From Entity Recognition to Entity Linking: A Survey of Advanced Entity Linking Techniques

Hong-Jie Dai[*1] Chi-Yang Wu[*1] Richard Tzong-Han Tsai[*2] Wen-Lian Hsu[*1]

[*1] Institute of Information Science, Academia Sinica
[*2] Department of Computer Science and Engineering, Yuan Ze University

Several research results have shown that specifying the information about certain entities is the most common information demand of information retrieval users. The needs should be answered by returning specific entities, their properties or related concepts instead of just any type of documents. While some search engines are capable of recognizing specific types of entities, true entity-oriented search still has a long way to go because of the high ambiguity in names across documents. Entity linking (EL) goes beyond the entity recognition task by linking a textual named entity mention to a knowledge base entry. It is a difficult task involving several challenges. This paper gives a survey of the EL tasks in the general and the biomedical domain. In addition, results of our latest EL work are provided for reference, which uncover new EL challenges found in biomedical text mining, along with discussions regarding their possible solutions.

1. Introduction

Finding information is one of the most common activities of Internet users. In most cases, query results are a mix of pages containing different entities that share the same name. In an ideal retrieval system, a user would simply input an entity or concept name and receive search results clustered according to the different entities/concepts that share that name. One method to approach such system is to include additional information in the indexed documents. Several experiments (R. Mihalcea & Moldovan, 2000; Rada Mihalcea & Moldovan, 2001; Sorden, Chang, & Nelson, 1999; Woods, 1997) with different indexed information have generated different and sometimes contradictory results. However, from these experiments one can conclude that although the problem of recognizing named entities (NEs) has been thoroughly evaluated within several shared tasks (Grishman & Sundheim, 1996; 2004; 2010; 2002; 2008), entity recognition results are still difficult to use directly because of the wide synonym and high ambiguity of variation in names across documents.

On the WWW, we need to deal with higher levels of ambiguity. Since there are so many documents on the Web, one single name will often refer to hundreds of different entities. But the distributions of mentions on the web are highly skewed. For each ambiguous name, there is usually one or two entities that dominate the vast majority of mentions sharing the same name (Sarmento, Kehlenbeck, Oliveira, & Ungar, 2009). For example, for an entity-oriented query, the Bing search engine[1] can group related pages organized by categories. The left part of Figure 1 shows an example of Bing search result of the query for the King of the Pop "Michael Jackson (MJ)" and his famous song "Blame it on the Boogie". The retrieved results look well. Unfortunately, if we want to find the same song performed by Michael George Jackson, the writer of the song, by using the query '"Michael George Jackson" Blame it on the Boogie', we can see that the retrieved videos are still biased to MJ's version (the right part of Figure 1) because MJ's version is more prominent. Obviously, for information retrieval/extraction (IR/IE) or question answering, these phenomena will harm their performance. Therefore, a highly accurate cross-document entity co-reference resolution or disambiguation algorithm is needed to increase the performance of IR/IE systems.

Figure 1: The revisit of the battle of the boogie.

Entity linking (EL) goes beyond NER task by linking a textual name of an entity to a knowledge base (KB) entry (McNamee, Dang, Simpson, Schone, & Strassel, 2009). Several preliminary results (Chu-Carroll & Prager, 2007; Chu-Carroll, Prager, Czuba, Ferrucci, & Duboue, 2006; Khalid, Jijkoun, & Rijke, 2008) have demonstrated that EL can improve search quality and question answering. This paper provides a survey of several EL-related researches and reports our recent progress on a biomedical EL task, which reveals several unexplored EL challenges.

2. Entity Linking

The EL problem comes up in many fields of research. In the database community, when merging multiple databases, an important step is to determine which records represent the same entity and should therefore be merged. Among database

Contact: Wen-Lian Hsu, Institute of Information Science, Academia Sinica, 128 Academia Road, Sec.2, Nankang, Taipei, Taiwan, ROC, +886-2-2788-3799, +886-2-2782-4814, hsu@iis.sinica.edu.tw

[1] http://www.bing.com/

researchers, this problem is described as object identification (Culotta & McCallum, 2005), and data de-duplication (Sarawagi & Bhamidipaty, 2002). In the AI community, the same EL problem is described as entity resolution (Bhattacharya & Getoor, 2007) and name matching (Bilenko, Mooney, Cohen, Ravikumar, & Fienberg, 2005). In the biomedical field, term identification (Krauthammer & Nenadic, 2004) and mapping (Aronson, 2001) or normalization (Hirschman, Colosimo, Morgan, & Yeh, 2005) are used to refer to the same concepts. Name identity uncertainty (Pasula, Marthi, Milch, Russell, & Shpitser, 2003), entity disambiguation (Dai, Tsai, & Hsu*, 2011; Dredze, McNamee, Rao, Gerber, & Finin, 2010) and record linkage (Winkler, 1999) are also commonly used to refer to this task (Elmagarmid, Ipeirotis, & Verykios, 2007).

There are several other tasks closely related to EL—for example, citation matching, the problem of deciding which citations correspond to the same publication (Lawrence, Giles, & Bollacker, 1999). Also similar is co-reference resolution or entity resolution: clustering entity mentions either within the same document or across multiple documents together, where each cluster corresponds to a single real-world entity (Dredze et al., 2010). This confusion of terminology has led to few cross-references between different research communities (Christen & Churches, 2002).

In this paper, we focus on the EL task in tracking entities that could be mentioned using different names and linking each of them to a unique database entry. In the following three subsections, we first introduce two tasks inspired the EL task of the Text Analysis Conference (TAC) Knowledge Base Population (KBP) (McNamee & Dang, 2009). A specific EL task, Gene Normalization (GN) (Hirschman et al., 2005) in the biomedical text mining is then described. Finally, an EL task based on the instance-based criterion is introduced (Dai, Chang, Tsai, & Hsu, 2011).

