Claims

1. A method for disambiguating a plurality of entity mentions in a document, the method comprising the following steps:

(a) parsing text of a document to obtain a plurality of entity mentions;

(b) determining one or more matched entity mention by matching each of the plurality of entity mentions with each entity listed in a database, wherein the database comprises verified entity relationship data indicating links between different entities, and the entity mention of the plurality of entity mentions having exact match with an entity listed in the database is determined as a matched entity mention;

(c) determining a score for each matched entity mention based on entity type and number of links the matched entity mention has with other matched entity mentions; and

(d) determining the matched entity mention with highest score as a hub entity, wherein the hub entity is deemed as a disambiguated entity mention,

wherein the score for each matched entity mention is determined based on entity type and number of links the matched entity mention has with other matched entity mentions by:

(e) determining, for each matched entity mention, one or more entities that are indicated as a first predetermined number of hops away from the corresponding matched entity in the verified entity relationship data;

(f) obtaining a weight value for each of the one or more entities, wherein the weight value was predetermined based on type of entity;

(g) if there are more than one entities determined for the matched entity mention, calculating for the matched entity mention, a weighted sum of the weight values of the more than one entities;

(h) if there are more than one matched entity mentions, identifying the matched entity mention that has the calculated weighted sum satisfying a first threshold as a hub entity candidate;

(i) determining a score for each hub entity candidate by

counting number of links between said hub entity candidate and each matched entity mention found in the verified entity relationship data, wherein the links to be counted are the links with a second predetermined number of hops between said hub entity candidate and each matched entity mention, or

making reference to the verified entity relationship data to count number of entities connected to each hub entity candidate within a second predetermined number of hops; and

(j) identifying the hub entity candidate with highest score as the hub entity.

2. The method of claim 1, wherein the method further comprises:

(a) identifying whether any matched entity mention is a large entity, wherein a matched entity mention is identified as a large entity if the number of links the matched entity has with other matched entity mentions is above or equal to a first threshold; and

(b) creating a first entity relationship map for the document using the hub entity and the one or more matched entity mention linked to the hub entity that is not the large entity.

3. The method of claim 2, wherein the method further comprises creating a second entity relationship map using the entities in the first entity relationship map, and entities that are one hop away from the entities in the first entity relationship map.

4. The method of claim 3, wherein the method further comprises:

(a) obtaining a predetermined number of most frequently found entity mentions from the parsed text of the document and fuzzy match each frequently found entity mentions against each entity listed in the database to return a list of disambiguated entity mention candidates;

(b) determining top candidates with matching similarity scores from the list of disambiguated entity mention candidates that are above or equal to a second threshold;

(c) creating, for each top candidate, a third entity relationship map using said top candidate and one or more entities one hop away from said top candidate;

(d) determining, for each top candidate, an expansion factor defining the extent of expansion in size of the second entity relationship map when the third entity relationship map is combined with the second entity relationship map; and

(e) determining the top candidate with expansion factor within an acceptable predetermined range of values as a disambiguated entity mention.

5. The method of claim 4, wherein the method further comprises:

updating the first entity relationship map by adding the disambiguated entity mention, which is the top candidate with expansion factor within the acceptable predetermined range of values, to the first entity relationship map; and

updating the second entity relationship map by adding one or more entity that is one hop away from the disambiguated entity mention in the verified entity relationship data, which is the top candidate with expansion factor within the acceptable predetermined range of values, to the second entity relationship map.

6. The method of any one of claim 4 or 5, wherein the second threshold is calculated by:

ranking all the disambiguated entity mention candidates according to matching similarity scores;

calculating a difference in matching similarity scores for each pair of ranked candidates consecutively, according to a formula, $D_n = C_n – C_{n+1}$, wherein

D_n = difference between candidate with *nth* rank and candidate with one rank below the *nth* rank,

C_n = candidate with *nth* rank,

C_{n+1} = candidate with one rank below the candidate with *nth* rank

n = 1, 2, and so on, wherein n = rank of candidate; and

setting, as the second threshold, the higher value of the matching similarity score of the pair of candidates with the largest calculated difference in matching similarity scores.

