An Entity-Topic Model for Entity Linking

Xianpei Han Le Sun Institute of Software, Chinese Academy of Sciences HaiDian District, Beijing, China. {xianpei, sunle}@nfs.iscas.ac.cn

Abstract

(EL)received Entity Linking has considerable attention in recent years. Given many name mentions in a document, the goal of EL is to predict their referent entities in a knowledge base. Traditionally, there have been two distinct directions of EL research: one focusing on the effects of mention's *context compatibility*, assuming that "the referent entity of a mention is reflected by its context"; the other dealing with the effects of document's topic coherence, assuming that "a mention's referent entity should be coherent with the document's main topics". In this paper, we propose a generative model - called entitytopic model, to effectively join the above two complementary directions together. By jointly modeling and exploiting the context compatibility, the topic coherence and the correlation between them, our model can accurately link all mentions in a document using both the local information (including the words and the mentions in a document) and the global knowledge (including the topic knowledge, the entity context knowledge and the entity name knowledge). Experimental results demonstrate the effectiveness of the proposed model.

1 Introduction

Entity Linking (EL) has received considerable research attention in recent years (McNamee & Dang, 2009; Ji et al., 2010). Given many name mentions in a document, the goal of EL is to predict their referent entities in a given knowledge base (KB), such as the *Wikipedia*¹. For example, as

shown in Figure 1, an EL system should identify the referent entities of the three mentions WWDC, Apple and Lion correspondingly are the entities Apple Worldwide Developers Conference, Apple Inc. and Mac OS X Lion in KB. The EL problem appears in many different guises throughout the areas of natural language processing, information retrieval and text mining. For instance, in many applications we need to collect all appearances of a specific entity in different documents, EL is an effective way to resolve such an information integration problem. Furthermore, EL can bridge the mentions in documents with the semantic information in knowledge bases (e.g., Wikipedia and *Freebase* 2), thus can provide a solid foundation for knowledge-rich methods.



Figure 1. A Demo of Entity Linking

Unfortunately, the accurate EL is often hindered by the *name ambiguity* problem, i.e., a name may refer to different entities in different contexts. For example, the name *Apple* may refer to more than 20 entities in Wikipedia, such as *Apple Inc.*, *Apple (band)* and *Apple Bank*. Traditionally, there have been two distinct directions in EL to resolve the *name ambiguity* problem: one focusing on the effects of mention's *context compatibility* and the other dealing with the effects of document's *topic coherence*. EL methods based on context

¹ www.wikipedia.org

² www.freebase.com

compatibility assume that "the referent entity of a mention is reflected by its context" (Mihalcea & Cosomai, 2007; Zhang et al., 2010; Zheng et al., 2010; Han & Sun, 2011; Kataria et al., 2011: Sen 2012). For example, the context compatibility based methods will identify the referent entity of the mention Lion in Figure 1 is the entity Mac OS XLion, since this entity is more compatible with its context words operating system and release than other candidates such as Lion(big cats) or Lion(band). EL methods based on topic coherence assume that "a mention's referent entity should be coherent with document's main topics" (Medelyan et al., 2008; Milne & Witten, 2008; Kulkarni et al., 2009; Han et al., 2011). For example, the topic coherence based methods will link the mention Apple in Figure 1 to the entity Apple Inc., since it is more coherent with the document's topic MAC OS X Lion Release than other referent candidates such as Apple (band) or Apple Bank.

In recent years, both of the above two EL directions have shown their effectiveness to some extent, and obviously they are *complementary* to each other. Therefore we believe that bring the above two directions together will enhance the EL performance. Traditionally, the above two directions are usually be brought together using a hybrid method (Zhang and Sim, 2011; Ratinov et al., 2011; Han et al., 2011), i.e., the context compatibility and the topic coherence are first separately modeled, then their EL evidence are combined through an additional model. For example, Zhang and Sim (2011) first models the context compatibility as a context similarity and the topic coherence as a similarity between the underlying topics of documents and KB entries, then these two similarities are combined through an additional SVM classifier for the final EL decision.

The main drawback of these hybrid methods, however, is that they model the context compatibility and the topic coherence *separately*, which makes it difficult to capture the *mutual reinforcement* effect between the above two directions. That is, the topic coherence and the context compatibility are highly correlated and their evidence can be used to reinforce each other in EL decisions. For example, in Figure 1, if the context compatibility gives a high likelihood the mention *Apple* refers to the entity *Apple Inc.*, then this likelihood will give more evidence for this document's topic is about MAC OS X Lion, and it in turn will reinforce the topic coherence between the entity MAC OS X Lion and the document. In reverse, once we known the topic of this document is about MAC OS X Lion, the context compatibility between the mention Apple and the entity Apple Inc. can be improved as the importance of the context words operating system and release will be increased using the topic knowledge. In this way, we believe that modeling the above two directions jointly, rather than separately, will further improve the EL performance by capturing the mutual reinforcement effect between the context compatibility and the topic coherence.

In this paper, we propose a method to *jointly* model and exploit the context compatibility, the topic coherence and the correlation between them for better EL performance. Specifically, we propose a generative probabilistic model – called *entity-topic model*, which can uniformly model the text compatibility and the topic coherence as the *statistical dependencies* between the *mentions*, the *words*, the underlying *entities* and the underlying *topics* of a document by assuming that each document is generated according to the following two assumptions:

1) **Topic coherence assumption**: All entities in a document should be centered around the main topics of the document. For example, the entity Apple Inc. tends to occur in documents about IT, but the entity Apple Bank will more likely to occur in documents about bank or investment.