2.1 Tasks Inspired Entity Linking

(1) Link-The-Wiki Track in INEX

The INitiative for Evaluation of XML retrieval (INEX) workshop conducted the Link-The-Wiki track in 2007 (D. W. Huang, Xu, Trotman, & Geva, 2008). The goal of the track is to facilitate the wikification process (Rada Mihalcea & Csomai, 2007): when a user creates a new article in the Wikipedia, an automatic wikification system selects a number of prospective anchor texts (keyword extraction), and multiple link destinations for each anchor (link disambiguation) for him. The system also offers prospective updates to related links in other (e.g. older) wiki articles, which may point to a best entry point within this newly created article. Therefore, links on each article can always be up-to-date with the latest existing information within the wiki system. The link disambiguation step in the Link-The-Wiki track is closely related to the EL task discussed in this paper.

In 2007 the track examined article-wide linking in the Wikipedia; starting from 2008, the track extended to include the Anchor to Best Entry Point task (W. C. D. Huang, Geva, & Trotman, 2009). The mean average precision and the interpolated precision-recall were used in the track for evaluation. According to the Wikipedia guidelines, if the anchor text occurred several times in a document, only one instance is likely to be anchored. Therefore, an anchor in a Wikipedia page may be defined by a user in several slightly different ways. Based on the article-wide evaluation scheme, the track adapts a proximate metric by considering an anchor or entry point as relevant if it is no more than *n* characters away from a point chosen by an assessor (W. C. D. Huang et al., 2009).

(2) International Workshop on Semantic Evaluations

Senseval is the international organization devoted to the evaluation of semantic analysis systems. Beginning with the fourth workshop, SemEval-2007 (Artiles, Gonzalo, & Sekine, 2007) included semantic analysis tasks outside of word sense disambiguation (WSD). For instance, the Web People Search task (Artiles et al., 2007) focused on the disambiguation of person names in the scenario of the WWW search engines. The motivation of this task is to automatically cluster the web search results according to the "different" people sharing a given person name.

The Web People Search task can also be viewed as a co-inter-document co-reference resolution problem and its goal is similar to WSD. Artiles et al. (2007) pointed out the two main differences between the Web People Search task and WSD. The first difference is that in contrast to the subtle or even conflicting boundaries between word senses in a dictionary, the boundary between people name is more distinct. The second difference is that WSD usually operates with a relatively small dictionary containing a predefined set of senses. However, in the Web People Search task, the number of actual people is unknown a priori, and it is in average much higher than that in the WSD task.

2.2 Entity Linking Tasks

As described in the previous subsection, beyond NER, the EL task focuses on recognizing entities and linking each of them to a unique database entry. There are three principal challenges in such EL: the same entity can be referred to by more than one name string (the name variation issue), the same name string can refer to more than one entity (the ambiguity issue), and that many mentioned entities may not appear in a KB (NIL), even for large KBs (the absence issue) (Ji & Grishman, 2011). Interests in these problems have grown rapidly among different NLP communities. The following subsections introduce two main EL-related tasks, KBP-EL and BioCreative GN.

(1) Entity Linking in Knowledge Base Population

In 2009, the KBP task was introduced in TAC. The goal of the task is to promote research in discovering facts about entities and expanding a structured KB with this information. The KBP task considers the scenario that we may need to gather information, which is scattered among documents of a large collection, about a certain NE. This requires the ability to identify the relevant documents and to integrate facts, which may be redundant, complementary or in conflict, coming from these documents. The extracted information can then be used to augment an existing database.

In the KBP-EL task, given an entity query that consists of a name string and a background article containing that name string, the EL system is required to provide the KB entry to which the name refers; or NIL if there is no such KB entry. The

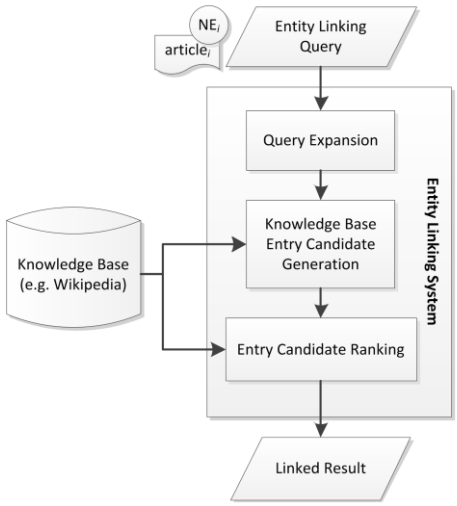


Figure 2: General entity linking system architecture for Knowledge Base Population.

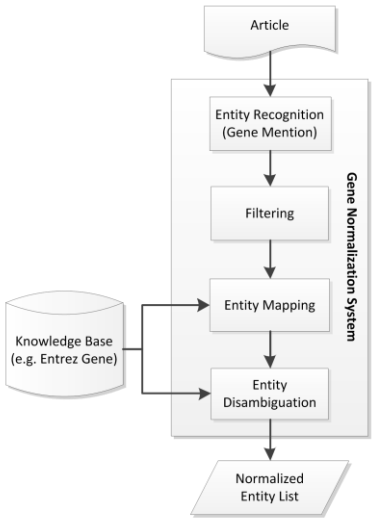


Figure 3: General gene normalization system architecture.

background article serves to disambiguate ambiguous name strings. Figure 2 depicts the architecture of general EL systems.

An EL query ($\langle NE_i, article_i \rangle$) is a request to the EL system for linking the textual entity mention NE_i in the given article $article_i$ to an entry in a KB. The architecture includes three steps: (1) query expansion: expand the query NE_i into a richer set of forms to improve the recall rate; (2) candidate generation: find all possible KB entries that the query might link to; (3) candidate ranking: rank the probabilities of all candidates and NIL answers. Ji and Grishman (2011) gave a detailed overview of the current state-of-the-art EL approaches and discussed the results of the evaluation.

