Computational Linguistics and Chinese Language Processing
Vol. 19, No. 1, March 2014, pp. 11-32

Joint Learning of Entity Linking Constraints Using a Markov-Logic Network

Hong-Jie Dai[*], Richard Tzong-Han Tsai[+], and Wen-Lian Hsu[#]

Abstract

Entity linking (EL) is the task of linking a textual named entity mention to a knowledge base entry. Traditional approaches have addressed the problem by dividing the task into separate stages: entity recognition/classification, entity filtering, and entity mapping, in which different constraints are used to improve the system's performance. Nevertheless, these constraints are executed separately and cannot be used interactively. In this paper, we propose an integrated solution to the task based on a Markov logic network (MLN). We show how the stage decision can be formulated and combined in an MLN. We conducted experiments on the biomedical EL task, gene mention linking (GML), and compared our model's performance with those of two other GML approaches. Our experimental results provide the first comprehensive GML evaluations from three different perspectives: article-wide precision/recall/F-measure (PRF), instance-based PRF, and question answering accuracy. This paper also provides formal definitions of all of the above EL tasks. Experimental results show that our method outperforms the baseline and state-of-the-art systems under all three evaluation schemes.

Keywords: Entity Linking, Entity Disambiguation, Markov Logic Network, Gene Normalization

1. Introduction

Developing a system that can identify entities, such as personal names and gene or disease mentions, and that can classify the relations between them is useful for several applications in natural language processing and knowledge acquisition. There are several possible uses for

[*] Graduate Institute of BioMedical Informatics, Taipei Medical University
E-mail: hjdai@tmu.edu.tw
[+] Dep. of Computer Science & Information Engineering, National Central University
E-mail: thtsai@csie.ncu.edu.tw
[#] Institute of Information Science, Academia Sinica
E-mail: hsu@iis.sinica.edu.tw

such a system in different fields, *e.g*., improving document retrieval for specific entities, relation extraction, and attribute assignment (*e.g*., gene ontology annotations). In these applications, recognized entities must be linked to unique database entries. McNamee and Dang (2009b) named the task of matching a textual entity mention to a knowledge base (KB) entry *Entity Linking* (EL). In Figure 1, we provide a biomedical abstract to illustrate this task. The abstract discusses the relationship of the gene "CD59" to other lymphocyte antigens.

TITLE: Structure of the **CD59**-encoding gene: further evidence of a relationship to *murine* lymphocyte antigen Ly-6 protein

ABSTRACT: The gene for **CD59** [**membrane inhibitor of reactive lysis** (**MIRL**), **protectin**], a phosphatidylinositol-linked surface glycoprotein that regulates the formation of the polymeric **C9 complex** of complement and that is deficient on the abnormal hematopoietic cells of *patients* with paroxysmal nocturnal hemoglobinuria, consists of four exons spanning 20 kilobases. … PMID [1381503]

Figure 1. An example of entity linking.

After EL, the gene mention "CD59" in the first sentence must be linked to ID 966 in the Entrez Gene database of PubMed. In the first sentence, the authors also listed other designations of the gene, including "membrane inhibitor of reactive lysis" and "protectin," and they defined "MIRL" as the abbreviation for "membrane inhibitor of reactive lysis." Linking these instances to the same entry is a problem related to the *name variations* issue. Furthermore, the gene "CD59" may exist in multiple species. For example, it appeared in the title of the abstract as a murine gene, but turns out to be referring to a human (patient) gene in the first sentence. Therefore, each gene must be linked to its own unique database entry. Since these instances are polysemous, they are considered *entity ambiguity* issues. Finally, the "C9 complex" in the first sentence is a protein complex, but the Entrez Gene database does not contain this type of entity. When an entity cannot be associated with any entries, it is called an *absence* issue (McNamee & Dang, 2009b), and those entities are referred to as "*Nils*".

Of all of the aforementioned issues, entity ambiguity is the most crucial problem (Dredze *et al*., 2010). Take the name "TP53" as an example. In the Entrez Gene database, there are over 300 proteins within over 20 species possessing the same name. Several disambiguation approaches have been proposed to address the problem. For example, Dredze *et al*. (2010) formulated the disambiguation task as a ranking problem and developed features to link entities to Wikipedia entries. Zhang *et al.* (2010) used an automatically generated corpus to train a binary classifier to reduce ambiguities. Dai *et al*. (2010) collected external knowledge for each entity and calculated likelihoods stating the similarity of the current text with the knowledge to improve the disambiguation performance.

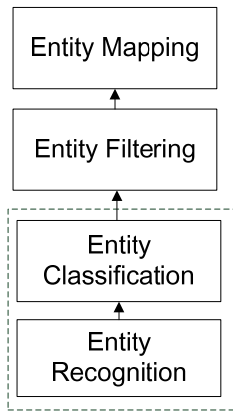


Figure 2. Stages in the bottom-up EL approach: Some works combine the entity recognition and the entity classification into one step.

Usually, a real-world EL system is constructed in a bottom-up manner, so it is necessary to make several decisions in different stages during the EL process. Figure 2 depicts the bottom-up process (Krauthammer *et al.*, 2004). *Entity recognition* marks single words (or several adjacent words) that indicate the presence of entities. As entity recognition does not determine the specific meaning of a concept, it is often combined with *Entity Classification*, which assigns entities to different classes, such as persons, genes, or diseases. After removing Nils (*Entity Filtering*), *Entity Mapping* maps entities to controlled database entries by calculating the similarities between the recognized entities and lexicon resources. This stage may resolve the entity ambiguity issue by a disambiguation process that uses contextual information to link entities to KB entries.

As shown in Figure 2, the traditional method for dealing with Nils has been to employ an additional step to filter out entities that have no corresponding entry in a KB. For example, Bunescu *et al.* (2006) filtered out mentions whose confidence scores are less than a fixed threshold. J Hakenberg *et al.* (2008) and Li *et al.* (2009) trained separate binary classifiers to validate linked mentions. Dredze *et al.* (2010) treated Nils as another KB entry candidate to train their EL ranking model.