2) **Context compatibility assumption**: *The context words of a mention should be centered on its referent entity.* For example, the words *computer, phone* and *music* tends to occur in the context of the entity *Apple Inc.*, meanwhile the words *loan, invest* and *deposit* will more likely to occur in the context of the entity *Apple Bank.*

In this way, the entity-topic model uniformly models the context compatibility, the topic coherence and the correlation between them as the dependencies between the observed information (the *mentions* and the *words*) in a document and the hidden information we want to know (the underlying *topics* and *entities*) through the global knowledge (including *the topic knowledge, the entity name knowledge* and *the entity context knowledge*). And the EL problem can now be decomposed into the following two inference tasks: 1) *Predicting the underlying topics and the underlying entities of a document* based on the observed information and the global knowledge. We call such a task the *prediction task*;

2) *Estimating the global knowledge from data*. Notice that the topic knowledge, the entity name knowledge and the entity context knowledge are all not previously given, thus we need to estimate them from data. We call such a task the *knowledge discovery task*.

Because the accurate inference of the above two tasks is intractable in our entity-topic model, this paper also develops an approximate inference algorithm – the Gibbs sampling algorithm to solve them.

Contributions. The main contributions of this paper are summarized below:

• We propose a generative probabilistic model, the *entity-topic model*, which can *jointly* model and exploit the context compatibility, the topic coherence and the correlation between them for better EL performance;

• We develop a *Gibbs sampling algorithm* to solve the two inference tasks of our model: 1) Discovering the global knowledge from data; and 2) Collectively making accurate EL decisions.

This paper is organized as follows. Section 2 describes the proposed entity-topic model. Section 3 demonstrates the Gibbs sampling algorithm. The experimental results are presented and discussed in Section 4. The related work is reviewed in Section 5. Finally we conclude this paper in Section 6.

2 The Entity-Topic Model for Entity Linking

In this section, we describe the proposed entitytopic model. In following we first demonstrate how to capture the context compatibility, the topic coherence and the correlation between them in the document generative process, then we incorporate the global knowledge generation into our model for knowledge estimation from data.

2.1 Document Generative Process

As shown in Section 1, we jointly model the context compatibility and the topic coherence as the statistical dependencies in the entity-topic model by assuming that all documents are generated in a topical coherent and context compatible way. In following we describe the document generative process.

In our model, each document d is assumed composed of two types of information, i.e., the *mentions* and the *words*. Formally, we represent a document as:

A **document** is a collection of M mentions and N words, denoted as $d = \{m_1, ..., m_M; w_1, ..., w_N\}$, with m_i the i^{th} mention and w_j the j^{th} word.

For example, the document in Figure 1 is represented as $d = \{WWDC, Apple, Lion; at, the, conference, ...\}$, where *WWDC*, *Apple*, *Lion* are the three mentions and the other are the words.

To generate a document, our model relies on three types of global knowledge, including:

• **Topic Knowledge** ϕ (*The entity distribution of topics*): In our model, all entities in a document are generated based on its underlying topics, with each topic is a group of semantically related entities. Statistically, we model each topic as a multinomial distribution of entities, with the probability indicating the likelihood an entity to be extracted from this topic. For example, we may have a topic $\phi_{Apple Inc.} = \{Steve Jobs^{0.12}, iPhone^{0.07}, iPod^{0.08}, ...\}$, indicating the likelihood of the entity *Steve Jobs* be extracted from this topic is 0.12, etc.

• Entity Name Knowledge ψ (*The name distribution of entities*): In our model, all name mentions are generated using the name knowledge of its referent entity. Specifically, we model the name knowledge of an entity as a multinomial distribution of its names, with the probability indicating the likelihood this entity is mentioned by the name. For example, the name knowledge of the entity *Apple Inc.* may be $\psi_{Apple Inc.} = \{Apple^{0.51}, Apple Computer Inc.^{0.10}, Apple Inc.^{0.07}, ...\}$, indicating that the entity *Apple Inc.* is mentioned by the name *Apple* with probability 0.51, etc.

• Entity Context Knowledge ξ (*The context* word distribution of entities): In our model, all context words of an entity's mention are generated using its context knowledge. Concretely, we model the context knowledge of an entity as a multinomial distribution of words, with the probability indicating the likelihood a word appearing in this entity's context. For example, we may have $\xi_{Apple Inc.} = \{phone^{0.07}, computer^{0.10}, IT^{0.06}, phone^{0.002}, ...\}$, indicating that the word computer appearing in the context of the entity Apple Inc. with probability 0.1, etc. Given the topic knowledge ϕ , the entity name knowledge ψ and the entity context knowledge ξ :

- 1. For each doc d in **D**, sample its topic distribution $\theta_d \sim Dir(\alpha)$;
- 2. For each of the M_d mentions m_i in doc d:
 - a) Sample a topic assignment $z_i \sim Mult(\theta_d)$;
 - b) Sample an entity assignment $e_i \sim Mult(\phi_{z_i})$;
 - c) Sample a mention $m_i \sim Mult(\psi_{e_i})$;
- 3. For each of the N_d words w_i in doc d:
 a) Sample a target entity it describes from d's referent entities a_i ~ Unif(e_{m1}, e_{m2}, ···, e_{md});
 b) Sample a describing word using a_i's context word distribution w_i ~ Mult(ξ_{ai}).