(2) Gene Normalization Task in BioCreative Workshop

BioCreative is a community-wide effort that promotes the development and evaluation of text-mining and IE systems applied in the biomedical domain. As one of the largest public biomedical text-mining competitions in biomedical fields, BioCreative has conducted several challenges and has released standard evaluation datasets for different tasks. To spur development in regards to the name variation and the ambiguity issues, BioCreative has held several open competitions for the GN task (Hirschman et al., 2005; Leitner et al., 2010; Lu et al., 2011; Morgan et al., 2008), which evaluates the ability of automated systems to generate a list of unique gene identifiers from PubMed abstracts.

Krauthammer and Nenadic (2004) differentiated three main steps for the successful identification of entities from literature: entity recognition, entity classification, and entity mapping. Several GN systems subsumed the steps and employed a variety of approaches to address the GN task (Dai, Lai, & Tsai, 2010; Hakenberg, Plake, Leaman, Schroeder, & Gonzalez, 2008; Wermter, Tomanek, & Hahn, 2009). In general, after gene mention recognition (GMR), the current top-performing systems include three main steps as shown in Figure 3: (1) filtering: filter out false positives (FPs) or NILs, (2) entity mapping: generate candidate database identifiers and (3) entity disambiguation. Some studies only focused on improving one of these steps. For example, Hakenberg *et al.* (2008) employed an isolated stage to filter out FPs, including protein families, groups or complexes. Tsuruoka *et al.* (2007) utilized logistic regression to improve the accuracy of entity mapping. Xu *et al.* (2007) proposed a knowledge-based disambiguation approach that combines features from text and knowledge sources via an IR method. Crim *et al.* (2005) used the maximum entropy model to classify valid identifiers from candidate identifier lists. Dai *et al.* (2010) collected external knowledge for each gene, such as chromosome locations, gene ontology terms, etc., and calculated the likelihoods stating the similarity of the current text with the knowledge to improve the disambiguation performance. Wang *et al.* (2010) focused on one source of entity ambiguity, the model organism, and developed a corpus for organism disambiguation.

(3) KBP-EL vs. BioCreative-GN

Despite the details of its goal, linking entities to KB entries of both the GN task and the KBP-EL subtask is technically the same. We can observe one significant difference when we look at Figure 2 and Figure 3—the input of the two tasks. In the KBP-EL subtask, the input includes an entity and the article in which the entity is mentioned. The given article is used to help disambiguate the given entity. In contrast, the input of GN task contains an article alone. The GN system needs to recognize all entities mentioned in the given article, link all of them to their corresponding database IDs and then return a list of linked IDs.

The GN process shows an important step in the curation process for the model organism databases: once an article is selected for curation, a curator will list the genes or gene products of interest in the article (Hirschman et al., 2005). This is a time-consuming step, but also a key step in an accurate search of the biological literature. Both curators and researchers would benefit from GN systems to speed up the process of linking literature to biological databases (Dowell, McAndrews-Hill, Hill, Drabkin, & Blake, 2009). On the other hand, systems developed for the EL task in KBP task can link a given entity to an existing KB entry (or conclude that the entity is a novel entity that has not been recorded in KB) and extract information from the given article to populate an existing KB. Such linking process can reduce redundant and conflicting information recorded in KB and

bridges the natural language processing and data mining/database communities.

2.3 Instance-based Entity Linking

All of the above EL-related tasks and their evaluation considered the EL task from the article-wide perspective. For example, in the BioCreative-GN task, a system is required to provide a list of gene mentions that exist in an abstract with their corresponding database identifiers. However, such article-wide linked result is insufficient for the assembly of the interrelations described in an article because in some cases, the same names described in an article may possess different identities.

To this end, our previous work (Dai, Chang, et al., 2011) considered the EL task from the perspective of the instance level. In contrast to the two aforementioned article-wide methods, an instance-based EL approach must link all gene mentions in the text, and also precisely give their exact occurrence information. Such instance-based EL results are important because they allow following application to infer the interrelationships among those linked entities, but this task is also more challenging than the traditional EL tasks. The instance-based EL task requires deeper linguistic analysis and domain dependent knowledge to infer each instance's identity.

The next subsection provides an overview of the challenges found in the instance-based EL, and we will provide our suggestions to address these challenges in Section 3.

(1) Challenges of the Instance-based Entity Linking

When considering EL tasks from the instance level, the first challenge is the lack of a suitable corpus for developing instance-based EL systems. In linguistics, a corpus is a large and structured set of texts used to do statistical analysis. It plays an important role to help generate linguistic rules or patterns, learn these rules or decision criteria automatically, and evaluate the results obtained by comparing them with a gold standard. All of the above mentioned EL-related tasks, which have used either Wikipedia text or PubMed articles as text corpora, are not suited for the finer level of granularity that allows applications to associate the extracted relationships between entities with correct identities. As far as we know, only two pioneering works aggressively tried to list all mentions' identities and made their datasets open available. The first is Cucerzan's dataset (2007)[2], which is compiled from two different sources: Wikipedia (350 articles) and MSBC news stories (20 articles). However, Cucerzan's data neither include NILs, nor does it provide the exact occurrence of all entities. The second dataset is released by Kulkarni *et al.* (2009), which is sampled from online news[3]. Unfortunately, this data set is small (only 103 news articles containing 7,544 Wikipedia entry annotations; the others are NIL annotations). There is no similar corpus available within the biomedical domain. Therefore, we undertook to compile an instance-based gene mention linking (IGML) corpus using a semi-automatic approach (*cf.* Section 3.5).

The second challenge is the lack of context information for disambiguating each individual instance. The main research directions in the traditional EL approaches relied on domain knowledge derived from entries' profiles and contextual features extracted within a predefined content window. Rule-based (Dai et al., 2010; Hakenberg et al., 2008), vector space models (Cucerzan, 2007) and machine learning approaches (Crim et al., 2005; Rada Mihalcea & Csomai, 2007; Milne & Witten, 2008) have been proposed to disambiguate entity mention individually. However, in some cases, the context is obscure. For example, considering the sentence "The synthetic replicate of **urocortin** was found to bind with high affinity to type 1 and type 2 CRF receptors and, based upon its anatomic localization within the brain, was proposed to be a natural ligand for the type 2 CRF receptors.". The sentence alone does not explicitly provide any clues to help computer program to determine the identity of the gene mention "urocortin", which has at least 8 ambiguous Entrez Gene IDs. One approach is to expand the context window used for disambiguation to the paragraph level. However, the paragraph described in a biomedical article usually mixes several pieces of information in its description, which may not be directly related to a target entity instance and leads traditional EL approaches to fail. Only two recent works (Han, Sun, & Zhao, 2011; Rastogi, Dalvi, & Garofalakis, 2011) started to deal with this issue.