Unfortunately, the separate-stage approach ignores possible dependencies among these stages and can result in error propagation. Continuing our example in Figure 1, in the EL stage, "MIRL" can be unambiguously linked to ID 996 with high confidence, because a search for the name in Entrez Gene returns only one match. Nevertheless, linking other mentions (*e.g.* "CD59" and "protectin") to ID 996 is not as easy, since "CD59" alone has 18 candidate entries. These names can be linked with more ease when considered as synonyms of MIRL. Nevertheless, a divergent filtering stage may filter out the entity mention "MIRL" because it is listed as an abbreviation of organization names, such as Mineral Industry Research Laboratory.

With a joint inference process, we can carry out both tasks simultaneously to avoid this type of error propagation (Poon *et al*., 2007).

Joint inference has become popular recently, because it allows features and constraints to be shared among different tasks. For example, J. R. Finkel *et al*. (2009) integrated parsing and named entity recognition into a joint model, whereas Dai *et al.* (2011) created a joint model for co-reference resolution and gene normalization and Liu *et al.* (2012) conducted entity recognition and normalization jointly for tweets. In this paper, we use the Markov Logic Network (MLN) (Richardson *et al*., 2006), a joint model that combines first order logic and Markov networks, to capture the bottom-up decisions derived from the process illustrated in Figure 2. This model captures the contextual information of the recognized entities for entity disambiguation, as well as the constraints used when linking an entity mention to a database entry. For example, an entity mention can only be linked to a database entry when the mention has not been recognized as a Nil.

Existing EL evaluation metrics assess a system's performance in terms of the effectiveness of database curation (Morgan *et al*., 2008) or question answering (QA) accuracy (McNamee, Dang, *et al*., 2009). In addition, we evaluate our system at a fine-grained entity by entity level. Such evaluation is more relevant to information extraction tasks, such as the bio-molecular event extraction task (Kim *et al*., 2009).

When considering EL tasks from the entity level, one challenge is the lack of contextual information for disambiguating each individual entity. The major scheme of traditional entity disambiguation approaches relies on domain knowledge derived from entities' profiles and contextual features extracted within a predefined content window. Rule-based (Dai *et al*., 2010; Jörg Hakenberg *et al*., 2008), vector space models (Cucerzan, 2007), and machine learning approaches (Crim *et al*., 2005; Mihalcea *et al*., 2007; Milne *et al*., 2008) have been proposed to disambiguate entity mentions individually. Nevertheless, the context is unclear under certain circumstances. Take the sentence "The synthetic replicate of **urocortin** can bind with high affinity to Type 1 and Type 2 CRF receptors" as an example. The sentence itself does not explicitly provide any clues to help computer programs determine the identity of the gene mention "urocortin", which has at least eight ambiguous Entrez Gene IDs. One approach is to expand the context window used for disambiguation to the paragraph level. Nevertheless, a paragraph described in a biomedical article usually incorporates several pieces of information in its description, which may not be related directly to a target entity instance and leads to the failure of traditional EL approaches.

Our idea of dealing with the challenge of deficient contextual information for disambiguating individual entity instances is to model dependencies among entities across sentences in the same paragraph. These dependencies have been ignored by most of the previous EL approaches. We refer to our approach as the collective EL, which is developed by

considering the relational information hidden among entities. In the following sections, we first give formal definitions of the EL tasks mentioned above, followed by an introduction of MLN and a description of the main ideas of the proposed EL method with the formulation of the collective EL approach.

2. Entity Linking Problem Definition

This section gives formal definitions of all related EL tasks.

Definition 1: Instance-based Entity Linking Problem

Let $M = (m_1, m_2, \ldots)$ denote a sequence of entities mentioned in an article A. The surface name of m_i is denoted by $Name(m_i)$. The named entity type of m_i is $EntityType(m_i)$. The surrounding context of m_i can be extracted by $Context(m_i)$. Given a KB containing a set of entries $ID = \{id_1, id_2, \ldots\}$, each of which organizes knowledge related to an entity, the instance-based EL problem is defined as finding a mapping function $LinkTo(m_i)$ that maps each m_i in M to a unique entry id_i in ID and satisfies the constraint $\left|LinkTo(m_i) : m_i \in M\right| = \left|M\right|$.

In instance-based gene mention linking (GML), only entities whose $EntityType(m_i)$ belong to "gene" are considered for evaluation. Both the gene normalization task in BioCreAtIvE (Morgan *et al.*, 2008) and the EL task in the KB population (McNamee & Dang, 2009a) can be subsumed into Definition 1. In BioCreAtIvE gene normalization, the developed system should satisfy the equation $\left|LinkTo(m_i) : m_i \in M\right| \le \left|M\right|$. We refer to this task as the article-wide EL problem.

Definition 2: Article-wide Entity Linking Problem

Let $M = \{m_1, m_2, \ldots\}$ denote a set of entities mentioned in A. Given the entries $ID = \{id_1, id_2, \ldots\}$ in a KB and the mapping function $LinkTo(m_i)$, the article-wide EL problem satisfies the constraint $\left|LinkTo(m_i) : m_i \in M\right| \le \left|M\right|$.

On the other hand, the KB population EL task only considers one certain entity m_i mentioned in A. We refer to this task as the article-wide "salient entity" linking problem, in accordance with the Wikipedia style manual, in which only the salient entity and its related entities should be linked in wikification. Excessive links would obstruct the readers in following the article by drawing attention away from important links (Mihalcea & Csomai, 2007).

Definition 3: Article-wide Salient Entity Linking Problem

Let $M = m_i$ denote the salient entity mentioned in A. Note that, in encyclopedia-style articles, $\left|M\right| = 1$ because the same surface name described in such articles should refer to the same instance. Given the entry set $ID = \{id_1, id_2, \ldots\}$ of a KB, the purpose of the article-wide salient EL problem is to find the mapping function $LinkTo(m_i)$ that links m_i to a unique entry id_i in E.

Note that, in the KB population EL subtask (pertained to the article-wide salient EL problem), the salient entity is given. Nevertheless, in the instance-based GML or the BioCreAtIvE gene normalization (pertaining to the article-wide EL problem) tasks, the systems must also deal with the entity recognition/classification problem.

3. First-order Logic and Markov Logic Networks

Markov logic is a statistical relational learning language based on first-order logic (FOL) and Markov networks. In this section, we consider FOL and Markov networks in terms of the GML task.

