(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 December 2015 (23.12.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/194934 A2

(51) International Patent Classification:
G06F 17/27 (2006.01)

(21) International Application Number:
PCT/MY2015/050043

(22) International Filing Date:
29 May 2015 (29.05.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
PI2014701649 18 June 2014 (18.06.2014) MY

(71) Applicant: **MIMOS BERHAD** [MY/MY]; Technology Park Malaysia, Kuala Lumpur, 57000 (MY).

(72) Inventors: **CHU, Benjamin Min Xian**; C/o Mimos Berhad, Technology Park Malaysia, Kuala Lumpur, 57000 (MY). **QIANG, Simon Liu**; C/o Mimos Berhad, Technology Park Malaysia, Kuala Lumpur, 57000 (MY). **LUKOSE, Dickson**; C/0 Mimos Berhad, Technology Park Malaysia, Kuala Lumpur, 57000 (MY).

(74) Agent: **H A RASHID, Ahmad Fadzlee**; Ipvolusi Sdn Bhd, A-3-3A, Centrio Pantai Hillpark, No. 1, Jalan Pantai Murni, Kuala Lumpur, 59200 (MY).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: A SYSTEM AND METHOD FOR ENTITY RESOLUTION

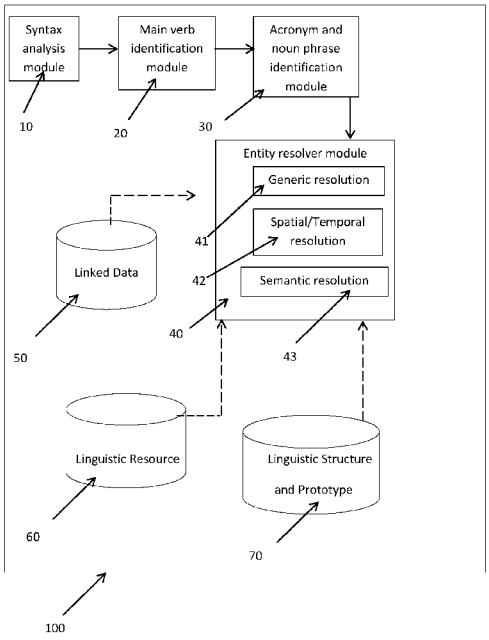


FIG. 1

(57) Abstract: The present invention relates to a system and method of entity resolution. The system (100) resolves an entity in a sentence by utilizing a generic resolution, a spatial and temporal resolution, and semantic resolution. The system (100) comprises of a syntax analysis module (10); a main verb identification module (20); an acronym and noun phrase identification module (30); an entity resolver module (40); a Linked Data Source (50); a Linguistic Resource module (60); and a Linguistic Structure and Prototype Repository (70).

WO 2015/194934 A2

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *of inventorship (Rule 4.17(iv))*

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

A SYSTEM AND METHOD FOR ENTITY RESOLUTION

FIELD OF INVENTION

The present invention relates to a system and method for entity resolution. More particularly, the present invention relates to a system and method for resolving entity by utilizing a generic resolution, a spatial and temporal resolution, and semantic resolution.

BACKGROUND OF THE INVENTION

An entity in natural language often display ambiguity as an expression can be interpreted or described with more than one meaning. For human, we may recognize the entity described by having a little background knowledge regarding the entity. On the other hand, this might not happen when dealing with computers or systems. In an entity resolution, natural language sentences are typed in by users and resolved against the known identities. However, the input entity may refer to multiple individual identities which may lead to undesired results. The main challenges are the similarity functions for comparing values and similarity thresholds need to be specified. Moreover, inaccurate entity disambiguation for polysemous entity may occur when performing without external background knowledge.

In regard to this, US Patent Publication No. US 2010/0161566 discloses techniques for adding entities to a group of entity resolution candidates by selecting entities that have a minimum threshold of similarity to a candidate, and allowing a greater number of resolutions in an entity resolution system. To resolve an incoming identity record, an initial group of candidates may be selected from known entities by identifying entities that match a candidate building attribute of the incoming identity record. Additional candidates may be selected by identifying entities with some information that is similar to one of the candidate entities. However, such system only creates new data or conjoins with the existing data in the database. It only depends on the previously stored data in the system. Thus, the entity resolution is limited to the entity in its storage.

Therefore, there is a need for a system and method of an entity resolution that addresses the drawbacks of the existing entity resolution system.

SUMMARY OF INVENTION

The present invention relates to a system and method of entity resolution. The system **(100)** resolves an entity in a sentence by utilizing a generic resolution, a spatial and temporal resolution, and semantic resolution. The system **(100)** comprises of a syntax analysis module **(10)**; a main verb identification module **(20)**; an acronym and noun phrase identification module **(30)**; an entity resolver module **(40)**; a Linked Data Source **(50)**; a Linguistic Resource module **(60)**; and a Linguistic Structure and Prototype Repository **(70)**.

Preferably, the entity resolver module **(40)** includes a generic resolution module **(41)** to perform the generic entity resolution on each acronym and each noun phrase in the acronym and noun phrase list by utilizing the Linked Data Source **(50)**; a spatial and temporal module **(42)** to perform the spatial and temporal entity resolution by extracting the spatial and temporal information from an animate and an inanimate noun phrase of the acronym and noun phrase list from the Linguistic Resource module **(60)**; and a semantic resolution module **(43)** to perform the semantic entity resolution by comparing semantic structures to a linguistic structure and prototype retrieved from the Linguistic Structure and Prototype Repository **(70)**.

The method for resolving entity automatically is characterised by the steps of performing syntax analysis to a sentence and transforming the sentence into a syntax structure; identifying a main verb of the syntax structure; identifying an acronym and a noun phrase of the syntax structure; performing entity recognition to identify the noun phrases of the sentence; and performing a generic resolution, a spatial and temporal resolution and a semantic resolution.

Preferably, the step of performing generic resolution includes choosing a candidate with the highest total accumulated match if the acronym has been processed; proceeding to the spatial and temporal resolution step after the candidate the highest total accumulated match is chosen; selecting an acronym from the acronym and noun phrase list if the acronym has not been processed; retrieving relevant candidates for the acronym from the Linked Data Source **(50)**; retrieving a semantic graph for each candidate from the Linked Data Source **(50)** if the candidates has not been processed; selecting a noun phrase from the acronym and noun phrase list if the noun phrase has been processed; retrieving a graph for the

noun phrase from the Linked Data Source **(50)**; matching the semantic graph of each candidate with semantic graph of the noun phrase; and choosing the semantic graph of the candidate that has the highest total match with the graph of the noun phrase.