Several previous EL works (Cucerzan, 2007; Kulkarni et al., 2009; Rada Mihalcea & Csomai, 2007) assumed that the same surface name described in an article always refer to the same instance. This assumption might be true in encyclopedia-style articles, such as Wikipedia, but is not suitable for biomedical articles. Based on our analysis, the same surface name annotated with more than one linked KB entries only occupies 6% of the articles of Cucerzan's Wikipedia data[4]. However, in our IGML corpus, 14.9% of the articles contain entity mentions with the same surface name but linked with an average of 2.93 different IDs. This phenomenon makes the instance-based EL in biomedical literature even more challenging.

2.4 EL Problem Definition

This section gives formal definitions of all of the above mentioned EL tasks.

Definition 1: Instance-based Entity Linking Problem Let $\mathrm{M} = (m_1, m_2, \dots)$ denotes a sequence of entities mentioned in an article A. The surface name of m_i is denoted by $Name(m_i)$. The NE type of the entity 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 $\mathrm{ID}_i = \{id_1, id_2, \dots\}$, 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 $|(LinkTo(m_i): m_i \in \mathrm{M})| = |\mathrm{M}|$.

In instance-based gene mention linking (GML) or instance-based GN, only the entities, whose $EntityType(m_i)$ belong to "gene", are considered for evaluation. Both the GN task in BioCreative and the EL task in KPB can be subsumed into

[2] Available at http://research.microsoft.com/users/silviu/WebAssistant/TestData

[3] Available at http://www.cse.iitb.ac.in/~soumen/doc/CSAW/

[4] In average, those names are linked with 2.09 Wikipedia entries.

Definition 1. In BioCreative-GN, the developed system should satisfy the equation $|\{LinkTo(m_i): m_i \in M\}| \leq |M|$. We refer to BioCreative-GN as the article-wide EL problem.

Definition 2: Article-wide Entity Linking Problem Let $M = \{m_1, m_2, \dots\}$ denotes a set of entities mentioned in A. Given the entries $ID_i = \{id_1, id_2, \dots\}$ in a KB and the mapping function $LinkTo(m_i)$, the article-wide EL problem satisfies the constraint $|\{LinkTo(m_i): m_i \in M\}| \leq |M|$.

On the other hand, the KBP-EL task only considers one certain entity m_i mentioned in A. The thesis refers the KBP EL task as the article-wide "salient entity" linking problem because based on the Wikipedia style manual, only the salient entity and its related entities should be linked in wikification; too many links would obstruct readers to follow the article by drawing attention away from important links (Rada Mihalcea & Csomai, 2007).

Definition 3: Article-wide Salient Entity Linking Problem Let $M = m_1$ denotes a salient entity mentioned in A. Given the entry set $ID_i = \{id_1, id_2, \dots\}$ 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 KBP-EL subtask (belonging to the article-wide salient EL problem), the salient entity is given. But in instance-based GML or the BioCreative GN (belonging to the article-wide EL problem) tasks, the systems must also deal with the NER problem.

3. Instance-based Entity Linking: What Works

Our idea to deal with the challenge of the lack of context information for disambiguation of individual entity instance is to model dependencies among entities across sentences in the same paragraph. These dependencies are ignored by most of the previous article-wide EL approaches. We refer our approach to as the collective entity disambiguation, which is developed by considering the relational information hidden among entities. In the following subsections, we first introduce the collective classification. We then describe the main ideas of the proposed collective entity disambiguation approach and use Markov logic (Domingos & Lowd, 2009) to implement a joint inference model that can model interweaved constraint found in the instance-based EL problem.

3.1 Collective Classification

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 many classification tasks, instances can be related, and the interrelationship can be used to improve the classification performance. The second instance-based EL challenge described in the previous subsection is an example. In most of the individual EL formulation, an intrinsic local classifier is employed to assign a probability to the linked ID of an individual mention instance independently of the linked IDs of other instances. A drawback of local classifiers is that, when they decide the linking ID of a mention instance, they cannot use information about the linked IDs and features of other mentions in the same article. Furthermore, there are strong dependencies among the unknown IDs of the instances, which could be either a true positive entity mention or a FP. 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.

Definition 4 Given a network N, an 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.

A model that can classify a set of interlinked nodes or objects using all three types of information described above is referred to as the collective classification (Sen et al., 2008). The *relational classification* (Preisach & Schmidt-Thieme, 2008) is used to denote an approach that concentrates on classifying relational data by using only the first two types of correlations.

3.2 Collective Entity Disambiguation

The collective entity disambiguation approach is based on two main ideas: salience in centering theory (Grosz, Weinstein, & Joshi, 1995) and transitivity (Ng, 2005).

(1) Discourse Salience

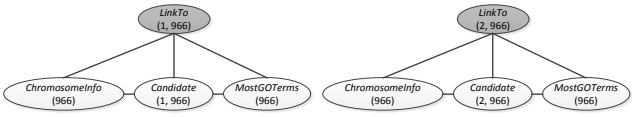
Discourse salience is a phenomenon that in a given discourse, there is precisely one entity that is the center of attention. Such entity is mentioned over and over again and makes it more salient than others. The collective disambiguation method utilizes this phenomenon to improve the instance-based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

(2) Transitivity

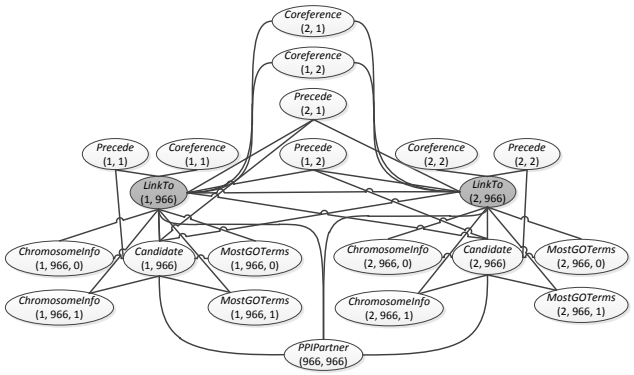
Similarly, the idea of *transitivity* allows us to express the concept that if two entities refer to the same concept, and one mention has been linked to a database entry, the other should also be linked to the same entry.