In FOL, the formulae are constructed using four types of symbols: constants, variables, functions, and predicates. For GML, a constant symbol may represent a gene mention (*e.g.* "CD59") or its unique database entry (*e.g.* the Entrez Gene ID "966"). If variables and constants are type-specific, their range can only cover objects of the corresponding type. To give an example, the variable *y*'s range covers all Entrez Gene database IDs. Predicate symbols are used to represent the relations between *terms*; for example, we can define the predicate, *LinkTo*(*x*, *y*), to indicate that a gene mention (the variable *x*) should be linked to an entry (the variable *y*). Formulae are constructed recursively from predicates applied to a tuple of terms by through use of logical connectives and quantifiers. Then, we can model the EL task by introducing a set of logical predicates. For instance, we can define the predicate *Candidate*(*i*, *j*) to indicate that the gene mention *i* can be mapped to an entry *j*. The predicate captures information about gene mentions and their corresponding candidate database entries. Through this predicate, we can infer whether a gene mention is unambiguous. Then, we can use the following formula

Formula 1: $\forall x \exists! id . Candidate(x, id) \Rightarrow LinkTo(x, id)$

to model the concept that, when a gene mention is mapped to only one entry, it should be linked to that entry. Note that we use the symbol, $\exists!$, to refer to a uniqueness quantification.

A first-order KB is a set of hard constraints on the set of ground atoms of predicates (or so-called *possible worlds*). If a world violates any formula, it has zero probability. In most domains, however, it is very difficult to derive non-trivial formulae that are always true. Markov logic softens these constraints to handle uncertainty by associating each formula with a weight that reflects the strength of a constraint. Ideally, if we could define a formula with a proper weight for its distribution, a world in which the formula is satisfied would have a higher probability than a world in which it is not. In Markov logic, a set of weighted formulae is called a MLN.

Definition 4: Markov Logic Network

An MLN *L* is a set of pairs (F_i, w_i), where F_i is a formula in FOL and w_i is a learned weight

corresponding to the F_i whose value is a real number. In combination with a finite set of constants $C = \{c_1, c_2, \ldots, c_{|C|}\}$, it defines a Markov network $M_{L,C}$ as follows: $M_{L,C}$ contains one node for each possible grounding of each predicate appearing in L. The value of the node is 1 if the ground predicate is true, otherwise it is 0.

Based on the definition, we can generate a graph structure of the ground Markov network where there is an edge between two nodes of $M_{L,C}$ if the corresponding ground atoms appear together in at least one grounding of one formula in L. Thus, the predicates in each ground formula form a clique in $M_{L,C}$. Each clique in the graph is associated with a potential function ϕ_i .The joint distribution of a set of variables X represented by $M_{L,C}$ then is defined by:

$$\mathrm{P}(X = x) = \frac{1}{Z}\prod_i \phi_i(x_{\{i\}})^{n_i(x)}$$

where $x_{\{i\}}$ is the state of the ith clique (*i.e.*, the exact values of the predicates that appear in that clique F_i), $n_i(x)$ is the number of true groundings of F_i in x, and $\phi_i(x_{\{i\}}) = e^{w_i}$. Z is the partition function given by $Z = \sum_{x \in X}\prod_i \phi_i(x_{\{i\}})$. Markov networks are often represented as log-linear models in which each clique is replaced by an exponential weighted sum of the features of the state, leading to

$$\mathrm{P}(X = x) = \frac{1}{Z}\exp\left(\sum_i w_i f_i(x)\right)$$

In our implementation, f_i is a binary feature, $f_i(x) \in \{0,1\}$.

4. The Proposed Entity Linking System for Gene Mention Linking

Figure 3 illustrates the input of the proposed EL system developed for GML and the FOL predicates defined for the corresponding bottom-up stages. The input is an article, such as a biomedical abstract. The given article is first processed by a gene mention recognizer to identify gene mention boundaries, and the employed gene mention mapper then maps each recognized gene to a list of candidate identifiers, based on a lexicon compiled from the Entrez Gene database.

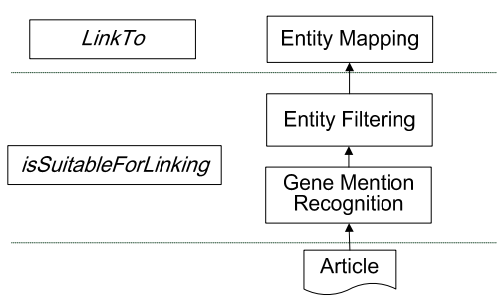


Figure 3. MLN hidden predicates defined for each stage.

Ideally, we should be able to treat all recognized gene mentions as candidates, and proceed directly to the entity mapping task. Nevertheless, the employed recognizer may generate false positive gene mentions. Such mentions can be classified into two types: those that do not belong to any entity class and those that belong to classes that are not the curation target (Nils). In GML, Nils appear when the gene mentions are DNA polymerases, or in a specific organism that is not considered. To capture the concept, we define the predicate *isSuitableForLinking*(*x*), which indicates that the gene mention *x* of the article is suitable for linking to an entry. For entity mapping, we use the predicate *LinkTo*(*x, id*) to represent that the gene mention *x* must be linked to the database entry *id*. As the objective of the entity mapping task is to determine a unique KB entry for each entity, we must define a formula to ensure that the constraint is satisfied. Regarding GML, we use the following formula to prevent a gene mention from associating with more than two identifiers.

Formula 2: Entity Mapping Constraint

$$LinkTo(x,id_i) \wedge id_i \neq id_j \Rightarrow \neg LinkTo(x,id_j)$$

4.1 Formulation of the Instance-based GML

Within the machine learning community, classification is typically done on each object independently without taking into account any underlying relation that connects the objects. In most of the individual EL formulations, an individual classifier is employed to assign a probability to the linked ID of an individual instance independently of the linked IDs of other instances. For example, the following formula expresses that, if the chromosome location information of the entity mention *x*, which has the KB entry *id* as its candidate ID, can be found in the surrounding text, *x* should be linked to *id*.

Formula 3: $hasChromosomeInfo(id) \wedge Candidate(x,id) \Rightarrow LinkTo(x,id)$

Other useful biographical information for locally disambiguating gene mentions includes tissues, gene ontology, and PPI. Some researchers (Hakenberg *et al*., 2007; Lai *et al*., 2009) have used this information for individual GML. Table 1 shows the observed predicates and formulae defined for this individual approach. We take *hasPPIPartnerRank* as another example. For the individual GML process, we define the following formula.