Preferably, the step of performing the spatial and temporal resolution includes retrieving all noun phrases of animate, inanimate and event type from Linguistic Resource module **(60)** and acronym and noun phrase list; determining whether each animate noun phrases have been processed based on the spatial and temporal resolution step; and proceeding to the semantic resolution step if the animate noun phrase has been processed; selecting the retrieved noun phrase of an animate type if the animate noun phrase noun has not been processed; selecting a noun phrase of an inanimate or event type; retrieving semantic candidates that match the inanimate or event type of the noun phrase from the Linked Data Source **(50)**; retrieving the semantic graph for the noun phrase, G_{NP} from the Linked Data Source **(50)**; selecting the candidate with the least distance or with the most relevancy weight if the semantic candidates have been processed; retrieving the graph for each of the candidates, G_c if the semantic candidates have not been processed; extracting spatial and temporal information of G_{NP} and G_c from the Linked Data Source **(50)**; determining the distance between G_{NP} and G_c locations if G_{NP} and G_c possess spatial information; and analysing the temporal information between G_{NP} and G_c and assigning a relevancy weight if G_{NP} and G_c don't possess spatial information.

Preferably, the step of performing the semantic resolution includes identifying and retrieving a main verb of a sentence from a main verb list; transforming the sentence into a semantic structure, G_s; retrieving all linguistic structures of the identified main verb from Linguistic Structure and Prototype Repository **(70)**; determining the semantic graph similarity between G_s and the retrieved linguistic structures; selecting the linguistic structure that matches the best to G_s; determining the concept type of the agent of the main verb; retrieving all prototypes which contain the subject with the same concept type from the Linguistic Structure and Prototype Repository **(70)**; determining the semantic graph similarity between G_s and all retrieved prototypes; selecting prototype that matches the best with G_s; extracting the matched entity from the prototype; and resolving entities based on the results of semantic resolution step.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and constitute a part of the specification, illustrate embodiments of the invention and, together with the description, serve to explain the principles of the invention.

FIG. 1 illustrates a system for entity resolution **(100)** according to an embodiment of the present invention.

FIG. 2 illustrates a flowchart of a method for entity resolution according to an embodiment of the present invention.

FIG. 3 illustrates a flowchart of substeps for generic entity resolution step of the method of **FIG. 2**.

FIG.4 illustrates an exemplary semantic graph of possible candidates for an acronym based on the method of **FIG. 2**.

FIG.5 illustrates a first exemplary semantic graph for a candidate retrieved from a Linked Data Source **(50)** of the system **(100)** of **FIG. 1**.

FIG.6 illustrates a second exemplary semantic graph for a candidate retrieved from a Linked Data Source **(50)** of the system **(100)** of **FIG. 1**.

FIG.7 illustrates an exemplary semantic graph for a noun phrase.

FIG. 8 illustrates a flowchart of substeps for spatial and temporal entity resolution step of the method of **FIG. 2**.

FIG. 9 illustrates an exemplary semantic graph for an animate noun phrase.

FIG. 10 illustrates an exemplary semantic graph for an inanimate noun phrase retrieved from a Linked Data Source **(50)** of the system **(100)** of **FIG. 1**.

FIGS. 11 (a-b) illustrate exemplary semantic graphs of spatial information of the inanimate noun phrase of **FIG. 10**.

FIG. 12 illustrates a flowchart of substeps for semantic entity resolution step of the method of **FIG. 2**.

DESCRIPTION OF THE PREFFERED EMBODIMENT

A preferred embodiment of the present invention will be described herein below with reference to the accompanying drawings. In the following description, well known functions or constructions are not described in detail since they would obscure the description with unnecessary detail.

FIG.1 shows a system for entity resolution **(100)** according to an embodiment of the present invention. The system **(100)** resolves an entity in a sentence by utilizing a generic resolution, a spatial and temporal resolution, and semantic resolution. The system (**100**) comprises of a syntax analysis module **(10)**; a main verb identification module **(20)**; an acronym and noun phrase identification module **(30)**; an entity resolver module **(40)**; a Linked Data Source **(50)**; a Linguistic Resource module **(60)**; and a Linguistic Structure and Prototype Repository **(70)**.

The syntax analysis module **(10)** is connected to the main verb identification module **(20)**. The syntax analysis module **(10)** is configured to transform a sentence input into a syntax structure, which is used as an input to the main verb identification module **(20)**.

The main verb identification module **(20)** is connected between the syntax analysis module **(10)** and the acronym and noun phrase identification module **(30)**. The main verb identification module **(20)** is used to identify the main verb of the syntax structure from the syntax analysis module **(10)**. The main verb identification module (**20**) outputs the identified main verb to the acronym and noun phrase identification module (**30**).

The acronym and noun phrase identification module **(30)** is connected to the main verb identification module **(20)** and the entity resolver module **(40)**. The acronym and noun phrase identification module **(30)** is configured to identify both acronyms and noun phrases from the syntax structure in order to produce the acronym and noun phrase list.

The entity resolver module **(40)** is connected to the acronym and noun phrase identification module **(30)**, the Linked Data Source **(50)**, the Linguistic Resource module **(60)** and the Linguistic Structure and Prototype Repository **(70)**. The entity resolver module **(40)** is configured to perform entity resolution of the acronym and noun phrase list based on a generic entity resolution, a spatial and temporal entity resolution, and semantic resolution. The entity resolver module **(40)** includes a generic resolution module **(41)**, a spatial and temporal resolution module **(42)**, and a semantic resolution module **(43)**. The generic resolution module **(41)** performs the generic entity resolution on each acronym and each noun phrase in the acronym and noun phrase list by utilizing the Linked Data Source **(50)**, wherein the Linked Data Source **(50)** comprises of data that contains real-world information which interlink with one another. The spatial and temporal resolution module **(42)** performs the spatial and temporal entity resolution by extracting the spatial and temporal information from an animate and an inanimate noun phrase of the acronym and noun phrase list from the Linguistic Resource module (**60**), wherein the Linguistic Resource module (**60**) comprises of different types of linguistics such as syntax, semantics and morphology The semantic resolution module **(43)** performs the semantic entity resolution by comparing semantic structures to a linguistic structure and prototype retrieved from the Linguistic Structure and Prototype Repository (**70**), wherein the Linguistic Structure and Prototype Repository (**70**) comprises of different kinds of language structures and prototypes that are used to categorize entities.

Referring to **FIG.2**, illustrates a flowchart of a method for entity resolution according to an embodiment of the present invention. Initially, a sentence is inputted by a user into the syntax analysis module **(10)** as in step **200**. The sentence is then transformed into a syntax structure as in step **203**. For example, based on a sentence inputted by the user which is “MJ found fame with the Chicago Bulls.”, the syntax analysis module (**10**) transforms the sentence into the syntax structure as provided in **Table 1**.

