4 Experiments

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

4.1 Knowledge Base

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

4.2 Data Sets

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

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

4.3 Evaluation Criteria

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

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

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

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

4.4 Baselines

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

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

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

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

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

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

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

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