Formula 4: Individual PPI

$$\exists! id.hasPPIPartnerRank(x,id,1) \wedge \exists w.hasWord(w) \wedge isPPIKeyword(w) \Rightarrow LinkTo(x,id)$$

implies that the gene mention *x* should be linked to the *id* that has the most PPI partners. The other predicates, including *hasGOTermRank* and *hasTissueRank*, follow a similar concept, in which the context is matched with the corresponding keywords to determine the frequency in the given abstract text.

Table 1. Observed predicates and formulae defined for entity mapping.

	Candidate(*x, id*) *hasChromosomeInfo*(*id*) *hasWord*(*w*): the abstract contain a word *w*. *isPPIKeyword*(*w*), *isPPIPartner*(id_1, id_2) *hasPPIPartnerRank* (*x, id, r*) *hasGOTermRank*(*x, id, r)* *hasTissueTermRank*(*x, id ,r)* *hasDictionaryMatchRank*(*x, id, r*) *hasPrecedingWord*(*x, w, l*), *hasFollowingWord*(*x, w, l*) *hasUnigramBetween*(*x, y, w*)
Variable Type	*x*: integer that refers to the *x*th gene mention in the given article (similarly *y* refers to the *y*th gene mention) *id*: an Entrez Gene ID, which refers to the linked KB entry. *w*: a word. *r*: integer that refers to the rank of the matching. *l*: integer that refers to a context window length.
Formulae	$Candidate(x,id) \wedge hasChromosomeInfo(id) \Rightarrow LinkTo(x,id)$ $hasWord(w) \wedge PPIKeyword(w) \wedge Candidate(x,id) \wedge \exists! id_i.MostPPIPartners(id_i) \wedge id_i = id \Rightarrow LinkTo(x,id)$ $Candidate(x,id) \wedge \exists! id_i.MostGOTerms(id_i) \wedge id_i = id \Rightarrow LinkTo(x,id)$ $Candidate(x,id) \wedge \exists! id_i.MostTissueTerms(id_i) \wedge id_i = id \Rightarrow LinkTo(x,id)$ $\exists! id.Candidate(x,id) \Rightarrow LinkTo(x,id)$ $\exists! u,id.Candidate(x,id) \wedge hasPrecedingWord(x+1,u) \wedge u = "(" \wedge hasUnigramBetween(x,x+1,u) \wedge hasFollowingWord(x+1,")") \Rightarrow LinkTo(x+1,id)$ $\exists! u,id.Candidate(x+1,id) \wedge hasPrecedingWord(x+1,u) \wedge u = "(" \wedge hasUnigramBetween(x,x+1,u) \wedge hasFollowingWord(x+1,")") \Rightarrow LinkTo(x,id)$

A drawback of individual EL classifiers is that, when they decide the linking entry of an entity, they cannot utilize information about the linked entries and features of other entities in the same article. Nevertheless, those entity instances can be related, and the interrelationship can be used to improve the EL performance. Furthermore, there are strong dependencies among the unknown IDs of the instances, which could either be a true positive entity mention or a Nil. These dependencies are highly nonlocal.

Collective classification refers to the task of inferring labels for a set of objects using not just their attributes, but also the relations among them (Sen *et al.*, 2008).

Definition 5: Collective Classification

Given a network *N*, a node *n* in *N*, and the label set *L*, there are three distinct feature types that can be utilized to determine the label *l* of *n*, where $l \in L$.

1. The observed features of *n*.
2. The observed features (including observed labels if they are known) of nodes in the neighborhood (related nodes) of *n*.
3. The unobserved labels of nodes in the neighborhood (related nodes) of *n*.

In our formulation for EL, for a given article, the candidate database entries of all recognized entities form the network *N*. A mention's candidate entry and order form the node *n* = (*id*, *order*) in *N*, and an edge exists between two nodes if they have dependencies. In this work, the dependencies are constructed based on two main ideas: the discourse salience property in centering theory (Grosz *et al*., 1995) and the protein-protein interaction (PPI) association.

4.1.1 Discourse Salience

Discourse salience is a phenomenon where, in a given discourse, there is precisely one entity that is the center of attention. This entity is mentioned over and over again, which makes it more salient than others. We utilize this phenomenon to improve the instance-based EL confidence. Suppose that *id* is a candidate database entry for several entities in a discourse, we then can assume that *id* is more salient than other database entries. If the EL system can link one of these mentions to *id* with high confidence, then the system is more likely to be able to link all of the other mentions to *id* as well.

4.1.2 Protein-protein Interaction

Similarly, the idea of employing the PPI association allows us to express the concept that a gene mention *y* should be linked to id_j if another gene mention *x* has been linked to id_i and id_i forms an interaction with id_j.

In order to capture the concepts above, the order of all individual instances described in an abstract are leveraged to build dependencies in our formulation. The lack of local contextual information then can be resolved by the constructed dependencies. In our work, the salience collective is written as follows in Markov logic.

Formula 5: Salience collective

$$Pr\,ecede(x,y) \wedge LinkTo(x,id) \wedge Candidate(y,id) \Rightarrow LinkTo(y,id)$$

If the database entry *id* is linked to an entity *x* that precedes the current mention *y* and *id* is a candidate entry of *y*, then the current entity *y* should also be linked to *id*. This formula is similar to the transition feature of the linear-chain conditional random fields (Lafferty *et al.*, 2001), which can be implemented in Markov logic as follows.

Transition feature: $Pr\,e\,cede(x,y) \wedge Label(x,+L) \Rightarrow Label(y,+L)$

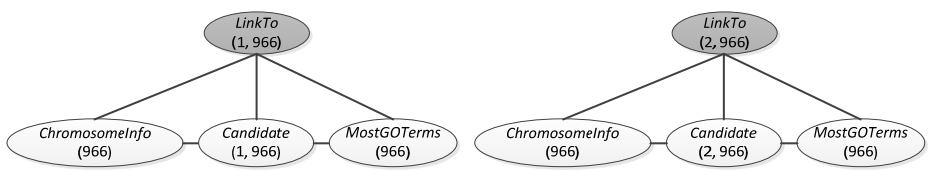
Note that the symbol "+" in the above formula directs the MLN learning algorithm to associate the formula with a different weight depending on variables containing the "+" notation.