#Text	Baseform	Syntactic relation	Syntax and morphology
1 MJ	mj	subj:>2	@SUBJ%NH ABBR NOM SG @SUBJ%NH ABBR NOM
2 found	Find	main:>0	@+FMAINV%VA V PAST
3 fame	Fame	obj:>2	@OBJ%NH N NOM SG
4 with	with	ins:>2	@ADVL%EH PREP
5 the	The	det:>7	@DN>%>N DET
6 Chicago	Chicago	attr:>7	@A>%N N NOM SG
7 Bulls	bull	pcomp:>4	@<P%NH N NOM PL
8 <s>	<s>		

Table 1

In step **201**, the main verb of the syntax structure is identified by the main verb identification module **(20)**. The main verb is identified by determining the base form of a word linked to "+FMAINV" in the syntax structure. Based on **Table 1**, the word "found" is linked to "+FMAINV" and thus, the main verb is identified as "find" that is the base form of the word "found".

Next, the acronym and noun phrase is identified by the acronym and noun phrase identification module **(30)** as in step **202**. The acronym is identified by determining the base form of a word linked to "ABBR" in the syntax structure while the noun phrase structure are identified by determining the base form of a word linked to "%>N" and "%NH". Based on **Table 1**, the word "MJ" is linked to "ABBR" and thus, the acronym is "MJ". The word "Chicago Bulls" is linked to "%>N" and "%NH". Thus the noun phrase identified is "Chicago Bulls". Alternatively, the syntactic relation as shown in **Table 1** may be used to identify the noun phrase. The acronym and noun phrase identification module **(30)** produces an acronym and noun phrase list based on the identified acronym and noun phrase as in step **204**.

Next, the entity recognition is performed to identify the noun phrases such as person, organization, location and so on from the sentence as in step **205**.

In step **209**, the generic resolution module **(41)** performs the generic entity resolution on each acronym and each noun phrase in the input sentence. Next, animate noun phrases and inanimate noun phrases of the input sentence are resolved using spatial and temporal information performed by the spatial and temporal resolution module **(42)** as in step **210**. In step **211**, the semantic resolution module **(43)** performs the semantic entity resolution on the input sentence by comparing semantic structures to a linguistic structure and prototype retrieved from the Linguistic Structure and Prototype Repository **(70)**. After going through all these steps, the entities of the input sentence are completely resolved as in step **212**.

Referring to **FIG. 3**, it shows the substeps for the generic entity resolution step as in step **209** of the method of **FIG. 2**. Initially, the generic resolution module **(41)** determines whether each acronym in the acronym and noun phrase list has been processed based on the generic entity resolution step as in decision **301**. If the acronym has been processed, the candidate with the highest total match overlap with the acronym is chosen as in step **311** and thereon, the process proceeds to the spatial and temporal entity resolution step (step **210**).

If an acronym has not been processed, the acronym is selected from the acronym and noun phrase list as in step **302**. Then, the generic resolution module **(41)** retrieves all relevant candidates from the Linked Data Source **(50)** as in step **303**. For example, if a sentence inputted by the user is "*MJ found fame with the Jackson 5*" and the acronym is identified as *MJ*, the generic resolution module **(41)** retrieves the candidates for the acronym *MJ* which is *Michael_Jackson* as Candidate #1 and *Michael_Jordan* as Candidate #2 as shown in **FIG. 4**.

Next, each candidate is determined whether the semantic graph for each candidate has been retrieved or not as in decision **304**. If all of the candidates have been processed, the step returns to step **301** which is to determine whether all the acronyms have been processed. Otherwise, a semantic graph for each candidate is retrieved from the Linked Data Source **(50)** in step **305.** For example, the graphs each of the candidate which are *Michael_Jackson* and *Michael_Jordan* are obtained as shown in **FIG. 5** and **FIG.6** respectively.

Thereon, the generic resolution module **(41)** determines whether each noun phrase in the acronym and noun phrase list has been processed based on the generic entity resolution step as in decision **306**. If all the noun phrases have been processed, the step returns to step **304** which is to determine whether all candidates have been processed. Otherwise, as in step **307**, a noun phrase is selected from the acronym and noun phrase list. In step **308**, the semantic graph for the noun phrase is retrieved from the Linked Data Source **(50)**. Based on the previous example, the noun phrase, *Jackson 5* is selected and its semantic graph retrieved from the Linked Data Source **(50)** as shown in **FIG. 7**. The semantic graph shows that *Michael_Jackson* is the former band member of Jackson 5 while *Marlon_Jackson* is the band member of Jackson 5.

Next, in step **309** the semantic graph of each candidate is matched with the semantic graph of the noun phrase. For each of the overlapping matches, the total match count is accumulated. In step **310** the candidate's semantic graph that has the best match overlap which is has the highest accumulated total match with the noun phrase graph is chosen. In the previous example, *Michael_Jackson* is chosen as its semantic graph has the highest total match with the Jackson 5's semantic graph. The step then returns to step **306** which is to determine whether each noun phrase in the acronym and noun phrase list has been processed.

Referring to **FIG. 8**, there is shown the substeps for the spatial and temporal entity resolution step as in step **210** of the method of **FIG. 2**. The substeps are described herein with reference to an example of a sentence inputted by the user, wherein the sentence is provided as below.

> "*The sell-out crowd in the Aberdeen exhibition centre is packed with young Scots who cheering a hero not much older than themselves. Against Rusedski, Andy Murray loses his first singles match, much to the crowd's dismay. But he's back the following evening to beat the British number two.*"

In steps **312, 313** and **314**, the spatial and temporal resolution module **(42)** extracts all noun phrases of animate type, inanimate type, and event type from the Linguistic Resource module **(60)** and the acronym and noun phrase list. Animate type is defined as an animate object or a person while inanimate type is defined as an

inanimate object or a place. The event type is defined as type of occasion or occurrence.

Next, in step **315**, the spatial and temporal resolution module **(42)** determines whether each animate noun phrase in the acronym and noun phrase list has been processed based on the spatial and temporal resolution step. If the animate noun phrase has been processed, the process proceeds to next step which is semantic resolution step (step **211)**.

Otherwise, a noun phrase of an animate type is selected as in step **316** which is previously retrieved in step **312**. In step **317**, a semantic graph for the selected noun phrase, G_{NP} is retrieved from the Linked Data Source **(50)**. Based on the example, the semantic graph for *Andy_Murray* is obtained as shown in **FIG.9**. The attributes of *Andy_Murray* such as label, residence, birthplace and description are obtained.