Using the above two collective features in entity disambiguation, the lack of context information can be smoothly resolved if one can model these dependencies in their model and keep these information when dealing with entities mentioned in different paragraphs. In the following section, we introduce how we model these dependencies by using a Markov logic network (MLN) and use the Markov logic[5] notations to express our ideas.

[5] Markov logic extends first-order logic by attaching weights to formulas. Please refer to Domingos & Lowd's work (2009) for details.



(a) Traditional individual entity disambiguation formulation developed for the article-wide EL problem.



(b) Collective entity disambiguation formulation.

Figure 4: Ground Markov network obtained by applying all of the three collective disambiguation formulae to the constants $x, y = \{1, 2\}$, $c = \{0, 1\}$, and $id = \{966\}$.

3.3 Collective Formulation of the Instance-based Gene Mention Linking Problem

The main difference between the proposed collective entity disambiguation and the individual disambiguation approach is the modeling of the dependencies among entities. In addition to the saliency and co-reference properties introduced in the previous subsection, we also model the correlation among entities. In GML, the correlation refers to the protein-protein interaction (PPI).

In our formulation, for all individual instances described in a paragraph, their orders are leveraged to build the dependencies including saliency, their PPI correlation and co-reference chains. For example, the saliency property is written as follows in Markov logic:

Formula 1: **Saliency** $Precede(x, y) \wedge LinkTo(x, id) \wedge Candidate(y, id) \Rightarrow LinkTo(y, id)$

In other words, 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. The formula is very similar to the transition feature of the linear-chain conditional random fields (CRF) (Lafferty, McCallum, & Pereira, 2001), which can be implemented in Markov logic as follows.

CRF transition feature: $Precede(x, y) \wedge Label(x, +L) \Rightarrow Label(y, +L)$

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

For the transitivity property, the model is required to infer whether or not the entities x and y are the same instances. The predicate $Coreference(x, y)$ is defined to capture the information. We can then define the formula

Formula 2: **Transitivity 1**

$Coreference(x, y) \wedge LinkTo(x, id_i) \Rightarrow LinkTo(y, id_i)$

to express the transitivity concept that if the xth and the yth gene mentions are co-reference pairs, then y should be also linked to id.

Finally, the PPI collective can be defined as the following formula, which captures the dependency 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:

Formula 3: **PPI** $LinkTo(x, id_i) \wedge Candidate(y, id_j) \wedge PPIPartner(id_i, id_j) \Rightarrow LinkTo(y, id_j)$

Figure 4 compares the ground Markov network (b) of our collective entity disambiguation with the traditional individual approach (a). In Figure 4 (a), the individual approach considers the likelihoods stating the similarity of the current context with domain knowledge of the recognized entity, including the chromosome location ($ChromosomeInfo(x, id, +c)$) and gene ontology ($MostGOTerms(x, id, +c)$) for individual disambiguation. Comparing Figure 4 (b) with (a), our collective entity disambiguation model captures the dependencies among entities in the same paragraph, allowing the information to be employed in the EL decision.

3.4 Joint Inference of the Entity Linking Stages

Biomedical researches nowadays are mainly focused on the human genome. However, humans are unsuitable for laboratory experiments. Therefore, mammalian model organisms that contain homologous human genes are frequently used in genomics studies. Consequently, the observed effects of the model organism's gene are at times inferred to the human genome, resulting in an aggregate of genes correlated with different species. Figure 5 is an example. This sentence is derived from a biomedical literature, and it describes the relationship of the homologous rat and human gene syntenin-1. This phenomenon leads to the third instance-based EL challenge described in the previous subsection.

> …Here, we demonstrate that rat **syntenin-1**, previously published as **syntenin-1** (**syntenin**), **mda-9**, or **TACIP18** in *human*, is a **neurofascin**-binding protein that exhibits a wide-spread tissue expression pattern with a relative maximum in brain. …

Figure 5: An example paragraph containing complex entity information.

Joint inference is a possible solution, because they make it possible for features and constraints to be shared among tasks and avoid error cascade and compound. It is important to employ the joint inference in the collective entity disambiguation model because all of the three collective properties described in the previous subsections could be trapped by the third instance-based EL challenge. For example, consider the transitivity property. A co-reference chain for the sentence in Figure 5 could be composed of {"(rat) syntenin-1", "syntenin-1", "syntenin", "mda-9", "TACIP18"}. However, the first gene is actually a rat gene, which must be removed from the chain. Directly employing the transitivity formulae (Formula 2) will lead to error cascades. One possible solution is to add additional constraints on the original formula:

Formula 4: **Transitivity 2**

$$Coreference(x,y) \wedge LinkTo(x,id_i) \wedge \neg\exists \boldsymbol{id_j}.\boldsymbol{LinkTo(y,id_j)} \Longrightarrow LinkTo(y,id_i)$$

The formula expresses that if the xth and the yth gene mentions are co-reference, and x is linked to id and y has not been linked, then y should be also linked to id. The formula captures the idea that we ask neighbors for help only if the context does not provide enough information for disambiguation.

Another solution is to add a filtering mechanism to remove the possible error edges in the collective model. For example, the following constraint can be defined to ensure that whenever x and y are a co-reference pair, they must be entities that should be linked.