We define the predicate *PPIPartner*(id_i, id_j), whose value is true if id_i and id_j form a PPI pair. We then use the following formula to capture the PPI association concept.

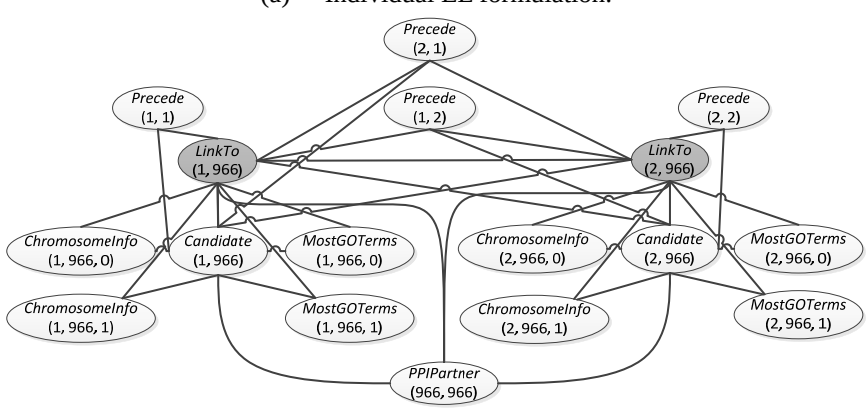
Formula 6: PPI

$$LinkTo(x,id_i) \wedge Candidate(y,id_j) \wedge PPIPartner(id_i,id_j) \Rightarrow LinkTo(y,id_j)$$

Based on these two collective formulae, Figure 4 compares the ground Markov network (b) of our collective formulation with the traditional individual approach (a). In Figure 4 (a), the individual approach considers the likelihoods stating the similarity of the current context with the domain knowledge of the recognized entity, including chromosome location (*ChromosomeInfo*) and gene ontology (*MostGOTerm*). Comparing Figure 4 (b) with (a), our collective formulation captures the dependencies among entities, allowing the information to be employed in the GML decision.



(a) Individual EL formulation.



(b) Collective EL formulation.

Figure 4. Ground Markov network obtained by applying example formulae to the constants x, y = {1, 2}, c = {0, 1}, and id = {966}.

4.2 Formulae for Entity Recognition/Classification and Filtering

The hidden predicate *isSuitableForLinking* captures the decisions made after the entity recognition/classification stage. When the gene mention *x* is linked to an identifier *id*, we employ the following formula to ensure that it is an entity suitable for linking.

Formula 7: $LinkTo(x,id) \Rightarrow isSuitableForLinking(x)$

Note that the formula models the bottom-up decision, as shown in Figure 2. The identifier *id* does not have to be linked to the gene mention *x* proposed by the entity recognition/classification stage. Nevertheless, the *id* cannot be assigned to the gene mention *x* that has not been proposed as a potential entity.

Our first formula for *isSuitableForLinking* treats all gene mentions as potential entities:

$hasName(x,n) \Rightarrow isSuitableForLinking(x)$.

The other formulae are constructed using the observed predicates defined in Tables 1 and 2 to check the contextual information. For example,

$hasFirstWord(x,+w) \wedge isSpeciesTerm(+w) \Rightarrow isSuitableForLinking(x)$

implies that the suitability of a certain gene mention for linking depends on whether or not the first word is a keyword for a certain species.

Table 2. Observed predicates for entity filtering.

hasName(*x*, *n*) *hasFirstWord*(*x*, *w*), *hasLastWord*(*x*, *w*) *hasPrefix*(*x*, *ch*, *d*, *l*): the *x*th gene mention has a prefix *ch* of length *l*, and the prefix's case is the same as its following character (*d* = 0) or different (*d* = 1). *isSpeciesTerm*(*w*), *isAllUpperCase*(*i*), *hasPartOfSpeech*(*x*, *k*, *p*) *isContainedMoreSpecificMentions*(*x*)	
Variable Type	*n*: a word or a sequence of words that refer to the surface name of a gene mention. *ch*: a character. *d*: an integer. *k*: the *k*th index of the gene mention. *p*: a part-of-speech

5. Experimental Results and Discussion

5.1 Experimental Setup

5.1.1 Evaluation Metrics

We use three metrics to evaluate our approach and compare it with other GML methods. The first and second metrics used the standard precision, recall, and F-measure metrics (PRF) at

two resolutions (article and instance).

Article-wide evaluation used the standard used in the BioCreAtIvE challenge (Hirschman *et al*., 2005), which is designed to determine an GML system's performance as an aid for the curation of biological databases. The GML system outputs a list of IDs for a given article, and this list is compared to the gold standard ID list. The PRF scores are calculated based on the sums of true/false positives/negatives (TP, TN, FP, FN).

Instance-based evaluation measures the GML performance at a fine-grained resolution (Dai *et al*., 2011). In contrast to the first metric, the PRF scores are calculated based on the sums of TP, TN, FP, and FN for all instances in the test dataset. Therefore, under this criterion, an FP can link a true gene mention to the wrong KB entry or link a false gene mention to any entry, while an FN can link a true gene mention to the wrong entry or fail to recognize a true gene mention. For TP/FP/FN, we need to determine when the predicted boundary matches that of the gold standard. Most pure entity recognition tasks use "exact-matching" as the primary criterion. Under this criterion, a candidate gene mention can only be counted as a TP if both its left and right boundaries fully coincide with the gold answer. In a real case, however, a gene mention can be tagged in several ways (*e.g.*, "serum <entity>LH</entity>" and "<entity>serum LH</entity>" are both correct), which are intrinsic to the annotation of any gene mention corpus, whether developed by humans or machines (Tsai *et al*., 2006), and may depend on the annotator's perspective (Franzén *et al*., 2002). Furthermore, for the GML task, the correctness of the linked entry is more important than its boundary. Therefore, we used the approximate-match to determine the boundary criterion. For example, a TP is counted when a machine-linked gene mention is a substring of the gold standard-linked gene mention or vice versa and the linked entry is equal to the gold entry.

The third metric, the mean accuracy across all queries, considers the QA perspective. EL is important in QA systems because the systems rely on data from multiple sources, so name ambiguity will lead to wrong answers and poor results. We adopt the evaluation metrics used in text analysis conference KB population (KBP) track 2009 (McNamee & Dang, 2009a) to report $\text{Accuracy}_{\text{micro}}$ and analyze the results from the QA perspective [1].