Thereon, each inanimate noun phrase is determined whether they have been matched with the semantic candidate retrieved from the Linked Data Source **(50)** as in decision **318**. If they have, the step repeats from step **315** which is to determine whether each animate noun phrase in the acronym and noun phrase list has been processed. Otherwise, in step **319** a noun phrase of inanimate or event type from the sentence is selected. Based on the example, the inanimate type selected from the sentence is *Aberdeen***.** Then, the semantic candidates that match for the inanimate noun phrase is retrieved from the Linked Data Source **(50)** in step **320**. Based on the example, the semantics candidates for Aberdeen retrieved are *Aberdeen,_Hong_Kong, Aberdeen,_New_South_Wales Aberdeen,_Scotland* and *Aberdeen,_Maryland* as shown in **FIG. 10.**

Next, the spatial and temporal resolution module **(42)** determines whether the semantic graphs for each of the semantic candidates have been retrieved as in decision **321**. If each semantic candidates have been processed, the candidate with the least distance or with the most relevancy weight with the animate noun phrase is selected as in step **322** and the step repeats from step **318** which is to determine whether each inanimate noun phrase has been processed. If each of the semantic candidates has not been processed, the graph for each candidate, G_c is retrieved as

in step **323.** Based on the example the, graphs for *Aberdeen, Hong_Kong* and *Aberdeen,_Scotland* are obtained as shown in **FIG. 11 (a-b)**.

From the G_c and G_{NP} obtained, spatial and temporal information is extracted from the Linked Data Source **(50)** as in step **325**. As shown **FIG.11 (a-b)**, the graphs of candidates for *Aberdeen* (*Aberdeen,_Hong_Kong and Aberdeen,_Scotland*) are retrieved such as the label, the latitude and the longitude. Then, G_{NP} and G_c are determined whether both of the graphs possess spatial information as in decision **324**.

If they do, the distance between the locations of G_{NP} and G_c is determined in step **327**. Based on the example, *Andy Murray* is living in *London, England* which is retrieved from the Linked Data Source **(50)**. Apparently the candidate *Aberdeen,_Scotland* is be the best candidate and relevant as it is nearer compared to *Aberdeen,_Hong_Kong*. If G_{NP} and G_c don't possess spatial information, the temporal information such as date and year are analysed and a relevancy weight is assigned as in step **326**. An example of a sentence inputted by the user is "*Roger won the Wimbledon Cup in 2003*". *Roger* is ambiguous and could possibly refer to *Roger Federer* or *Roger Taylor* who are both well-known tennis players. *Roger* is identified as an animate noun phrase G_{NP}. The temporal attribute such as the birthdate is taken into consideration. Based on the example, the temporal attribute of the sentence is year *2003*. The temporal attribute for each of the candidate of the animate type noun phrase, G_c , *Roger Federer* and *Roger Taylor* birthdate then are extracted from the Linked Data Source **(50)**. The most recent have higher weight assigned. For this sentence, *Roger Federer* born in the year 1981 is closer to 2003 if compared to *Roger Taylor* who is born in *1941*. After step **326** and step **327**, the step returns to step **321** which to determine whether each of the semantic candidates have been processed or not.

Referring to **FIG. 12**, there is shown the substeps for the semantic entity resolution step as in step **211** of the method of **FIG. 2**. For example, if a sentence is inputted by the user is "*Roger won the Wimbledon Cup in 2003*". The base form of the main verb of the sentence is identified in step **338** which is *win* and retrieved from the main verb list using morphology analysis. Morphology analysis is done by analysing each of the words in a sentence in order to identify type of speech the word

belongs to, whether it is a noun, verb, adjective, preposition and so on. Morphology analysis is also able to identify the base form of a noun or a verb. Next, the sentence is transformed into a semantic structure, G_s in step **339**. Based on the example, the semantic structure, G_s is as shown below:

G_s: [T:"Roger"] ←(agnt) ←[win]-{
(thme)→[tournament:"Wimbledon Cup"];
(ptim) →[year:2003];
}.

Thereon, the semantic resolution module **(43)** retrieves all linguistic structures of the identified main verb, *win* from the Linguistic Structure & Prototype Repository **(70)** as in step **340**. Based on the example given, the linguistic structures obtained as shown below:

Linguistic Structure #1:

[person] ←(agnt) ←[win] →(thme) →[tournament] →(loc) →[location]

Linguistic Structure #2:

[person] ←(agnt) ←[win]-{
(thme)→[tournament];
(ptim) →[year];
}.

Next, in step **341**, the semantic graph similarity is determined between G_s and each of the linguistic structure retrieved. The best match of G_s to the retrieved linguistic structure is selected as in step **342**. From this example, linguistic structure #2 is selected due to the similarities of the structure between the linguistic structure and G_s. From the linguistic structure that has been selected, the concept type of the agent of the main verb is determined in step **343**. In this example, the agent of the main verb of the sentence is *Roger* and concept type that refers to *Roger* is "animate being". The prototypes that contain the subject word with the same concept type, *Roger Taylor* and *Roger Federer* are retrieved from the Linguistic Structure & Prototype Repository **(70)** in step **344**. The prototypes obtained are as shown below:

Prototype #1:

[tennis-player:"Roger Taylor"] ←(agnt) ←[win] →(thme) →[tournament:"Copenhagen Open"].

Prototype #2:

[tennis-player:"Roger Federer"] ←(agnt) ←[win]-{
(thme) →[tournament:"Wimbledon Cup"];
(ptim) →[year:"2003"];
}

The semantic graph similarity is determined between G_s and all the retrieved prototypes as in step **345**. The prototype that matches the best with G_s is selected as in step **346**. Based on the example, the prototype#2 is selected as the best match.

[tennis-player:"Roger Federer"] ←(agnt) ←[win]-{
(thme) →[tournament:"Wimbledon Cup"];
(ptim) →[year:"2003"];
}

After the best match is found, the semantic resolution module **(43)** extracts the matched entity in step **347**. Based on the example, *[tennis-player:"Roger Federer"]* is selected as resolved entity for *Roger*.

Finally, the entities of the input sentence are resolved by using the results of the semantic resolution step **(210)** as in step **348**.

While embodiments of the invention have been illustrated and described, it is not intended that these embodiments illustrate and describe all possible forms of the invention. Rather, the words used in the specifications are words of description rather than limitation and various changes may be made without departing from the scope of the invention.