Figure 2. The document generative process, with *Dir(.)*, *Mult(.)* and *Unif(.)* correspondingly *Dirichlet*, *Multinomial* and *Uniform distribution*

Given the entity list $\mathbf{E} = \{e_1, e_2, ..., e_E\}$ in the knowledge base, the word list $\mathbf{V} = \{w_1, w_2, ..., w_v\}$, the entity name list $\mathbf{K} = \{n_1, n_2, ..., n_K\}$ and the global knowledge described in above, the generation process of a document collection (corpus) $\mathbf{D} = \{d_1, d_2, ..., d_D\}$ is shown in Figure 2. To demonstrate the generation process, we also demonstrate how the document in Figure 1 can be generated using our model in following steps:

Step 1: The model generates the topic distribution of the document as $\theta_d = \{Apple Inc.^{0.45}, Operating System(OS)^{0.55}\};$

Step 2: For the three mentions in the document:

i. According to the topic distribution θ_d , the model generates their topic assignments as z_1 =*Apple Inc.*, z_2 = *Apple Inc.*, z_3 = *OS*;

ii. According to the topic knowledge $\phi_{Apple Inc.}$, ϕ_{OS} and the topic assignments z_1 , z_2 , z_3 , the model generates their entity assignments as $e_1 = Apple$ Worldwide Developers Conference, $e_2 = Apple Inc.$, $e_3 = Mac OS X Lion$;

iii. According to the name knowledge of the entities *Apple Worldwide Developers Conference*, *Apple Inc.* and *Mac OS X Lion*, our model generates the three mentions as m_1 =WWDC, m_2 = *Apple*, m_3 = *Lion*;

Step 3: For all words in the document:

i. According to the referent entity set in document $\mathbf{e}_{d} = \{Apple \ Worldwide \ Developers \ Conference, \ Apple \ Inc., \ Mac \ OS \ X \ Lion\}, the model generates the target entity they describes as$

 a_3 =Apple Worldwide Developers Conference and a_4 =Apple Inc.;

ii. According to their target entity and the context knowledge of these entities, the model generates the context words in the document. For example, according to the context knowledge of the entities *Apple Worldwide Developers Conference*, the model generates its context word $w_3 = conference$, and according to the context knowledge of the entity *Apple Inc.*, the model generates its context word $w_4 = introduces$.

Through the above generative process, we can see that all entities in a document are extracted from the document's underlying topics, ensuring the topic coherence; and all words in a document are extracted from the context word distributions of its referent entities, resulting in the context compatibility. Furthermore, the generation of topics, entities, mentions and words are highly correlated, thus our model can capture the correlation between the topic coherence and the context compatibility.

2.2 Global Knowledge Generative Process

The entity-topic model relies on three types of global knowledge (including the topic knowledge, the entity name knowledge and the entity context knowledge) to generate a document. Unfortunately, all three types of global knowledge are unknown and thus need to be estimated from data. In this paper we estimate the global knowledge through Bayesian inference by also incorporating the knowledge generation process into our model.

Specifically, given the topic number T, the entity number E, the name number K and the word number V, the entity-topic model generates the global knowledge as follows:

1) $\phi|\beta \sim Dir(\beta)$

For each topic z, our model samples its entity distribution ϕ_z from an *E*-dimensional Dirichlet distribution with hyperparameter β .

2)
$$\psi | \gamma \sim Dir(\gamma)$$

For each entity e, our model samples its name distribution ψ_e from a *K*-dimensional Dirichlet distribution with hyperparameter γ .

3) $\xi | \delta \sim Dir(\delta)$

For each entity e, our model samples its context word distribution ξ_e from a V-dimensional Dirichlet distribution with hyperparameter δ .

Finally, the full entity-topic model is shown in Figure 3 using the plate representation.



Figure 3. The plate representation of the entitytopic model

2.3 The Probability of a Corpus

Using the entity-topic model, the probability of generating a corpus $\mathbf{D} = \{d_1, d_2, ..., d_D\}$ given hyperparameters α , β , γ and δ can be expressed as:

$$P(\mathbf{D}; \alpha, \beta, \gamma, \delta) = \prod_{d} P(\mathbf{m}_{d}, \mathbf{w}_{d}; \alpha, \beta, \gamma, \delta)$$

$$= \prod_{d} \sum_{\mathbf{e}_{d}} P(\mathbf{e}_{d} | \alpha, \beta) P(\mathbf{m}_{d} | \mathbf{e}_{d}, \gamma) P(\mathbf{w}_{d} | \mathbf{e}_{d}, \delta)$$

$$= \int_{\phi} P(\phi | \beta) \int_{\psi} P(\psi | \gamma) \prod_{d} \sum_{\mathbf{e}_{d}} P(\mathbf{m}_{d} | \mathbf{e}_{d}, \psi)$$

$$\times \int_{\xi} P(\xi | \delta) \sum_{\mathbf{a}_{d}} P(\mathbf{a}_{d} | \mathbf{e}_{d}) P(\mathbf{w}_{d} | \mathbf{a}_{d}, \xi)$$

$$\times \int_{\theta} P(\theta | \alpha) P(\mathbf{e}_{d} | \theta, \phi) d\theta d\xi d\psi d\phi \qquad (2.1)$$

where $\mathbf{m}_{\mathbf{d}}$ and $\mathbf{e}_{\mathbf{d}}$ correspondingly the set of mentions and their entity assignments in document d, $\mathbf{w}_{\mathbf{d}}$ and $\mathbf{a}_{\mathbf{d}}$ correspondingly the set of words and their entity assignments in document d.