5.1.2 Datasets

In the experiments, we used the training and test sets (281 and 262 abstracts, respectively) released by the BioCreAtIvE gene normalization task (Morgan *et al*., 2008) for article-wide evaluation. The corpus contains annotations for human genes that are linked to IDs in the Entrez Gene database. Although the gold answers contain each ID's surface name, they do not give the exact location of the corresponding gene mention in the abstract. To obtain

[1] For details please refer to http://apl.jhu.edu/~paulmac/kbp.html.

fine-grained evaluation results, our in-lab biologists compiled an instance-based GML corpus by annotating the exact location and the boundary of the IDs' gene mentions (Dai *et al*., 2012). After compiling the corpus, we performed three-fold cross validation (CV) on the training dataset to optimize the weights and formulae and to evaluate its performance on the test set.

In QA evaluation, as defined in KBP, the associated document is used to provide contextual information that might be useful for linking. We paired the surface names with their corresponding documents as the input query. For each query, the corresponding gold answer could be 1) an Entrez Gene ID or 2) a Nil in cases where our biologists annotated the entity as a mention without associating it with any IDs. Table 3 shows the generated query/answer pairs based on the BioCreAtIvE Corpus.

Table 3. The statistics of the generated query/answer pairs on the BioCreAtIvE corpus.

Dataset	# of queries	# of Nil	# of Entities
Training	1073	87	1132
Test	1266	66	1154

5.1.3 Model Configurations

To assess the performance of our models and determine the possible gains that can be achieved by considering a collective model and a joint model of the bottom-up stages, we designed two configurations. The first configuration was the collective model, which used all of the disambiguation formulae defined in Section 4.1 (denoted as **CM**). The constructed Markov network resembles Figure 4 (b) with additional grounding for the predicates and individual formulae defined in Table 1. The second configuration further included the formulae defined in Section 4.2 on CM to build a joint model (denoted as **JCM**). This work used the 1-best Margin Infused Relax online learning Algorithm (MIRA) (McDonald *et al*., 2005) for learning weights and employed cutting plane inference (Riedel, 2008) with integer linear programming as its base solver for inference at test time as well as during the MIRA online learning process.

In addition, we compared the first configuration (CM) with two GML approaches: Lai *et al*. (2009)'s rule-based approach and Crim *et al*. (2005)'s maximum entropy (ME) approach, which handled the GML task as an individual classification problem. The ME approach was adopted with the individual features described in Section 4.1.

Finally, to compare the stage-based approach with the second configuration (JCM), which performs joint filtering and linking, we trained a separate ME model via the features designed for the *isSuitableForLinking* stage in Figure 3. This model then was combined with the first configuration, the Rule-based approach and ME approach, which we denote as CM_{stage},

Rule-based$_{stage}$, and ME$_{stage}$, respectively.

For the above configurations, we employed Lai *et al.*'s system[2] to recognize all gene mentions and generate mapped candidate IDs for each mention. All configurations were based on the same candidate ID sets.

In the next sub-section, we first discuss the fine-grained resolution results. Then, we derive BioCreAtIvE's evaluation results by simply merging the linked identifiers in all indices and removing duplicated identifiers. Finally, the results are displayed from the QA perspective.

Table 4. The three-fold CV results on the training set using instance-based criterion. Our models are highlighted in bold.

Config.	P (%)	R (%)	F (%)	Diff (F)
No Disambiguation	81.0	48.0	60.3	-
Salience Collective	79.9	49.6	61.3	+1.0
PPI Collective	79.3	51.2	62.2	+1.9
Rule-based	71.7	54.0	61.6	+1.3
ME	79.8	48.9	60.5	+0.5
CM	73.5	55.9	63.5	+3.2
Rule-based$_{stage}$	71.7	54.0	61.6	+1.3
ME$_{stage}$	86.4	46.9	60.8	+0.8
CM$_{stage}$	73.5	55.9	63.5	+3.2
JCM	79.9	54.9	65.1	+4.8

5.2 Experiment Results

Table 4 shows the fine-grained results derived on the training set. The employed system's linking performance without applying any disambiguation approaches is shown in the first row (No Disambiguation.) Our model can simulate the same PRF scores when only Formula 1 is applied. The last column shows the improvement of F-score over the baseline after implementing different GML disambiguation methods.

It can be observed that, by adding the salience collective (Formula 5) without any disambiguation formulae and domain knowledge, the recall rate is improved by 1.6%, which results in an improved F-score. This demonstrates that a scientific article often contains repetitive information, such as key genes, which can be captured by the formula. Furthermore,

[2] The employed system can be downloaded from https://sites.google.com/site/potinglai/downloads.

the PPI collective combining the domain knowledge achieves a higher PRF-score, even outperforming the Rule-based and ME.

Table 5. Results derived on the test set.

Metrics	Fine-grained Resolution (%)				Aids for Curation (%)			
Config.	P	R	F	Diff	P	R	F	Diff
No Disambiguation	80.7	56.3	66.3	0	77.3	71.5	74.3	0
Salience Collective	79.5	59.0	67.7	+1.4	77.2	71.3	74.1	-0.2
Rule-based	72.9	63.9	68.1	+1.8	82.6	83.4	83.0	+8.7
ME	79.2	58.2	67.1	+0.8	88.8	79.0	83.6	+9.3
CM	73.8	64.3	68.7	+2.4	86.1	83.0	84.5	+10.2
Rule-based$_{stage}$	73.7	64.2	68.7	+2.4	84.1	83.7	83.9	+9.6
ME$_{stage}$	80.2	58.4	67.6	+1.3	90.2	79.0	84.3	+10
CM$_{stage}$	74.3	64.3	69.0	+2.7	87.9	83.2	85.5	+11.2
JCM	77.5	63.7	69.9	+3.6	87.7	83.8	85.7	+11.4

The results derived on the test set by fine-grained resolution and aided for curation and QA are shown in Tables 5 and 6, respectively. In summary, we observe that the collective GML method consistently outperforms the compared methods under the three criteria. Moreover, the comparison of our joint model (JCM) and the stage models shows that the joint model performs better under all evaluation metrics.