CLAIMS

1. A system for entity resolution **(100)** comprises of:
   a) a syntax analysis module **(10)** to transform the sentence inputted by the user into a syntax structure,
   b) a main verb identification module **(20)** to identify the main verb of the syntax structure,
   c) an acronym and noun phrase identification module **(30)** to identify the acronym and noun phrases from the syntax analysis module **(10)**,
   d) an entity resolver module **(40)** to perform entity resolution,
   e) a Linked Data Source **(50)** to provide data that contains real-world information which interlink with one another,
   f) a Linguistic Resource module **(60)** to provide different types of linguistics such as syntax, semantics and morphology, and
   g) a Linguistic Structure and Prototype Repository **(70)** to provide the different kinds of language structures and prototypes that are used to categorize entities,

   wherein the system (**100**) is characterised by the entity resolver module **(40)** includes:
   i. a generic resolution module **(41)** to perform the generic entity resolution on each acronym and each noun phrase in the acronym and noun phrase list by utilizing the Linked Data Source **(50)**;
   ii. a spatial and temporal module **(42)** to perform the spatial and temporal entity resolution by extracting the spatial and temporal information from an animate and an inanimate noun phrase of the acronym and noun phrase list from the Linguistic Resource module (**60**); and
   iii. a semantic resolution module **(43)** to perform the semantic entity resolution by comparing semantic structures to a linguistic structure and prototype retrieved from the Linguistic Structure and Prototype Repository **(70)**.

2. A method for resolving entity automatically is characterised by the steps of:
   a) performing syntax analysis to a sentence and transforming the sentence into a syntax structure;
   b) identifying a main verb of the syntax structure;

c) identifying an acronym and a noun phrase of the syntax structure;
d) performing entity recognition to identify the noun phrases of the sentence; and
e) performing a generic resolution, a spatial and temporal resolution and a semantic resolution.

3. The method as claimed in claim **2**, wherein the step of performing generic resolution includes:
a) choosing a candidate with the highest total accumulated match if the acronym has been processed;
b) proceeding to the spatial and temporal resolution step after the candidate the highest total accumulated match is chosen;
c) selecting an acronym from the acronym and noun phrase list if the acronym has not been processed;
d) retrieving relevant candidates for the acronym from the Linked Data Source **(50)**;
e) retrieving a semantic graph for each candidate from the Linked Data Source **(50)** if the candidates has not been processed;
f) selecting a noun phrase from the acronym and noun phrase list if the noun phrase has been processed;
g) retrieving a graph for the noun phrase from the Linked Data Source **(50)**;
h) matching the semantic graph of each candidate with semantic graph of the noun phrase; and
i) choosing the semantic graph of the candidate that has the highest total match with the graph of the noun phrase.

4. The method as claimed in claim **2**, wherein the step of performing the spatial and temporal resolution includes:
a) retrieving all noun phrases of animate, inanimate and event type from Linguistic Resource module **(60)** and acronym and noun phrase list;
b) determining whether each animate noun phrases have been processed based on the spatial and temporal resolution step; and
c) proceeding to the semantic resolution step if the animate noun phrase has been processed;

d) selecting the retrieved noun phrase of an animate type if the animate noun phrase noun has not been processed;
e) selecting a noun phrase of an inanimate or event type;
f) retrieving semantic candidates that match the inanimate or event type of the noun phrase from the Linked Data Source **(50)**;
g) retrieving the semantic graph for the noun phrase, G_{NP} from the Linked Data Source **(50)**;
h) selecting the candidate with the least distance or with the most relevancy weight if the semantic candidates have been processed;
i) retrieving the graph for each of the candidates, G_c if the semantic candidates have not been processed;
j) extracting spatial and temporal information of G_{NP} and G_c from the Linked Data Source **(50)**;
k) determining the distance between G_{NP} and G_c locations if G_{NP} and G_c possess spatial information; and
l) analysing the temporal information between G_{NP} and G_c and assigning a relevancy weight if G_{NP} and G_c don't possess spatial information.

5. The method as claimed in claim **2**, wherein the step of performing the semantic resolution includes:
a) identifying and retrieving a main verb of a sentence from a main verb list;
b) transforming the sentence into a semantic structure, G_s;
c) retrieving all linguistic structures of the identified main verb from Linguistic Structure and Prototype Repository **(70)**;
d) determining the semantic graph similarity between G_s and the retrieved linguistic structures;
e) selecting the linguistic structure that matches the best to G_s;
f) determining the concept type of the agent of the main verb;
g) retrieving all prototypes which contain the subject with the same concept type from the Linguistic Structure and Prototype Repository **(70)**;
h) determining the semantic graph similarity between G_s and all retrieved prototypes;

i) selecting prototype that matches the best with G_s;

j) extracting the matched entity from the prototype; and

k) resolving entities based on the results of semantic resolution step.