3 Inference using Gibbs Sampling

In this section, we describe how to resolve the entity linking problem using the entity-topic model. Overall, there were two inference tasks for EL: 1) The prediction task. Given a document d, predicting its *entity assignments* ($\mathbf{e}_{\mathbf{d}}$ for mentions and $\mathbf{a}_{\mathbf{d}}$ for words) and *topic assignments* ($\mathbf{z}_{\mathbf{d}}$). Notice that here the EL decisions are just the prediction of per-mention entity assignments ($\mathbf{e}_{\mathbf{d}}$).

2) The knowledge discovery task. Given a corpus $\mathbf{D}=\{d_1, d_2, ..., d_D\}$, estimating the global knowledge (including the entity distribution of topics ϕ , the name distribution ψ and the context word distribution ξ of entities) from data.

Unfortunately, due to the heaven correlation between topics, entities, mentions and words (the correlation is also demonstrated in Eq. (2.1), where the integral is intractable due to the coupling between θ , ϕ , ψ and ξ), the accurate inference of the above two tasks is intractable. For this reason, we propose an approximate inference algorithm – the Gibbs sampling algorithm for the entity-topic model by extending the well-known Gibbs sampling algorithm for LDA (Griffiths & Steyvers, 2004). In Gibbs sampling, we first construct the posterior distribution $P(\mathbf{z}, \mathbf{e}, \mathbf{a} | \mathbf{D})$, then this posterior distribution is used to: 1) estimate θ , ϕ , ψ and ξ ; and 2) predict the entities and the topics of all documents in D. Specifically, we first derive the joint posterior distribution from Eq. (2.1) as:

$$P(\mathbf{z}, \mathbf{e}, \mathbf{a} | \mathbf{D}) \propto P(\mathbf{z}) P(\mathbf{e} | \mathbf{z}) P(\mathbf{m} | \mathbf{e}) P(\mathbf{a} | \mathbf{e}) P(\mathbf{w} | \mathbf{a})$$

where

$$P(\mathbf{z}) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^D \frac{\prod_t \Gamma(\alpha + C_{dt}^{DT})}{\Gamma(T\alpha + C_{d*}^{DT})} \qquad (3.1)$$

is the probability of the joint topic assignment \mathbf{z} to all mentions \mathbf{m} in corpus D, and

$$P(\mathbf{e}|\mathbf{z}) = \left(\frac{\Gamma(E\beta)}{\Gamma(\beta)^E}\right)^T \prod_{t=1}^T \frac{\prod_e \Gamma(\beta + C_{te}^{TE})}{\Gamma(E\beta + C_{t*}^{TE})} \quad (3.2)$$

is the conditional probability of the joint entity assignments \mathbf{e} to all mentions \mathbf{m} in corpus D given all topic assignments \mathbf{z} , and

$$P(\mathbf{m}|\mathbf{e}) = \left(\frac{\Gamma(K\gamma)}{\Gamma(\gamma)^{K}}\right)^{E} \prod_{e=1}^{E} \frac{\prod_{m} \Gamma(\gamma + C_{em}^{EM})}{\Gamma(K\gamma + C_{e*}^{EM})} \quad (3.3)$$

is the conditional probability of all mentions **m** given all per-mention entity assignments **e**, and

$$P(\mathbf{a}|\mathbf{e}) = \prod_{d=1}^{D} \prod_{e \in \mathbf{e_d}} \left(\frac{C_{de}^{DE}}{C_{d*}^{DE}}\right)^{C_{de}^{DA}}$$
(3.4)

is the conditional probability of the joint entity assignments \mathbf{a} to all words \mathbf{w} in corpus D given all per-mention entity assignments \mathbf{e} , and

$$P(\mathbf{w}|\mathbf{a}) = \left(\frac{\Gamma(V\delta)}{\Gamma(\delta)^{V}}\right)^{E} \prod_{e=1}^{E} \frac{\prod_{w} \Gamma(\delta + C_{ew}^{EW})}{\Gamma(V\delta + C_{e*}^{EW})} \quad (3.5)$$

is the conditional probability of all words **w** given all per-word entity assignments **a**. In all above formulas, $\Gamma(.)$ is the Gamma function, C_{dt}^{DT} is the times topic *t* has been assigned for all mentions in document *d*, $C_{d*}^{DT} = \sum_{t} C_{dt}^{DT}$ is the topic number in document *d*, and C_{te}^{TE} , C_{em}^{EM} , C_{de}^{DE} , C_{de}^{DA} , C_{ew}^{EW} have similar explanation.

Based on the above joint probability, we construct a Markov chain that converges to the posterior distribution $P(\mathbf{z}, \mathbf{e}, \mathbf{a} | \mathbf{D})$ and then draw samples from this Markov chain for inference. For entity-topic model, each state in the Markov chain is an assignment (including *topic assignment to a mention, entity assignment to a mention* and *entity assignment to a word*). In Gibbs sampling, all assignments are sequentially sampled conditioned on all the current other assignments. So here we only need to derive the following three fully conditional assignment distributions:

- 1) $P(z_i = t | \mathbf{z}_{-i}, \mathbf{e}, \mathbf{a}, \mathbf{D})$: the topic assignment distribution to a mention given the current other topic assignments \mathbf{z}_{-i} , the current entity assignments \mathbf{e} and \mathbf{a} ;
- P(e_i = e|z, e_{-i}, a, D) : the entity assignment distribution to a mention given the current entity assignments of all other mentions e_{-i}, the current topic assignments z and the current entity assignments of context words a;
- 3) $P(a_i = e | \mathbf{z}, \mathbf{e}, \mathbf{a}_{-i}, \mathbf{D})$: the entity assignment distribution to a context word given the current entity assignments of all other context words \mathbf{a}_{-i} , the current topic assignments \mathbf{z} and the current entity assignments \mathbf{e} of mentions.