Table 6. QA results on the test set.

Config.	Accuracy
No Disambiguation	65.7
Salience Collective	67.2
Rule-based	72.4
ME	66.8
CM	73.1
Rule-based$_{stage}$	73.0
ME$_{stage}$	67.0
CM$_{stage}$	73.3
JCM	73.5

We also observe that adding the salience collective reduces the recall rate in the aid for curation evaluation. According to our analysis, the salience collective improves the recall in

the fine-grained evaluation. In contrast, for database curation, the collective tends to improve the overall precision. Nevertheless, it reduces the recall. By removing the salience collective from CM, the P improved by 0.8% but the R reduced by 0.7% in the test set. This phenomenon is reasonable because adding the dependency causes the model to link mentions with previous linked IDs.

Finally, the results also reveal the performance gap (approximately 15.8%) when we want to employ the GML system, which is evaluated in terms of the database curation criterion with 80+% F-score on advanced IE tasks, such as relation or event extraction.

5.3 Discussion

Evaluating EL from the fine-grained perspective allows us to analyze the task in detail. In this section, we describe our findings and propose potential research directions.

One advantage of employing MLN in our EL modeling is that it is easy to model arbitrary longer range dependencies, as expressed by Formula 5 and Formula 6. It is difficult to model such dependencies using ME. As shown in Tables 4 and 5, adding the collectives improves the fine-grained EL performance.

Another advantage of our GML approach is that it is flexible and can be applied quickly in real-world applications. The EL task usually is defined as linking a mention to a unique entry. Nonetheless, in the biomedical field, there are some mention descriptions that cannot be linked to unique IDs. The following are some examples extracted from our corpus:

1. **ABCB9 protein** appears to be most highly expressed in the Sertoli cells of the seminiferous tubules in *mouse and rat testes*.
2. cDNA cloning and chromosomal localization of the *human and mouse* isoforms of **Ksp-cadherin**.
3. **p63** was detected in a variety of *human and mouse* tissues.

The GML system cannot link each of the gene mentions in the above sentences to just one ID. Our model can deal with these cases by modifying the constraint of Formula 2 with a larger cardinality or introducing additional formulae to determine the cardinal constraint dynamically.

Our experiment results also raise an interesting question: What causes the huge performance gap between the fine-grained and database curation evaluations? A closer look at the bottom-up approach is useful in answering this question. Several works have studied the boundary issue in entity recognition (J. Finkel *et al.*, 2005; Tsai *et al*., 2006), and this issue was found to have a significant effect on the performance of GML. For example, consider the following sentence:

"<entity id=3083>Hepatocyte growth factor (HGF) activator</entity> is a serine protease responsible for proteolytic activation of <entity id=3082>HGF</entity> in response to tissue injury"

All of the employed gene mention recognition systems and the three open available gene mention recognition systems[3,4,5] separate the first gene mention (ID:3083) into at least one mention ("hepatocyte growth factor" or "HGF"). The incorrect boundary leads to errors in the entity mapping stage, and it could result in the extraction of an incorrect self-activation event: <entity id=3082>HGF</entity> activates <entity id=3082>HGF</entity>. An experiment conducted on the test set showed that our MLN model could achieve an F-score of 79.4% from the fine-grained IE perspective if we replaced the predicted mentions' boundaries with their corresponding overlapped gold standard boundaries. These results show that a hybrid approach combined with entity-centric boundary expansion is required before entity mapping. For instance, if we input the example sentence to a syntactic parser like Enju[6] and find that the adjacent words "Hepatocyte growth factor (HGF) activator" belong to the same noun phrase and the word "activator" is a legal suffix for a gene mention, it implies that we can expand the boundary.

The result also motivates us to reconsider the bottom-up EL approach. Are the results of entity recognition/classification a prerequisite for GML? We raise this question because, under the bottom-up approach, the entity mapping process still needs to deal with the boundary issue in order to generate more candidate identifiers, as shown by the previous example sentence. Moreover, the disambiguation process needs to look for knowledge, such as species information, surrounding the gene mention's boundary, which is usually located in the same noun phrase. It has been shown that joint learning of multiple types of linguistic structures in models can produce more consistent outputs. A feasible approach would be to treat noun phrases as potential candidate gene mentions and employ a mapping algorithm to generate identifiers for each noun phrase. Using the proposed approach to model biographical information and the dependencies between noun phrases, we can perform joint learning and inference for gene mention recognition and linking. This issue will be the major direction of our future research. For chunking parsing, there are several openly available tools, such as GENIA tagger[7], OpenNLP[8], and Lingpipe[9] package. Kang *et al*. (2011) have reported that the OpenNLP package performs noun-phrase chunking, best among the six state-of-the-art chunkers specifically for the biomedical domain. Therefore, we will use the OpenNLP

[3] http://pages.cs.wisc.edu/~bsettles/abner/

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[5] http://cbioc.eas.asu.edu/banner/webBasedBannerStart.html

[6] http://www-tsujii.is.s.u-tokyo.ac.jp/enju/demo.html

[7] http://www.nactem.ac.uk/GENIA/tagger/

[8] http://opennlp.apache.org/

[9] http://alias-i.com/lingpipe/

package as the first step to generate candidate noun phrases, and we may consider combining more results from different chunkers through a voting strategy to further improve the chunking performance.

Table 7. The hardest queries in the test set.

<query id, name, docid>
<#1, AIP1, 9647693>
<#2, TR1, 10455115>
<#3, PAGE1, 9651357>
<#4, UGT2B11, 8333863>

Finally, we report our observations on the hardest queries under QA evaluation. Table 7 lists the queries that could not be answered correctly by any of the employed methods. After checking each error event carefully, we found that our model outputs Nil for Query #1 due to the absence of gene biographical information in the context. The gene mention of Query #2 is an abbreviation, and the query is affected by a problem similar to the ambiguous acronym problem discussed in the KBP 2009 track (McNamee, Dang, *et al*., 2009). Our model fails to output correct IDs for #3 and #4 because no distinction is made between matches of official symbols and synonyms when searching for candidate IDs. In our current work, the matches of official symbols and those of synonyms share the same predicate. We believe that appending more predicates and formulae corresponding to these two types of matches will improve our system's accuracy.