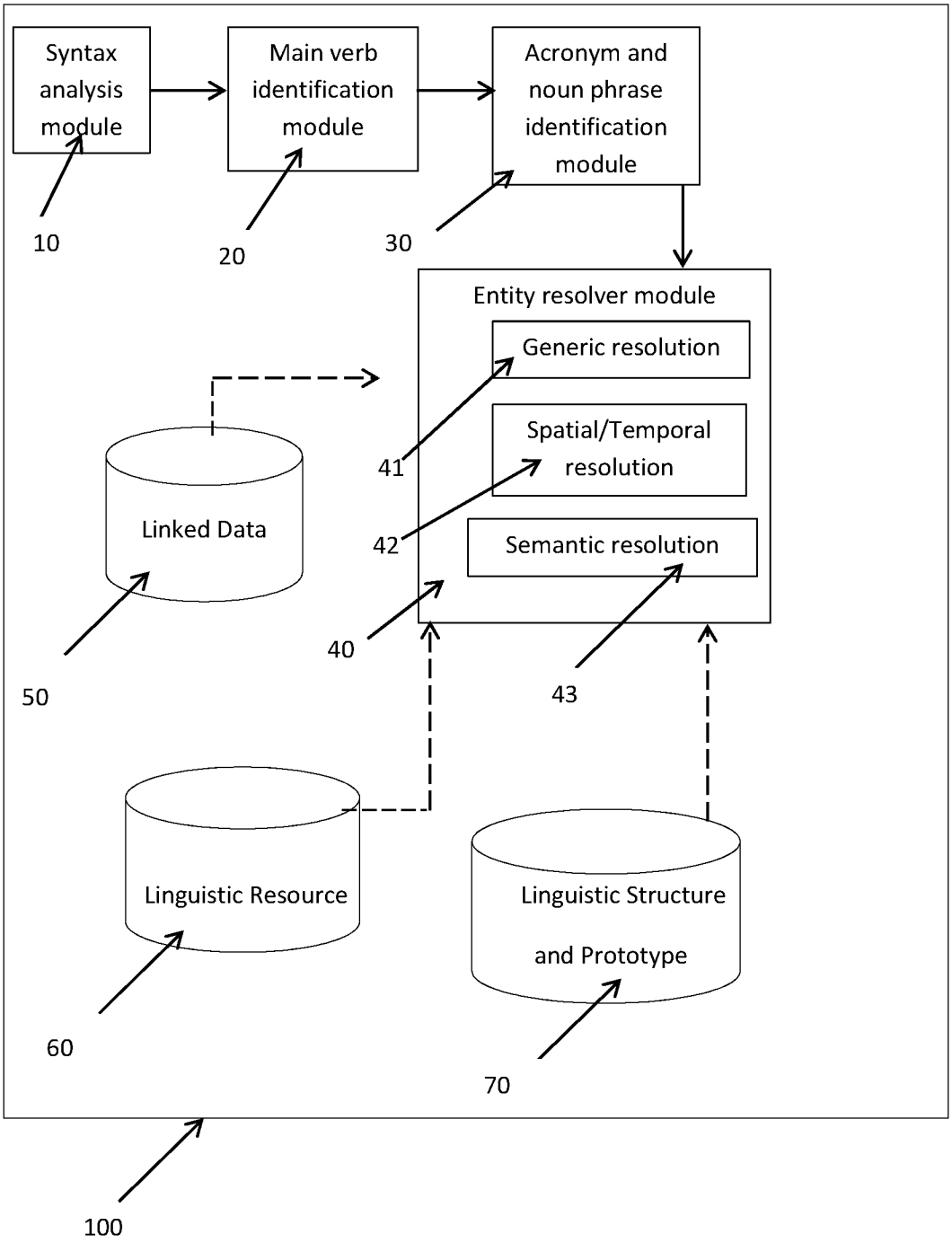

Syntax analysis module
Main verb identification module
Acronym and noun phrase identification module
10
20
30
Entity resolver module
Generic resolution
Spatial/Temporal resolution
Semantic resolution
41
42
Linked Data
40
43
50
Linguistic Resource
Linguistic Structure and Prototype
60
70
100

FIG. 1

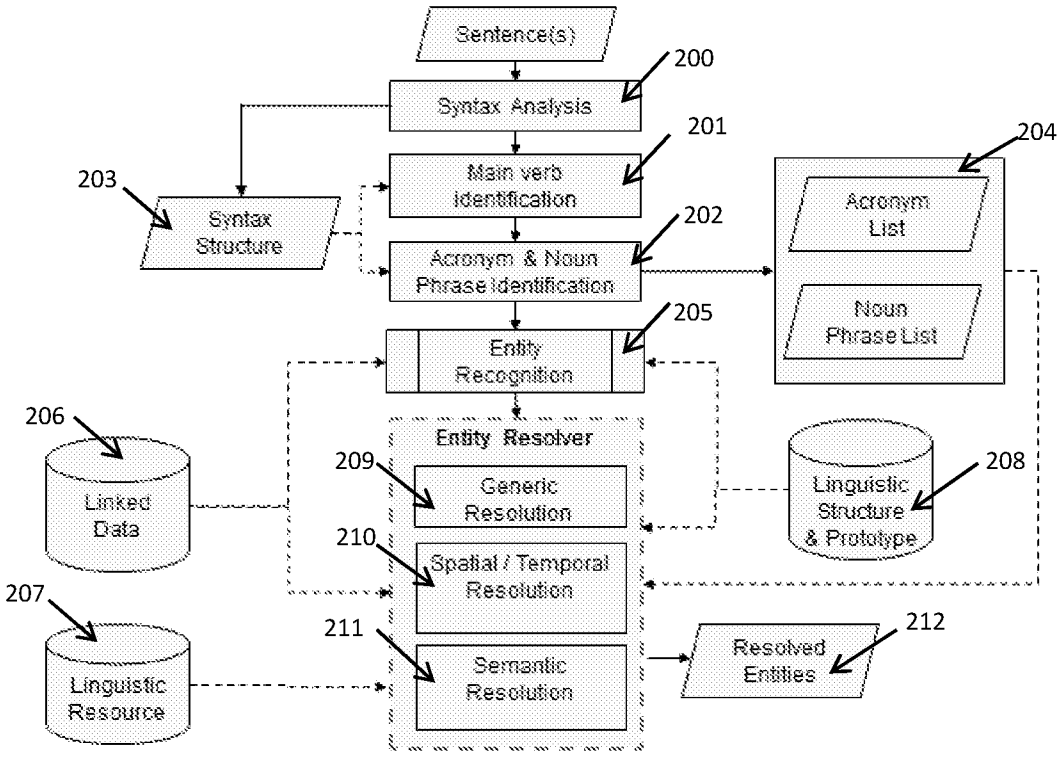

Sentence(s)
200
Syntax Analysis
201
Main verb Identification
203
Syntax Structure
202
Acronym & Noun Phrase Identification
204
Acronym List
Noun Phrase List
205
Entity Recognition
206
Linked Data
Entity Resolver
209
Generic Resolution
210
Spatial / Temporal Resolution
211
Semantic Resolution
208
Linguistic Structure & Prototype
207
Linguistic Resource
212
Resolved Entities