Using the Formula 3.1-3.5, we can derive the above three conditional distributions as (where m_i is contained in doc *d*):

$$P(z_i = t | \mathbf{z}_{-i}, \mathbf{e}, \mathbf{a}, \mathbf{D}) \propto \frac{C_{(-i)dt}^{DT} + \alpha}{C_{(-i)d*}^{DT} + T\alpha} \times \frac{C_{(-i)te}^{TE} + \beta}{C_{(-i)t*}^{TE} + E\beta}$$

where the topic assignment to a mention is determined by the probability this topic appearing in doc d (the 1st term) and the probability the referent entity appearing in this topic (the 2nd term);

$$\begin{split} P(e_i &= e | \mathbf{z}, \mathbf{e}_{-i}, \mathbf{a}, \mathbf{D}) \quad \propto \\ \frac{C_{(-i)te}^{TE} + \beta}{C_{(-i)t*}^{TE} + E\beta} \times \frac{C_{(-i)em}^{EM} + \gamma}{C_{(-i)e*}^{EM} + K\gamma} \times \big(\frac{C_{(-i)de}^{DE} + 1}{C_{(-i)de}^{DE}}\big)^{C_{de}^{DA}} \end{split}$$

where the entity assignment to a mention is determined by the probability this entity extracted from the assigned topic (the 1^{st} term), the probability this entity is referred by the name *m* (the 2^{nd} term) and the contextual words describing this entity in doc *d* (the 3^{rd} term);

$$P(a_i = e | \mathbf{z}, \mathbf{e}, \mathbf{a}_{-i}, \mathbf{D}) \propto \frac{C_{de}^{DE}}{C_{d*}^{DE}} \times \frac{C_{(-i)ew}^{EW} + \delta}{C_{(-i)e*}^{EW} + V\delta}$$

where the entity assignment to a word is determined by the number of times this entity has been assigned to mentions in doc d (the 1st term) and the probability the word appearing in the context of this entity (the 2nd term).

Finally, using the above three conditional distributions, we iteratively update all assignments of corpus \mathbf{D} until coverage, then the global knowledge is estimated using the final assignments, and the final entity assignments are used as the referents of their corresponding mentions.

Inference on Unseen Documents. When unseen documents are given, we predict its entities and topics using the incremental Gibbs sampling algorithm described in (Kataria et al., 2011), i.e., we iteratively update the entity assignments and the topic assignments of an unseen document as the same as the above inference process, but with the previously learned global knowledge fixed.

Hyperparameter setting. One still problem here is the setting of the hyperparameters α, β, γ and δ . For α and β , this paper empirically set the value of them to $\alpha = 50/T$ and $\beta = 0.1$ as in Griffiths & Steyvers(2004). For γ , we notice that $K\gamma$ is the number of pseudo names added to each entity, when $\gamma = 0$ our model only mentions an entity using its previously used names. Observed that an entity typically has a fixed set of names, we set γ to a small value by setting $K\gamma = 1.0$. For δ , we notice that $V\delta$ is the number of pseudo words added to each entity, playing the role of smoothing its context word distribution. As there is typically a relatively loose correlation between an entity and its context words, we set δ to a relatively large value by fixing the total smoothing words added to each entity, a typical value is $V\delta = 2000$.

4 Experiments

In this section, we evaluate our method and compare it with the traditional EL methods. We first explain the experimental settings in Section 4.1-4.4, then discuss the results in Section 4.5.

4.1 Knowledge Base

In our experiments, we use the Jan. 30, 2010 English version of Wikipedia as the knowledge base, which contains over 3 million entities. Notice that we also take the general concepts in Wikipedia (such as *Apple, Video, Computer*, etc.) as entities, so the entity in this paper may not strictly follow its definition.

4.2 Data Sets

There are two standard data sets for EL: $IITB^3$ and TAC 2009 EL data set (McNamee & Dang, 2009), where IITB focuses on *aggressive recall* EL and TAC 2009 focuses on EL on *salient mentions*. Due to the collective nature of our method, we mainly used the **IITB** as the primary data set as the same as Kulkarni et al.(2009) and Han et al.(2011). But we also give the EL accuracies on the TAC 2009 in Sect. 4.5.4 as auxiliary results.

Overall, the IITB data set contains 107 web documents. For each document, the name mentions' referent entities in Wikipedia are manually annotated to be as exhaustive as possible. In total, 17,200 name mentions are annotated, with 161 name mentions per document on average. In our experiments, we use only the name mentions whose referent entities are contained in Wikipedia.

4.3 Evaluation Criteria

This paper adopted the same performance metrics used in the Kulkarni et al. (2009), which includes **Recall**, **Precision** and **F1**. Let M^* be the golden standard set of the EL results (each EL result is a pair (m, e), with m the mention and e its referent entity), M be the set of EL results outputted by an EL system, then these metrics are computed as:

$$Precision = \frac{|M \cap M^*|}{|M|}$$
$$Recall = \frac{|M \cap M^*|}{|M^*|}$$

where two EL results are considered equal if and only if both their mentions and referent entities are equal. As the same as Kulkarni et al.(2009), *Precision* and *Recall* are averaged across documents and overall *F1* is used as the primary performance metric by computing from average *Precision* and *Recall*.