6. Conclusions

In this paper, we give formal definitions for EL tasks, including instance-based EL, article-wide EL, and article-wide salient EL. We then present a novel approach that employs MLN to jointly model bottom-up decisions in a specific EL task-GML. A collective formulation for instance-based GML is introduced with several useful formulae, including the dependencies among IDs, which can be used for GML disambiguation. Moreover, the benefit of predicting suitable mentions and their IDs jointly in contrast to the stage-based approach is illustrated, which selects mentions before linking IDs. Our experiments provide the first comprehensive gene mention evaluation results from three different perspectives and highlight problems that need to be addressed in the future, including the assignment of non-unique identifiers, the boundary issue, and the direction for joint entity recognition and linking.

Acknowledgement

This research was supported in part by the National Core Facility Program for Biotechnology (Taiwan Bioinformatics Consortium of Taiwan) grant NSC-100-2319-B-010-002, the National Science Council of Taiwan grant NSC-102-2218-E-038-001, as well as the research grant TMU101-AE1-B55 of Taipei Medical University.

Reference

Bunescu, R., & Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Crim, J., McDonald, R., & Pereira, F. (2005). Automatically Annotating Documents with Normalized Gene Lists. *BMC Bioinformatics*, *6*(Suppl 1), S13.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic.

Dai, H.-J., Chang, Y.-C., Tsai, R. T.-H., & Hsu, W.-L. (2011). Integration of gene normalization stages and co-reference resolution using a Markov logic network. *Bioinformatics*, *27*(18), 2586-2594.

Dai, H.-J., Lai, P.-T., & Tsai, R. T.-H. (2010). Multistage Gene Normalization and SVM-Based Ranking for Protein Interactor Extraction in Full-Text Articles. *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS*, *7*(3), 412-420.

Dai, H.-J., Wu, C.-Y., Tsai, R. T.-H., & Hsu, W.-L. (2012). From Entity Recognition to Entity Linking: A Survey of Advanced Entity Linking Techniques. In *Proceedings of the 26th Annual Conference of the Japanese Society for Artificial Intelligence*, Yamaguchi, Japan.

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 (COLING 2010)*, Beijing.

Finkel, J., Dingare, S., Manning, C., Nissim, M., Alex, B., & Grover, C. (2005). Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics*, *6*(Suppl 1), S5.

Finkel, J. R., & Manning, C. D. (2009, June). Joint parsing and named entity recognition. In *Proceedings of NAACL 2009*.

Franzén, K., Eriksson, G., Olsson, F., Asker, L., Lidén, P., & Cöster, J. (2002). Protein names and how to find them. *International Journal of Medical Informatics*, *67*(1-3), 49-61. doi: Doi: 10.1016/s1386-5056(02)00052-7

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203-225.

Hakenberg, J., Plake, C., Leaman, R., Schroeder, M., & Gonzalez, G. (2008). Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, *24*(16), 126-132. doi: 10.1093/bioinformatics/btn299

Hakenberg, J., Plake, C., Leaman, R., Schroeder, M., & Gonzalez, G. (2008). Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, *24*, i126 - 132.

Hakenberg, J. o., Royer, L., Plake, C., Strobelt, H., & Schroeder, M. (2007). Me and my friends: gene mention normalization with background knowledge. In *Proceedings of Second BioCreAtIvE Challenge Evaluation Workshop*.

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, *6*(1).

Kang, N., van Mulligen, E. M., & Kors, J. A. (2011). Comparing and combining chunkers of biomedical text. *J Biomed Inform*, *44*(2), 354-360. doi: 10.1016/j.jbi.2010.10.005

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., & Tsujii, J. i. (2009). Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, Boulder, Colorado.

Krauthammer, M., & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, *37*(6), 512-526.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*.

Lai, P.-T., Bow, Y.-Y., Huang, C.-H., Dai, H.-J., Tsai, R. T.-H., & Hsu, W.-L. (2009). Using Contextual Information to Clarify Gene Normalization Ambiguity. In *Proceedings of the IEEE International Conference on Information Reuse and Integration (IEEE IRI 2009)*, Las Vegas, USA.

Li, Y., Lin, H., & Yang, Z. (2009). Incorporating rich background knowledge for gene named entity classification and recognition. *BMC Bioinformatics*, *10*(1), 223.

Liu, X., Zhou, M., Wei, F., Fu, Z., & Zhou, X. (2012). Joint Inference of Named Entity Recognition and Normalization for Tweets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea.

McDonald, R., Crammer, K., & Pereira, F. (2005). Online Large-Margin Training of Dependency Parsers. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Michigan.

McNamee, P., & Dang, H. T. (2009a). Overview of the TAC 2009 Knowledge Base Population Track. In *Proceedings of the Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland USA.

McNamee, P., & Dang, H. T. (2009b). Overview of the TAC 2009 Knowledge Base Population Track. In *Proceedings of the Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland USA.

McNamee, P., Dang, H. T., Simpson, H., Schone, P., & Strassel, S. M. (2009). An Evaluation of Technologies for Knowledge Base Population. In *Proceedings of Test Analysis Conference 2009 (TAC 09)*, Gaithersburg, Maryland USA.

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*, Lisbon, Portugal.

Milne, D., & Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, Napa Valley, California, USA.

Morgan, A. A., Lu, Z., Wang, X., Cohen, A. M., Fluck, J., Ruch, P., . . . Hirschman, L. (2008). Overview of BioCreative II gene normalization. *Genome Biology*, *9*(Suppl 2), S3.

Poon, H., & Domingos, P. (2007). *Joint inference in information extraction*.

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning*, *62*(*Special Issue: Multi-Relational Data Mining and Statistical Relational Learning*), 107-136.

Riedel, S. (2008). Improving the Accuracy and Efficiency of MAP Inference for Markov Logic. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*, Helsinki, Finland.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., & Eliassi-Rad, T. (2008). Collective classification in network data. *AI Magazine*, *29*(3), 93.

Tsai, R. T.-H., Wu, S.-H., Chou, W.-C., Lin, C., He, D., Hsiang, J., . . . Hsu, W.-L. (2006). Various criteria in the evaluation of biomedical named entity recognition. B*MC Bioinformatics*, *7*(92), 14.

Zhang, W., Su, J., Tan, C. L., & Wang, W. T. (2010). Entity Linking Leveraging Automatically Generated Annotation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China.