FIG. 2

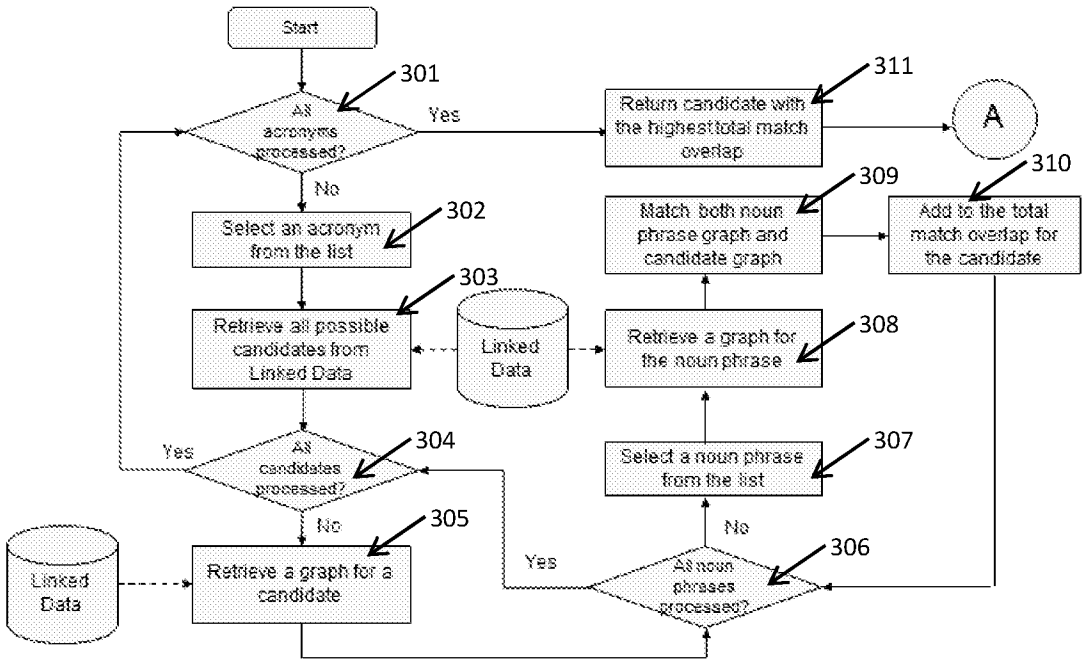


FIG. 3

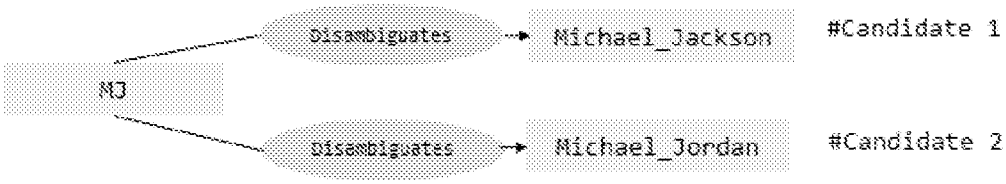


FIG. 4

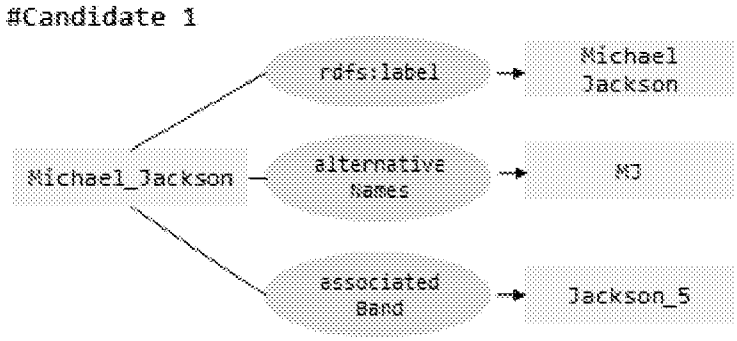


FIG. 5

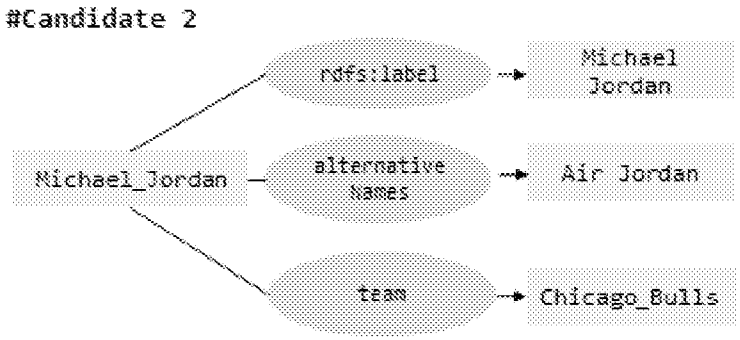

#Candidate 2
rdfs:label
Michael Jordan
Michael_Jordan
alternative Names
Air Jordan
team
Chicago_Bulls

FIG. 6

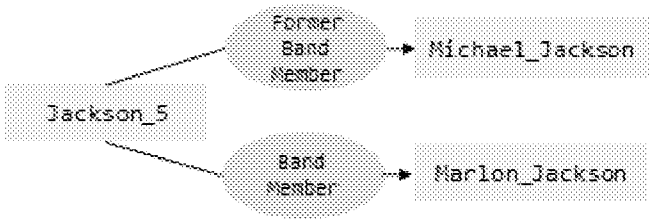

Former Band Member
Michael_Jackson
Jackson_5
Band Member
Marlon_Jackson

FIG. 7

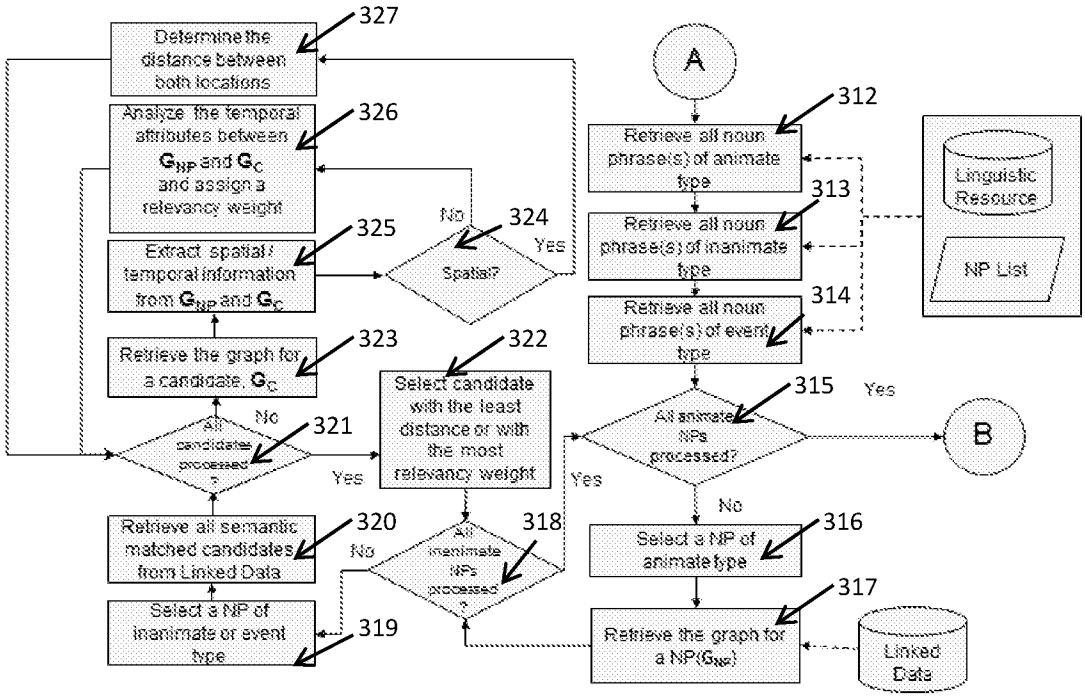

Determine the distance between both locations
327
Analyze the temporal attributes between G_{NP} and G_C and assign a relevancy weight
326
Extract spatial/ temporal information from G_{NP} and G_C
325
Spatial?
324
No
Yes
Retrieve the graph for a candidate, G_C
323
All candidates processed ?
321
No
Yes
Select candidate with the least distance or with the most relevancy weight
322
Retrieve all semantic matched candidates from Linked Data
320
Select a NP of inanimate or event type
319
All inanimate NPs processed ?
318
No
A
Retrieve all noun phrase(s) of animate type
312
Retrieve all noun phrase(s) of inanimate type
313
Retrieve all noun phrase(s) of event type
314
Linguistic Resource
NP List
All animate NPs processed?
315
Yes
B
Yes
No
Select a NP of animate type
316
Retrieve the graph for a NP(G_{NP})
317
Linked Data