7. The method according to claim 6, wherein calculating of the difference in matching similarity scores for each pair of ranked candidates consecutively is performed only for ranked candidates with a matching similarity score above or equal to a third threshold.

8. The method according to any one of claims 1 to 7, wherein the first predetermined number of hops is one hop.

9. The method according to any one of claims 1 to 8, wherein the second predetermined number of hops is one hop.

10. The method of any one of claims 1 to 9, wherein the weight value was predetermined based on the types of entities as follows:

Employee of an entity;
Executive director of the entity;
Shareholder of the entity; and
Subsidiary entity of the entity;
where the Executive director has the greatest weight value, the Subsidiary entity has the least weight value, and the Employee and Shareholder have the same weight value or weight values between the greatest weight value and the least weight value.

11. The method of any one of claims 5 to 10, wherein a fuzzy full text search is performed against the document using each disambiguated entity mention and each of the one or more entities that is one hop away from the disambiguated entity mention in the updated second entity relationship map, wherein a fuzzy matched string in the document found via the fuzzy full text search with a predetermined acceptable matching similarity score that is not listed as a disambiguated entity mention in the updated second entity relationship map is updated as a disambiguated entity mention.

12. The method of any one of the claims 4 to 11, wherein the expansion factor of each top candidate is calculated based on a size of the third entity relationship map, a size of the second entity relationship map, an intersection size and a union size,

wherein the size of the third entity relationship map is determined by the number of disambiguated entities in the third entity relationship map,

wherein the size of the second entity relationship map is determined by the number of entities in the second entity relationship map,

wherein the intersection size is determined by the number of entities that are present in both the second entity relationship map and the third entity relationship map,

wherein the union size is obtained by:

adding the number of entities that are present in the second entity relationship map and the number of entities that are present in the third entity relationship map, and subtracting

the number of entities that are present in both the second entity relationship map and the third entity relationship map.

13. The method of any one of the preceding claims, wherein the method comprising a step of categorizing the plurality of entity mentions to filter entity mentions that are categorized as junk mention, wherein a junk mention is a predetermined invalid entity deemed to be not a corporate entity or an individual entity.

14. An apparatus for disambiguating a plurality of entity mentions in a document, wherein the apparatus comprises:

a processor for executing instructions in a memory to operate the apparatus to:

(a) parse text of a document to obtain a plurality of entity mentions;

(b) determine one or more matched entity mention by matching each of the plurality of entity mentions with each entity listed in a database, wherein the database comprises verified entity relationship data indicating links between different entities, and the entity mention of the plurality of entity mentions having an exact match with an entity listed in the database is determined as a matched entity mention;

(c) determine a score for each matched entity mention based on entity type and number of links the matched entity mention has with other matched entity mentions; and

(d) determine the matched entity mention with highest score as a hub entity, wherein the hub entity is deemed as a disambiguated entity mention,

wherein the apparatus is operable to:

(e) determine, for each matched entity mention, one or more entities that are indicated as a first predetermined number of hops away from the corresponding matched entity in the verified entity relationship data;

(f) obtain a weight value for each of the one or more entities, wherein the weight value was predetermined based on type of entity;

(g) if there are more than one entities determined for the matched entity mention, calculate for the matched entity mention, a weighted sum of the weight values of the more than one entities;

(h) if there are more than one matched entity mentions, identify the matched entity mention that has the calculated weighted sum satisfying a first threshold as a hub entity candidate;

(i) determine a score for each hub entity candidate by

counting number of links between said hub entity candidate and each matched entity mention found in the verified entity relationship data, wherein the links to be counted are the links with a second predetermined number of hops between said hub entity candidate and each matched entity mention, or

making reference to the verified entity relationship data to count number of entities connected to each hub entity candidate within a second predetermined number of hops; and

(j) identify the hub entity candidate with highest score as the hub entity.

15. A system for disambiguating a plurality of entity mentions in a document, wherein the system comprises the apparatus as claimed in claim 14.