR as much as it does to MaxEnt, indicating our joint framework can generalize well in different RE models.

[3] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

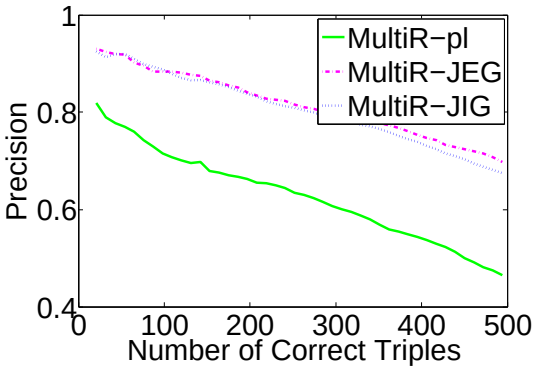

Figure 4: The numbers of correct triples v.s. the precisions for approaches with MultiR extractor.

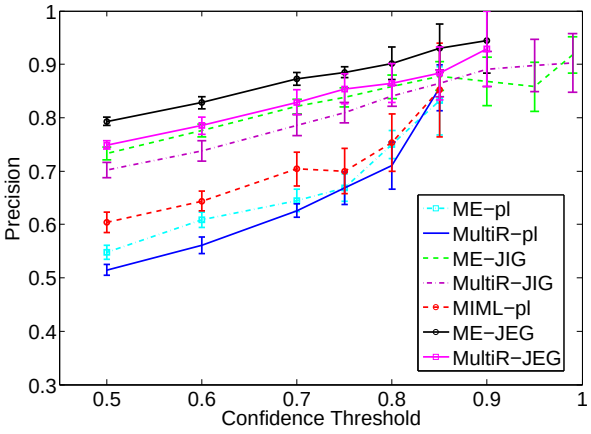

Figure 5: The precisions of different models under different confidence thresholds. The error bars represents the standard deviations of the results.

We further plot Precision-NoC curves for *MultiR-JEG* and *MultiR-JIG* in Figure 4, showing that our framework can result in better performance and smoother curves with MultiR extractor. It is interesting to see that with MultiR extractor, the two kinds of expectations perform comparably.

5.5 Results with Confidence Estimations

Now, we will investigate the results from another perspective with the help of confidence estimations. We calculate the precisions of the competitors and our approaches on different confidence thresholds from 0.5 to 1. The results are summarized in Figure 5. Note that the results across different approaches are not directly comparable, we put them in the same figure only to save space.

In Figure 5, intuitively, as the confidence threshold goes up, the extraction precisions should increase, indicating triples with higher confidences are more likely to be correct. However,

lower thresholds tend to result in estimations with smaller standard derivations due to those precisions are estimated over much more triples than those with higher thresholds, which means the randomness will be smaller.

On the other hand, our joint frameworks provide more evidences that can be used to well capture the reliability of an extraction. For example, the precisions of *Multir-JIG* and *ME-JIG* both stay around 85% when the confidence is higher than 0.85, with about 120 correct triples, indicating that by setting a proper threshold, we can obtain considerable amount of high quality knowledge facts at an acceptable precision, which is crucial for KBP. However, we cannot harvest such amount of high quality knowledge facts from the other three pipeline systems.

6 Conclusions

In this paper, we propose a joint framework for the task of populating KBs with new knowledge facts, which performs joint inference on two subtasks, maximizes their preliminary scores, fulfills the type expectations of relations and avoids global incompatibilities with respect to all local predictions to find an optimal assignment. Experimental results show that our framework can significantly eliminate the error propagations in pipeline systems and outperforms competitive pipeline systems with state-of-the-art RE models. Regarding the explicit argument type expectations and the implicit ones, the latter can result in a higher overall precision, while the former performs better in acquiring high quality knowledge facts with higher confidence control, indicating that if the KB has a well-defined schema we can use explicit type requirements for the KBP task, and if not, our model can still perform well by mining the implicit ones. Our framework can also generalize well with other preliminary RE models. Furthermore, we assign extraction confidences to all extracted facts to facilitate further applications. By setting a suitable threshold, our framework can populate high quality reliable knowledge facts to existing KBs.

For future work, we will address the NIL issue of EL where we currently assume all entities should be linked to a KB. It would be also interesting to jointly model the two subtasks through structured learning, instead of joint inference only. Currently we only use the coherence of extracted triples and the KB to estimate confidences, which would be nice to directly model the issue in a joint model.

Acknowledgments

We would like to thank Heng Ji, Kun Xu, Dong Wang and Junyang Rao for their helpful discussions and the anonymous reviewers for their insightful comments that improved the work considerably. This work was supported by the National High Technology R&D Program of China (Grant No. 2012AA011101, 2014AA015102), National Natural Science Foundation of China (Grant No. 61272344, 61202233, 61370055) and the joint project with IBM Research. Any correspondence please refer to Yansong Feng.

References

ACE. 2004. The automatic content extraction projects. http://projects.ldc.upenn.edu/ace.