4.4 Baselines

We compare our method with five baselines which are described as follows:

Wikify!. This is a context compatibility based EL method using vector space model (Mihalcea & Csomai, 2007). *Wikify!* computes the context compatibility using the word overlap between the mention's context and the entity's Wikipedia entry.

EM-Model. This is a statistical context compatibility based EL method described in Han & Sun(2011), which computes the compatibility by integrating the evidence from the entity popularity, the entity name knowledge and the context word distribution of entities.

M&W. This is a relational topic coherence based EL method described in Milne & Witten(2008). M&W measures an entity's topic coherence to a document as its average semantic relatedness to the *unambiguous* entities in the document.

CSAW. This is an EL method which combines context compatibility and topic coherence using a hybrid method (Kulkarni et al., 2009), where context compatibility and topic coherence are first separated modeled as context similarity and the sum of all *pair-wise* semantic relatedness between the entities in the document, then the entities which can maximize the weighted sum of the context compatibility and the topic coherence are identified as the referent entities of the document.

EL-Graph. This is a graph based hybrid EL method described in Han et al. (2011), which first models the context compatibility as text similarity and the topic coherence of an entity as its node importance in a referent graph which captures all mention-entity and entity-entity relations in a document, then a random walk algorithm is used to collectively find all referent entities of a document.

Except for *CSAW* and *EL-Graph*, all other baselines are designed only to link the salient name mentions (i.e., key phrases) in a document. In our experiment, in order to compare the EL performances on also the non-salient name mentions, we push these systems' recall by reducing their respective importance thresholds of linked mentions.

³ http://www.cse.iitb.ac.in/~soumen/doc/QCQ/

4.5 Experimental Results

4.5.1 Overall Performance

We compared our method with all the above five baselines. For our method, we estimate the global knowledge using all the articles in the Jan. 30, 2010 English version of Wikipedia, and totally there were 3,083,158 articles. For each article, the mentions within it are detected using the methods described in Medelyan et al.(2008) and all terms in an article are used as context words, so a term may both be a mention and a context word. The topic number of our model is T = 300 (will be empirically set in Sect 4.5.2). To train the entitytopic model, we run 500 iterations of our Gibbs sampling algorithm to converge. The training time of our model is nearly one week on our server using 20 GB RAM and one core of 3.2 GHz CPU. Since the training can be done offline, we believe that the training time is not critical to the realworld usage as the online inference on new document is very quick. Using the above settings, the overall results are shown in Table 1.

	Precision	Recall	F1
Wikify!	0.55	0.28	0.37
EM-Model	0.82	0.48	0.61
M&W	0.80	0.38	0.52
CSAW	0.65	0.73	0.69
EL-Graph	0.69	0.76	0.73
Our Method	0.81	0.80	0.80

Table 1. The overall results on IITB data set

From the overall results in Table 1, we can see that: 1) By *jointly* modeling and exploiting the context compatibility and the topic coherence, our method can achieve competitive performance: ① compared with the context compatibility baselines Wikify! and EM-Model. our method correspondingly gets 43% and 19% F1 improvement; 2 compared with the topic coherence baselines M&W, our method achieves 28% F1 improvement; 3 compared with the hybrid baselines CSAW and EL-Graph, our method correspondingly achieves 11% and 7% F1 improvement.

2) Compared with the *context compatibility only* and the *topic coherence only* methods, the main advantage of our method is that, rather than only achieved high entity linking precision on *salient* mentions, it can also effectively link the

non-salient mentions in a document: this is demonstrated in our method's significant *Recall* improvement: a 32~52% Recall improvement over baselines *Wikify!*, *EM-Model* and *M&W*. We believe this is because a document usually contains little evidence for EL decisions on non-salient mentions, so with either only context compatibility or only topic coherence the evidence is not enough for EL decisions on these non-salient mentions, and bring these two directions together is critical for the accurate EL on these mentions.

3) Compared with the *hybrid* methods, the main advantage of our method is the improvement of EL precision (a 11~16% improvement over baselines *CSAW* and *EL-Graph*), we believe this is because: ① Our method can further capture the mutual reinforcement effect between the context compatibility and the topic coherence; ② The traditional hybrid methods usually determine the topic coherence of an entity to a document using *all entities* in the document, in comparison our method uses only *the entities in the same topic*, we believe this is more reasonable for EL decisions.

4.5.2 Parameter Tuning

One still parameter of our method is the topic number *T*. An appropriate *T* will distribute entities into well-organized topics, in turn it will capture the co-occurrence information of entities. Figure 4 plots the *F1* at different *T* values. We can see that the *F1* is not very sensitive to the topic number and with T = 300 our method achieves its best *F1* performance.



Figure 4. The FI vs. the topic number T

4.5.3 Detailed Analysis

In this section we analyze why and how our method works well in detail. Generally, we believe the main advantages of our method are:

1) **The effects of topic knowledge**. One main advantage of our model is that the topic knowledge

can provide a *document-specific entity prior* for EL. Concretely, using the topic knowledge and the topic distribution of documents, the prior for an entity appearing in a document d is highly related to the document's topics:

$$P(e|d) = \sum_{z} P(z|d)P(e|z)$$

This prior is obviously more reasonable than the "information less prior" (i.e., all entities have equal prior) or "a global entity popularity prior" (Han & Sun, 2011). To demonstrate, Table 2-3 show the 3 topics where the *Apple Inc.* and the fruit *Apple* have the largest generation probability P(e|z) from these topics. We can see that the topic knowledge can provide a reasonable prior for entities appearing in a document: the *Apple Inc.* has a large prior in documents about *Computer, Video* and *Software*, and the fruit *Apple* has a large prior in documents about *Wine, Food* and *Plant*.