Formula 5:

$$Coreference(x,y) \Longrightarrow SuitablyLink(x) \wedge SuitablyLink(y)$$

In this formula, the predicate $SuitablyLink(x)$ indicates that the xth gene mention of the article should be linked with a database entry. This constraint can remove FP candidates from the co-reference chain to reduce the cascade of errors.

The same concept can also be applied on EL. For example, the following formula ensures that, whenever the xth entity is linked to a database entry id, it must be an entity suitable for linking.

Formula 6: $LinkTo(x,id) \Longrightarrow SuitablyLink(x)$

The concept of the formula is that the database entry id does not have to be linked to the entity x proposed by the entity recognition stage; however, the id cannot be assigned to the xth gene mention that has not been proposed as a potential entity.

3.5 Instance-based Gene Mention Linking Corpus

As described in the previous section, we have compiled an instance-based gene mention linking (IGML) corpus. Table 1 shows the annotated corpus statistics. Note that in Table 1, the last row (# of IDs per mention) shows that, in some cases, a gene mention is annotated with more than one Entrez Gene ID by our annotators. This occasions occurs when the preceding context of the gene mention suggests that it is describing more than one gene at a time (e.g. …it is found that *human and rat* **syntenin-1**…, …the *mammalian* **syntenin-1**…)

In contrast to most of the previous EL works, which separated the entity recognition and the entity disambiguation stages in their evaluation, we have considered evaluating the combined results of the recognition and disambiguation stages to investigate their interrelationship by using an instance-based evaluation metrics (Dai, Tsai, et al., 2011). This work follows the same procedure in constructing an evaluation corpus to evaluate the collective entity disambiguation approach: we employed a gene mention tagger on the IGML corpus to recognize all gene mention candidates, which may include FPs. The employed gene mention tagger achieved an F-score of 85.8% on the BioCreative II gene mention tagging corpus (Smith et al., 2008). However, it only achieves an F-score (F) of 69.01% with 58.31% precision (P) and 84.51% recall (R) on the IGML corpus when the exact full matching boundary criterion is used.

Table 1: IGML corpus statistics.

Dataset	Training Set	Test Set
Numbers of articles	282	262
Numbers of gene mentions	2,813	3,143
Numbers of linked Entrez Gene IDs	2,861	3,187
Numbers of words per article	215.86	228.91
Numbers of mentions per article	10.01	12.00
Numbers of words per mention	1.52	1.35
Numbers of IDs per mention	1.02	1.01

The exact/partial entity mapping approach was then employed to generate candidate IDs for each entity (Dai et al., 2010). The official lexicon released by the BioCreative II GN task was used in our mapping procedure, which contains 32,975 Entrez Gene entries and their possible gene/protein names. In the lexicon, on average, each ID has 5.55 synonyms, and each synonym has 1.12 IDs. Compared with Wikipedia, the dataset used by (Rada Mihalcea & Csomai, 2007), in which each entry has 3.21 synonyms. The GN lexicon lists more synonyms, which can increase coverage but hurt precision by increasing ambiguity.

For each mention *m* in a sentence *s* recognized by the tagger and the set of Entrez Gene ID candidates for *m* mapped by the entity mapping stage, a procedure searched *s* for the first human annotated mention *n* overlapping with *m* and set *n*'s human annotated ID as *m*'s true Entrez Gene ID. Other candidates were set as *m*'s incorrect IDs. The compiled corpus can then be used as a training dataset for a binary classifier based on the individual entity disambiguation approach as shown in Figure 4 (a) or the collective disambiguation approach for the instance-based EL problem. Table 2 shows the properties of the final generated corpus. In this table, the best/worst GML performance is obtained by linking the ambiguous entity mention with correct/wrong linking answer every time if the gold linked ID is listed as one of the entity's candidate linking identifier. The optimal GML performance is the best linking performance without considering whether or not the linking IDs are listed as candidates.

Table 2: Properties of the gene mention linking corpus after employing the entity recognition and the entity mapping. The following results are only focused on human genes.

Dataset (P/R/F (%))	Training Set	Test Set
GMR performance	55.3/83.4/66.5	66.2/82.7/65.1
Best GML performance	83.0/56.1/67.0	83.0/66.0/73.5
Worst GML performance	71.1/48.1/57.4	70.8/56.3/62.7
Optimal GML performance	83.3/57.8/68.2	83.2/66.7/74.1
# of correct recognized genes	2,101	2,398
# of recognized mentions per article	13.51	17.09
# of words per mention	1.84	1.56
# IDs per mention	1.52	1.50

For the filtering corpus, again, for each mention *m* in a sentence *s* recognized by the gene mention tagger, the procedure checked whether or not the boundaries of the mention *m* matched with the IGML's human annotated boundaries. All matched mentions are regarded as true positives while the others are true negatives. Note that the employed gene mention tagger only achieved 55%~65% precision, indicating that around half of the recognized mentions are FPs.

Table 3: The collective disambiguation results using instance-based criterion.

	Training set (%)			Test set (%)		
Configuration	P	R	F	P	R	F
No disambiguation (P-oriented)	**80.4**	48.6	**60.6**	**80.7**	56.3	**66.3**
No disambiguation (R-oriented)	64.7	**56.3**	60.2	66.3	**66.0**	66.2
Random baseline	68.4	51.6	58.8	68.3	59.8	63.8
Saliency discourse	79.2	50.2	61.5	79.5	59.0	67.7
Protein-protein interaction	**79.4**	**51.1**	**62.2**	**80.1**	**59.8**	**68.5**
Transitivity	78.5	49.5	60.7	78.6	58.8	67.2
All collective formulae	79.1	52.0	62.8	78.4	61.0	68.6
All collective formulae + Filtering	**79.3**	52.0	**62.9**	**78.8**	61.0	**68.8**
All individual formulae	74.9	**54.3**	**62.9**	75.7	**61.7**	68.0
Collective + individual	74.5	**55.7**	**63.7**	74.9	64.8	69.5
Collective + individual + Filtering	**79.9**	54.9	**65.1**	**77.8**	**65.3**	**71.0**

Finally, to generate the co-reference resolution corpus, we treated gene mentions generated by the tagger containing the corresponding same gold linked ID as co-references. All of the following experiments are then conducted on the constructed corpus and evaluated on the instance-based criteria. As shown in Table, the upper bound PRF-scores of the following experiment is 83.3/57.8/68.2 and 83.2/66.7/74.1 on training and test set respectively.