Andrew Carlson, Justin Betteridge, Byran Kisiel, Burr Settles, Estevam Hruschka Jr., and Tom Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 1306–1313. AAAI Press.

Taylor Cassidy, Zheng Chen, Javier Artiles, Heng Ji, Hongbo Deng, Lev-Arie Ratinov, Jing Zheng, Jiawei Han, and Dan Roth. 2011. Entity linking system description. In *TAC2011*.

Zheng Chen and Heng Ji. 2011. Collaborative ranking: A case study on entity linking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 771–781, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liwei Chen, Yansong Feng, Songfang Huang, Yong Qin, and Dongyan Zhao. 2014. Encoding relation requirements for relation extraction via joint inference. In *Proceedings of the 52nd Annual Meeting on Association for Computational Linguistics*, ACL 2014, pages 818–827, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *EMNLP*.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 431–439, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Coling2010*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radu Florian, Hongyan Jing, Nanda Kambhatla, and Imed Zitouni. 2006. Factorizing complex models: A case study in mention detection. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 473–480. Association for Computational Linguistics.

Radu Florian, John F Pitrelli, Salim Roukos, and Imed Zitouni. 2010. Improving mention detection robustness to noisy input. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 335–345. Association for Computational Linguistics.

Xianpei Han and Le Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of ACL*, HLT '11, pages 945–954, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xianpei Han, Le Sun, and Jun Zhao. 2011. Collective entity linking in web text: a graph-based method. In *SIGIR*, SIGIR '11, pages 765–774, New York, NY, USA. ACM.

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 ACL-HLT - Volume 1*, HLT '11, pages 541–550, Stroudsburg, PA, USA. ACL.

Heng Ji, Ralph Grishman, and Hoa Dang. 2011. Overview of the tac2011 knowledge base population track. In *Proceedings of TAC*.

Rohit J. Kate and Raymond J. Mooney. 2010. Joint entity and relation extraction using card-pyramid parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 203–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting on Association for Computational Linguistics*, ACL 2014, pages 402–412, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qi Li, Sam Anzaroot, Wen-Pin Lin, Xiang Li, and Heng Ji. 2011. Joint inference for cross-document information extraction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2225–2228, New York, NY, USA. ACM.

Thomas Lin, Mausam, and Oren Etzioni. 2012. No noun phrase left behind: Detecting and typing unlinkable entities. In *Proceedings of the 2012 EMNLP-CoNLL*, EMNLP-CoNLL '12, pages 893–903, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fang Liu and Jun Zhao. 2012. Sweat2012: Pattern based english slot filling system for knowledge base population at tac 2012. In *TAC2012*.

Mike Mintz, Steven Bills, Rion Snow, and Dan 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 IJCNLP of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011.

Sean Monahan and Dean Carpenter. 2012. Lorify: A knowledge base from scratch. In *TAC2012*.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *ACL*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, pages 148–163. Springer Berlin / Heidelberg.

Sebastian Riedel, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 2013. Relation extraction with matrix factorization and universal schemas. In *Joint Human Language Technology Conference/Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '13)*, June.

Benjamin Roth, Grzegorz Chrupala, Michael Wiegand, Mittul Singh, and Dietrich Klakow. 2012. Generalizing from freebase and patterns using cluster-based distant supervision for tac kbp slotfilling 2012. In *TAC2012*.

Prithviraj Sen. 2012. Collective context-aware topic models for entity disambiguation. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 729–738, New York, NY, USA. ACM.

Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum. 2013. Joint inference of entities, relations, and coreference. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 1–6, New York, NY, USA. ACM.

Ang Sun, Xin Wang, Sen Xu, Yigit Kiran, Shakthi Poornima, Andrew Borthwick, , and Ralph Grishman. 2012. Intelius-nyu tac-kbp2012 cold start system. In *TAC2012*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *EMNLP-CoNLL*, pages 455–465. ACL.

Suzanne Tamang, Zheng Chen, and Heng Ji. 2012. Entity linking system and slot filling validation system. In *TAC2012*.

Michael Wick, Sameer Singh, Ari Kobren, and Andrew McCallum. 2013. Assessing confidence of knowledge base content with an experimental study in entity resolution. In *AKBC2013*.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective cross-document relation extraction without labelled data. In *Proceedings of EMNLP*, EMNLP '10, pages 1013–1023, Stroudsburg, PA, USA. ACL.

Shubin Zhao and Ralph Grishman. 2005. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 419–426, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. 2008. Large-scale parallel collaborative filtering for the netflix prize. In *Proceedings of the 4th International Conference on Algorithmic Aspects in Information and Management*, AAIM '08, pages 337–348, Berlin, Heidelberg. Springer-Verlag.

Yiping Zhou, Lan Nie, Omid Rouhani-Kalleh, Flavian Vasile, and Scott Gaffney. 2010. Resolving surface forms to wikipedia topics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1335–1343, Stroudsburg, PA, USA. Association for Computational Linguistics.