Topic(Computer)	Topic(Video)	Topic(Software)
Computer	Video	Computer software
CPU	Mobile phone	Microsoft Windows
Hardware	Mass media	Linux
Personal computer	Music	Web browser
Computer memory	Television	Operating system

Table 2. The 3 topics where the *Apple Inc*. has the largest P(e|z)

Topic(Wine)	Topic(Food)	Topic(Plant)
Wine	Food	Plant
Grape	Restaurant	Flower
Vineyard	Meat	Leaf
Winery	Cheese	Tree
Apple	Vegetable	Fruit

Table 3. The 3 topics where the fruit *Apple* has the largest P(e|z)

2) The effects of a fine-tuned context model. The second advantage of our model is that it provides a statistical framework for fine-tuning the context model from data. To demonstrate such an effect, Table 4 compares the EL performance of 1) the entity-topic model with no context model is used (No Context), i.e., we determine the referent entity of a mention by deleting the 3rd term of the formula $P(e_i = e | \mathbf{z}, \mathbf{e}_{-i}, \mathbf{a}, \mathbf{D})$ in Section 3; (2) with the context model estimated using the entity's Wikipedia page (Article Content), ③ with the context model estimated using the 50 word window of all its mentions in Wikipedia (Mention *Context*) and; ④ with the context model in the original entity-topic model (Entity-Topic Model). From Table 4 we can see that a fine-tuned context model will result in a 2~7% F1 improvement.

Context Model	F1
No Context	0.73
Article Content	0.75
Mention Context	0.78
Entity-Topic Model	0.80

Table 4. The *F1* using different context models

3) **The effects of joint model**. The third advantage of our model is that it *jointly* model the context compatibility and the topic coherence, which bring two benefits: ① the mutual reinforcement between the two directions can be captured in our model; ② the context compatibility and the topic coherence are uniformly modeled and jointly estimated, which makes the model more accurate for EL.

4.5.4 EL Accuracies on TAC 2009 dataset

We also compare our method with the top 5 EL systems in TAC 2009 and the two state-of-the-art systems (EM-Model and EL-Graph) on TAC 2009 data set in Figure 5 (For EL-Graph and our method, a NIL threshold is used to detect whether the referent entity is contained in the knowledge base, if the knowledge base not contains the referent entity, we assign the mention to a NIL entity). From Figure 5, we can see that our method is competitive: 1) Our method can achieve a 3.4% accuracy improvement over the best system in TAC 2009; 2) Our method, EM-Model and EL-Graph get very close accuracies (0.854, 0.86 and 0.838 correspondingly), we believe this is because: ① The mentions to be linked in TAC data set are mostly salient mentions; 2 The influence of the NIL referent entity problem, i.e., the referent entity is not contained in the given knowledge base: Most referent entities (67.5%) on TAC 2009 are NIL entity and our method has no special handling on this problem, rather than other methods such as the EM-Model, which affects the overall performance of our method.



Figure 5. The EL accuracies on TAC 2009 dataset

5 Related Work

In this section, we briefly review the related work of EL. Traditionally, the context compatibility based methods link a mention to the entity which has the largest compatibility with it. Cucerzan (2007) modeled the compatibility as the cosine similarity between the vector space representation of mention's context and of entity's Wikipedia entry. Mihalcea & Csomai (2007), Bunescu & Pasca (2006), Fader et al. (2009), Gottipati et al.(2011) and Zhang et al.(2011) extended the vector space model with more information such as the entity category and the acronym expansion, etc. Han & Sun (2011) proposed a generative model which computes the compatibility using the evidences from entity's popularity, name distribution and context word distribution. Kataria et al.(2011) and Sen (2012) used a latent topic model to learn the context model of entities. Zheng et al. (2010), Dredze et al. (2010), Zhang et al. (2010), Zhou et al. (2010) and Ji & Chen(2011) employed the ranking techniques to further take relations between candidate entities into account.

On the other side, the *topic coherence* based methods link a mention to the entity which are most coherent to the document containing it. Medelyan et al. (2008) measured the topic coherence of an entity to a document as the weighted average of its relatedness to the unambiguous entities in the document. Milne and Witten (2008) extended Medelyan et al. (2008)'s coherence by incorporating commonness and context quality. Bhattacharya and Getoor (2006) modeled the topic coherence as the likelihood an entity is generated from the latent topics of a document. Sen (2012) modeled the topic coherence as the groups of co-occurring entities. Kulkarni et al. (2009) modeled the topic coherence as the sum of all pair-wise relatedness between the referent entities of a document. Han et al.(2011) and Hoffart et al.(2011) modeled the topic coherence of an entity as its node importance in a graph which captures all mention-entity and entity-entity relations in a document.

6 Conclusions and Future Work

This paper proposes a generative model, the entitytopic model, for entity linking. By uniformly modeling context compatibility, topic coherence and the correlation between them as statistical dependencies, our model provides an effective way to jointly exploit them for better EL performance.

In this paper, the entity-topic model can only link mentions to the previously given entities in a knowledge base. For future work, we want to overcome this limit by incorporating an *entity discovery* ability into our model, so that it can also discover and learn the knowledge of previously unseen entities from a corpus for linking name mentions to these entities.

Acknowledgments

The work is supported by the National Natural Science Foundation of China under Grants no. 90920010 and 61100152. Moreover, we sincerely thank the reviewers for their valuable comments.