(1) Collective Disambiguation Performance

Table 3 shows the performance of the collective disambiguation method. The experiment first conducted ten-fold cross validation on the training set of the compiled corpus to evaluate the performance of the three collective disambiguation formulae, the discourse salience (*cf.* Formula 1), the transitivity (*cf.* Formula 4), and the PPI (*cf.* Formula 3). The entire training set was then used to train a MLN model and evaluate its performance on the test set.

The first three rows show the baseline results. The first two rows are the performance without applying any disambiguation approaches; for which all mentions with only one candidate ID were directly treated as answers, and entities with more than one candidate ID were discarded (to optimize precision; P-oriented) or kept (for maximal recall; R-oriented). For each candidate gene mention, the third baseline "Random baseline" randomly selects one of the candidate mention's possible candidate IDs as the linked ID.

As shown in Table 3, by adding the salient discourse property of the centering theory and the transitivity property of co-reference resolutions without any domain knowledge, the recall rate is improved and results in an improved F-score. Furthermore, the PPI collective combining the domain knowledge achieves the highest PRF-scores, even outperforming all individual rules on the test set. However, we observed that if the constraint in our transitivity formula was removed (using Formula 2 instead of Formula 4), the precision rate is improved to 79.7% but the recall is dropped to 56.4%; F-score (66.1%) is lower than No disambiguation baselines. This result shows the effect of the third challenge, in which a single surface name can refer to different instances in biomedical literature.

Table 3 also shows the performance by adding all of the collective formulae and Formula 5 (All collective formulae + Filtering), and Formula 6 (Collective+Individual+Filtering). The results show that joint inference of the filtering and co-reference can improve the precision and result in an improved F-score.

4. Conclusions

In this paper, we introduced an overview of the EL tasks in two different research domains with different objectives and gave formal definitions for those tasks, including instance-based EL, article-wide EL and article-wide salient EL. We analyzed the reasons which have made instance-based EL a more challenging task, discussed our observations, and suggested possible solutions to address these challenges of biomedical text mining. We believe that these results would be helpful for current and new EL researchers, and could facilitate the progress of IE and QA researches in general.

References

[Aronson 2001] Aronson, A., Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. JOURNAL OF BIOMEDICAL INFORMATICS, 35, 17-21, 2001.

[Artiles 2007] Artiles, J., Gonzalo, J., & Sekine, S., The SemEval-2007 WePS evaluation: establishing a benchmark for the web people search task. Paper presented at the Proceedings of the 4th International Workshop on Semantic Evaluations, Prague, Czech Republic, 2007

[Bhattacharya 2007] Bhattacharya, I., & Getoor, L., Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 5, 2007

[Bilenko 2005] Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S., Adaptive name matching in information integration. Intelligent Systems, IEEE, 18(5), 16-23, 2005

[Christen 2002] Christen, P., & Churches, T., Febrl-Freely extensible biomedical record linkage: Joint computer science technical report series, 2002

[Chu-Carroll 2007] Chu-Carroll, J., & Prager, J., An experimental study of the impact of information extraction accuracy on semantic search performance. Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal, 2007

[Chu-Carroll 2006] Chu-Carroll, J., Prager, J., Czuba, K., Ferrucci, D., & Duboue, P., Semantic search via XML fragments: a high-precision approach to IR. Paper presented at the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA, 2006

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

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

[Culotta 2005] Culotta, A., & McCallum, A. Joint deduplication of multiple record types in relational data. Paper presented at the Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM'05), New York, NY, USA, 2007

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

[Dai 2010] Dai, H.-J., Lai, P.-T., & Tsai, R. T.-H., 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, 2010

[Dai 2011] Dai, H.-J., Tsai, R. T.-H., & Hsu[*], W.-L., Entity Disambiguation Using a Markov-Logic Network. Paper presented at the Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), Chiang Mai, Thailand, 2011

[Domingos 2009] Domingos, P., & Lowd, D., Markov Logic: An Interface Layer for Artificial Intelligence: Morgan and Claypool Publishers, 2009

[Dowell 2009] Dowell, K. G., McAndrews-Hill, M. S., Hill, D. P., Drabkin, H. J., & Blake, J. A., Integrating text mining into the MGI biocuration workflow. Database (Oxford), 2009, bap019. doi: 10.1093/database/bap019, 2009

[Dredze 2010] Dredze, M., McNamee, P., Rao, D., Gerber, A., & Finin, T., Entity Disambiguation for Knowledge Base Population. Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, 2010

[Elmagarmid 2007] Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S., Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering, 19(1), 1-16. doi: 10.1109/tkde.2007.9, 2007

[Grishman 1996] Grishman, R., & Sundheim, B., Message Understanding Conference-6: a brief history. Paper presented at the Proceedings of the 16th conference on Computational linguistics - Volume 1, Copenhagen, Denmark, 1996

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

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

[Han 2011] Han, X., Sun, L., & Zhao, J., Collective entity linking in web text: a graph-based method. Paper presented at the Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, Beijing, China, 2011

[Hirschman 2005] Hirschman, L., Colosimo, M., Morgan, A., & Yeh, A., Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics, 6(Suppl 1), S11, 2005

[Huang 2008] Huang, D. W., Xu, Y., Trotman, A., & Geva, S., Overview of INEX 2007 Link the Wiki Track. In F. Norbert, K. Jaap, L. Mounia & T. Andrew (Eds.), Focused Access to XML Documents (pp. 373-387): Springer-Verlag, 2008