FIG. 8

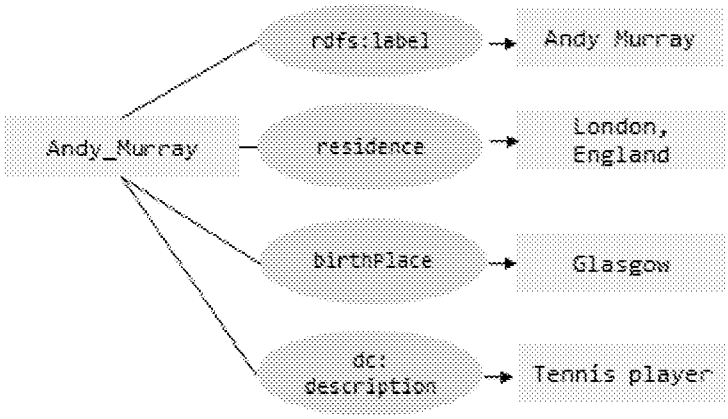

Andy_Murray
rdfs:label
Andy Murray
residence
London, England
birthPlace
Glasgow
dc: description
Tennis player

FIG. 9

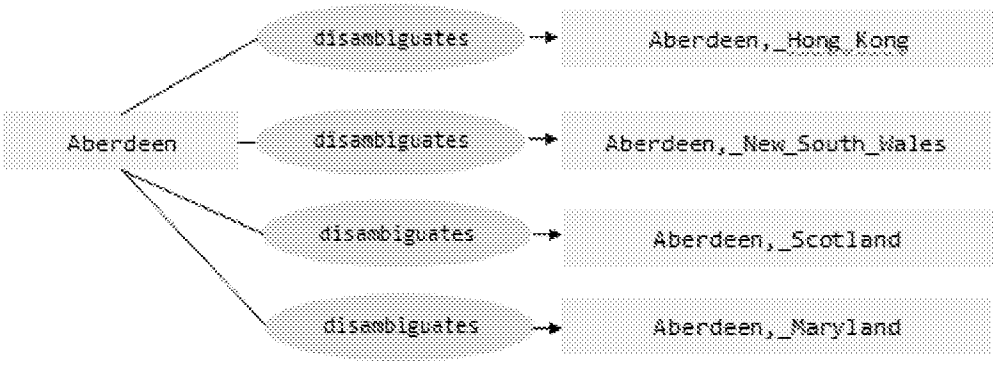


FIG. 10

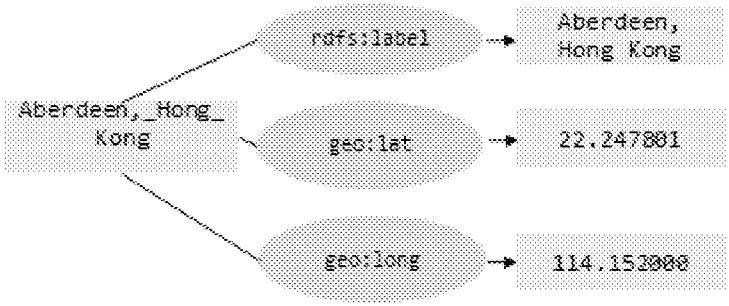


FIG. 11a

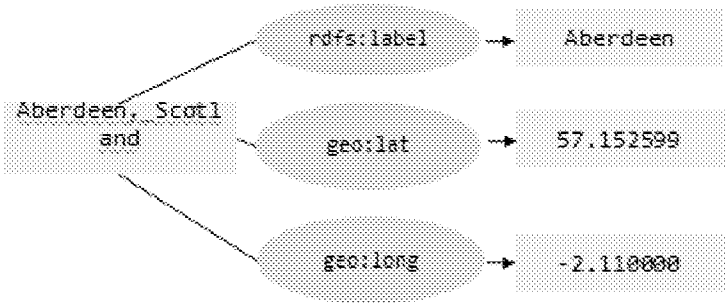


FIG. 11b

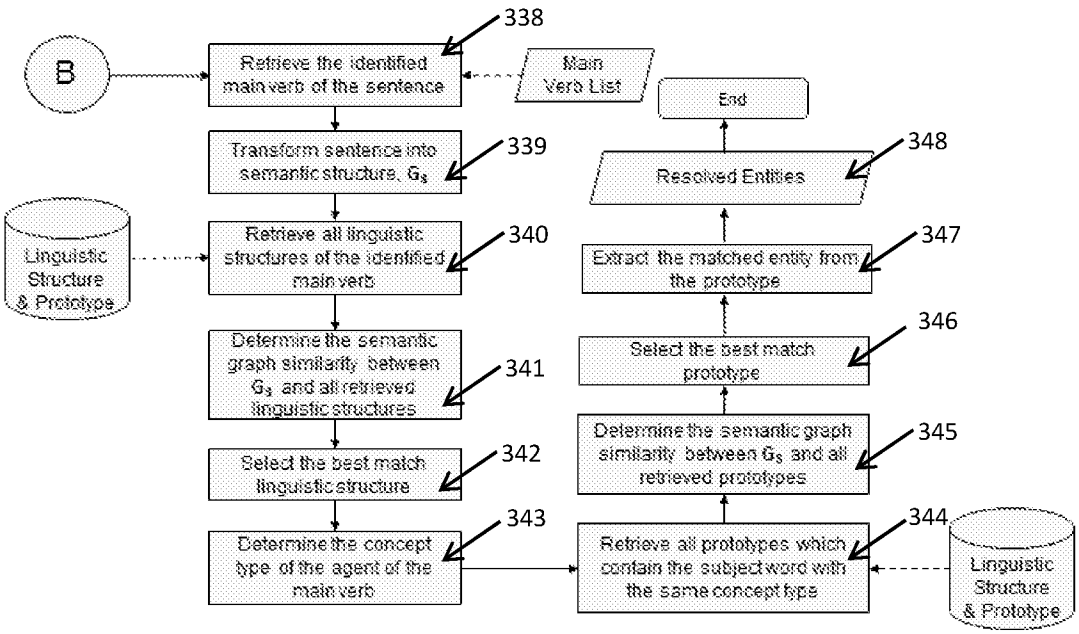

B
Retrieve the identified main verb of the sentence
338
Main Verb List
Transform sentence into semantic structure, G_s
339
Linguistic Structure & Prototype
Retrieve all linguistic structures of the identified main verb
340
Determine the semantic graph similarity between G_s and all retrieved linguistic structures
341
Select the best match linguistic structure
342
Determine the concept type of the agent of the main verb
343
Retrieve all prototypes which contain the subject word with the same concept type
344
Linguistic Structure & Prototype
Determine the semantic graph similarity between G_s and all retrieved prototypes
345
Select the best match prototype
346
Extract the matched entity from the prototype
347
Resolved Entities
348
End

FIG. 12