References

- Adafre, S. F. & de Rijke, M. 2005. *Discovering missing links in Wikipedia*. In: Proceedings of the 3rd international workshop on Link discovery.
- Bhattacharya, I. and L. Getoor. 2006. *A latent dirichlet model for unsupervised entity resolution*. In: Proceedings of SIAM International Conference on Data Mining.
- Blei, D. M. and A. Y. Ng, et al. (2003). *Latent dirichlet allocation*. In: The Journal of Machine Learning Research 3: 993--1022.
- Bunescu, R. & Pasca, M. 2006. Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of EACL, vol. 6.
- Brown, P., Pietra, S. D., Pietra, V. D., and Mercer, R. 1993. *The mathematics of statistical machine translation: parameter estimation*. Computational Linguistics, 19(2), 263-31.
- Chen, S. F. & Goodman, J. 1999. An empirical study of smoothing techniques for language modeling. In Computer Speech and Language, London; Orlando: Academic Press, c1986-, pp. 359-394.
- Cucerzan, S. 2007. *Large-scale named entity disambiguation based on Wikipedia data*. In: Proceedings of EMNLP-CoNLL, pp. 708-716.
- De Beaugrande, R. A. and W. U. Dressler. 1981. Introduction to text linguistics, Chapter V, Longman London.
- Dredze, M., McNamee, P., Rao, D., Gerber, A. & Finin, T. 2010. Entity Disambiguation for Knowledge Base Population. In: Proceedings of the 23rd International Conference on Computational Linguistics.

- Fader, A., Soderland, S., Etzioni, O. & Center, T. 2009. Scaling Wikipedia-based named entity disambiguation to arbitrary web text. In: Proceedings of Wiki-AI Workshop at IJCAI, vol. 9.
- Gottipati, S., Jiang, J. 2011. *Linking Entities to a Knowledge Base with Query Expansion*. In: Proceedings of EMNLP.
- Griffiths, T. L. and M. Steyvers. 2004. *Finding scientific topics*. In: Proceedings of the National Academy of Sciences of the United States of America.
- Han, X., Sun, L. and Zhao J. 2011. *Collective Entity Linking in Web Text: A Graph-Based Method*. In: Proceedings of 34th Annual ACM SIGIR Conference.
- Han, X. and Sun, L. 2011. A Generative Entity-Mention Model for Linking Entities with Knowledge Base. In: Proceedings of ACL-HLT.
- Hoffart, J., Yosef, M. A., et al. 2011. *Robust Disambiguation of Named Entities in Text*. In: Proceedings of EMNLP.
- Jelinek, Frederick and Robert L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In: Proceedings of the Workshop on Pattern Recognition in Practice.
- Kataria, S. S., Kumar, K. S. and Rastogi, R. 2011. *Entity Disambiguation with Hierarchical Topic Models*. In: Proceedings of KDD.
- Kulkarni, S., Singh, A., Ramakrishnan, G. & Chakrabarti, S. 2009. Collective annotation of Wikipedia entities in web text. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 457-466.
- Li, X., Morie, P. & Roth, D. 2004. *Identification and tracing of ambiguous names: Discriminative and generative approaches*. In: *Proceedings of the National Conference on Artificial Intelligence*, pp. 419-424.
- McNamee, P. & Dang, H. T. 2009. Overview of the *TAC 2009 Knowledge Base Population Track*. In: Proceeding of Text Analysis Conference.
- Ji, H., et al. 2010. *Overview of the TAC 2010 knowledge base population track*. In: Proceedings of Text Analysis Conference.
- Ji, H. and Chen, Z. 2011. *Collaborative Ranking: A Case Study on Entity Linking*. In: Proceedings of EMNLP.
- Milne, D. & Witten, I. H. 2008. Learning to link with Wikipedia. In: Proceedings of the 17th ACM

conference on Conference on information and knowledge management.

- Milne, D., et al. 2006. *Mining Domain-Specific Thesauri from Wikipedia: A case study*. In Proc. of IEEE/WIC/ACM WI.
- Medelyan, O., Witten, I. H. & Milne, D. 2008. *Topic indexing with Wikipedia*. In: Proceedings of the AAAI WikiAI workshop.
- Mihalcea, R. & Csomai, A. 2007. *Wikify!: linking documents to encyclopedic knowledge*. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 233-242.
- Pedersen, T., Purandare, A. & Kulkarni, A. 2005. Name discrimination by clustering similar contexts. Computational Linguistics and Intelligent Text Processing, pp. 226-237.
- Ratinov, L. and D. Roth, et al. 2011. *Local and Global Algorithms for Disambiguation to Wikipedia*. In: Proceedings of ACL.
- Sen, P. 2012. Collective context-aware topic models for entity disambiguation. In Proceedings of WWW '12, New York, NY, USA, ACM.
- Zhang, W., Su, J., Tan, Chew Lim & Wang, W. T. 2010. Entity Linking Leveraging Automatically Generated Annotation. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).
- Zheng, Z., Li, F., Huang, M. & Zhu, X. 2010. *Learning* to *Link Entities with Knowledge Base*. In: The Proceedings of the Annual Conference of the North American Chapter of the ACL.
- Zhou, Y., Nie, L., Rouhani-Kalleh, O., Vasile, F. & Gaffney, S. 2010. *Resolving Surface Forms to Wikipedia Topics*. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 1335-1343.
- Zhang, W. and Sim, Y. C., et al. 2011. *Entity Linking* with Effective Acronym Expansion, Instance Selection and Topic Modeling. In: Proceedings of IJCAI.