[Huang 2009] Huang, W. C. D., Geva, S., & Trotman, A., Overview of the INEX 2008 Link the Wiki Track. In S. Geva, J. Kamps & A. Trotman (Eds.), Advances in Focused Retrieval (Vol. 5631, pp. 314-325): Springer Berlin / Heidelberg, 2009

[Ji 2011] Ji, H., & Grishman, R., Knowledge base population: Successful approaches and challenges. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 2011

[Jin-Dong 2004] Jin-Dong, K., Tomoko, O., Yoshimasa Tsuruoka, Y. T., & Collier, N., Introduction to the bio-entity recognition task at JNLPBA. Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04), 70-75, 2004

[Khalid 2008] Khalid, M. A., Jijkoun, V., & Rijke, M. d., The impact of named entity normalization on information retrieval for question answering. Paper presented at the Proceedings of the IR research, 30th European conference on Advances in information retrieval (ECIR'08), 2008

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

[Kulkarni 2009] Kulkarni, S., Singh, A., Ramakrishnan, G., & Chakrabarti, S., Collective annotation of wikipedia entities in web text. Paper presented at the Proceeding of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) Paris, France, 2009

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

[Lawrence 1999] Lawrence, S., Giles, C. L., & Bollacker, K. D., Autonomous citation matching. Paper presented at the Proceedings of the third annual conference on Autonomous Agents, New York, NY, USA, 1999

[Leitner 2010] Leitner, F., Mardis, S. A., Krallinger, M., Cesareni, G., Hirschman, L. A., & Valencia, A., An Overview of BioCreative II.5. IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 7(3), 385-399, 2010

[Lu 2011] Lu, Z., Kao, H.-Y., Wei, C.-H., Huang, M., Liu, J., Hsu, C.-J. K. C.-N., . . . Wilbur, W. J., The gene normalization task in BioCreative III. BMC Bioinformatics, 12(Suppl 9), S2, 2011

[McNamee 2009] McNamee, P., & Dang, H. T., Overview of the TAC 2009 Knowledge Base Population Track. Paper presented at the Proceedings of the Second Text Analysis Conference (TAC 2009), Gaithersburg, Maryland USA, 2009

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

[Mihalcea 2007] Mihalcea, R., & Csomai, A., Wikify!: linking documents to encyclopedic knowledge. Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal, 2007

[Mihalcea 2000] Mihalcea, R., & Moldovan, D., Semantic indexing using WordNet senses, 2000

[Mihalcea 2001] Mihalcea, R., & Moldovan, D., Document indexing using named entities. Studies in Informatics and Control, 10(1), 21-28, 2001

[Milne 2008] Milne, D., & Witten, I. H., Learning to link with wikipedia. Paper presented at the Proceedings of the 17th ACM conference on Information and knowledge management, Napa Valley, California, USA, 2008

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

[Ng 2005] Ng, V., Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. Paper presented at the Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), University of Michigan, USA, 2005

[Pasula 2003] Pasula, H., Marthi, B., Milch, B., Russell, S., & Shpitser, I., Identity uncertainty and citation matching. ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 1425-1432, 2003

[Preisach 2008] Preisach, C., & Schmidt-Thieme, L., Ensembles of relational classifiers. Knowledge and Information Systems, 14(3), 249-272. doi: 10.1007/s10115-007-0093-3, 2008

[Rastogi 2011] Rastogi, V., Dalvi, A. N., & Garofalakis, A. M., Large-scale collective entity matching. Proceedings of the VLDB Endowment, 4(4), 208-218, 2011

[Rebholz-Schuhmann 2010] Rebholz-Schuhmann, D., Yepes, A. J. J., Mulligen, E. M. v., Kang, N., Kors, J., Milward, D., . . . Hahn, U., CALBC silver standard corpus. Paper presented at the Proceedings of the 3rd International Symposium on Languages in Biology and Medicine, Jeju Island, South Korea, 2010

[Sang 2002] Sang, E. F. T. K., Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. Paper presented at the proceedings of the 6th conference on Natural language learning - Volume 20, 2002

[Sarawagi 2002] Sarawagi, S., & Bhamidipaty, A., Interactive deduplication using active learning. Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02), Edmonton, Alberta, Canada, 2002

[Sarmento 2009] Sarmento, L., Kehlenbeck, A., Oliveira, E., & Ungar, L., An Approach to Web-Scale Named-Entity Disambiguation. Paper presented at the Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition, Leipzig, Germany, 2009

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

[Smith 2008] Smith, L., Tanabe, L. K., Ando, R. J. n., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., . . . Wilbur, W. J., Overview of BioCreative II gene mention recognition. Genome Biology, 9(Suppl 2), S2, 2008

[Sorden 1999] Sorden, N. N., Chang, H. F., & Nelson, S. J., Automated Indexing of Gene Symbols. Proceedings of the American Medical Informatics Association Symposium (AMIA '99), Washington, DC, 1999

[Tsuruoka 2007] Tsuruoka, Y., McNaught, J., Tsujii, J., & Ananiadou, S., Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics, 23(20), 2768-2774, 2007

[Wang 2010] Wang, X., Tsujii, J. i., & Ananiadou, S., Disambiguating the species of biomedical named entities using natural language parsers. Bioinformatics, 26(5), 661-667, 2010

[Wermter 2009] Wermter, J., Tomanek, K., & Hahn, U., High-performance gene name normalization with GENO. Bioinformatics, 25(6), 815-821. doi: 10.1093/bioinformatics/btp071, 2009

[Winkler 1999] Winkler, W. E., The state of record linkage and current research problems, 1999

[Woods 1997] Woods, W. A., Conceptual Indexing: A Better Way to Organize Knowledge Technical Report SMLI TR-97-61. Mountain View, CA, USA: Sun Microsystems, Inc., 1997

[Xu 2007] Xu, H., Fan, J.-W., Hripcsak, G., Mendonça, E. A., Markatou, M., & Friedman, C., Gene symbol disambiguation using knowledge-based profiles. Bioinformatics, 23(8), 1015-1022. doi: 10.1093/bioinformatics/btm056, 2007