Title: A METHOD AND SYSTEM FOR AUTOMATED RELATION DISCOVERY FROM TEXTS

Abstract: The present invention provides a system (100) for discovering relations between texts in sentence of a machine-readable document. The system comprises a text preprocessor (101) and a relation discovery module (102). The text preprocessor (101) processes the documents to identify and extract entities, noun phrases and verb from thereof. The relation discovery module (102) discovers the relation through a generic and semantic relation extraction for unstructured and structured texts to resolves intra-sentential and inter-sentential contexts.
A Method and System for Automated Relation Discovery from Texts

Field of the Invention

[0001] The present invention relates to information extraction. More specifically, the present invention relates to a system and method for automated relation discovery from texts.

Background

[0002] Given typed entities and relations, it would be desired to be able to infer implicit contexts from both structured and unstructured collections of texts. To be able to achieve that, one need to be able to extract the entities and their relations with each other from natural language texts. There are two main challenges: 1) Extraction of entities from structured/unstructured texts; 2) Extraction of typed relations between entities from structured/unstructured texts; and 3) Intra-sentential relation extraction.

[0003] Conventional approach relation discovery from natural language texts requires substantial amount of training examples or tagged datasets, i.e. supervised approach. Such supervised approach offers poor extraction quality of entities from texts; as it will either heavily dependent on the availability of the annotated data, or the use of predefined gazetteer lists do not accommodate the diversity to address for multiple domains.

[0004] US patent publication no. US2009/00 19032 A1, in FIG 2, showed a table below as an example for annotated training corpus used by a conventional extraction method known in the art.
Training Corpus

"We found that TP53 is a lung cancer gene"
"Smoking is bad for your lungs"

Annotated Training Corpus

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>Sentence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>Label</td>
</tr>
<tr>
<td>We</td>
<td>common (c)</td>
</tr>
<tr>
<td>found</td>
<td>common (c)</td>
</tr>
<tr>
<td>that</td>
<td>common (c)</td>
</tr>
<tr>
<td>TP53</td>
<td>GENE (G)</td>
</tr>
<tr>
<td>is</td>
<td>common (c)</td>
</tr>
<tr>
<td>a</td>
<td>common (c)</td>
</tr>
<tr>
<td>lung</td>
<td>disease (d)</td>
</tr>
<tr>
<td>cancer</td>
<td>disease (d)</td>
</tr>
<tr>
<td>gene</td>
<td>common (c)</td>
</tr>
</tbody>
</table>

[0005] Details of the illustrations are provided in the US publication and therefore not elaborate herein.

[0006] When it comes to the extraction of typed relations between entities from structured/unstructured texts, the state-of-the-art approaches frequently involve parsing that depends on syntax structures (such as part-of-speech 'POS' tags, dependency parse trees, etc).

[0007] Further, relation extraction techniques thus far focus on intra-sentential context (i.e. within a sentence). The challenge, which is also the keen interest, is to extract relations between entities across sentences, in the inter-sentential context. On the overall outset of the problems, there are mainly two: granularity (e.g. intra-sentential level vs. inter-sentential level), heterogeneity of the texts (structured vs. unstructured documents) and the use of domain specific sources versus open domains.
In accordance with one aspect of the present invention, there is provided a system for carrying out a relation extraction from a machine-readable document having sentences. The system comprises a text preprocessing module for lemmatizing the sentences into tokenized text and identifying paragraphs from the document; a coreference resolution module configured to resolve all possible anaphors; an entity recognition module for extracting entities from the sentence; an entity resolution and disambiguation module adapted for resolving all ambiguous entities / noun phrases, acronyms and abbreviations; a relation extraction module configured to operably carrying out a generic relation extraction and a semantic relations extraction for extracting triples for a sentence based on the extracted entities, wherein a weighted ranking score is computed for each triple for a sentence for selecting and storing an most appropriate triple.

In one embodiment, each triple is computed with a basic ranking score by matching the identified concepts with the respective verb schemas, and weighted with their respective recorded popularity scores to obtain the weighted ranking score for each triple, the highest ranking score of which is selected as the most appropriate triple.

In another embodiment, the generic relation extraction among entities in an intra-sentential context extracts noun phrases and verb from linguistic resources to match the sentence with predefined patterns, the linguistic resources and a Linked Data is searched for possible properties to generate corresponding relation the sentence refers to.
In yet another embodiment, the generic relation extraction among entities in an inter-sentential context retrieves a list of entities and noun phrases from a repository and another list of entities, noun phrases that are not within the same paragraph and the connecting verbs in order to extract and match schemas based on the verb identified from the linguistic resources to identify relations from a knowledge base, the linguistic resources is searched for possible triples based on the properties found.

Further, the semantic relations extraction for entities in an intra-sentential context matches predefined patterns to extract and match schemas based on the identified verb, and searches entities or noun phrases through a linked database.

In yet a further embodiment, the semantic relations extraction for entities in an inter-sentential context retrieves a list of entities and noun phrases from a repository and another list of entities, noun phrases that are not within the same paragraph and the connecting verbs in order to extract and match schemas based on the verb identified from the linguistic resources.

In another aspect of the present invention, there is provided a method of carrying out a relation extraction from a machine-readable document having sentences. The method comprises lemmatizing texts of the sentences into tokenized text and identifying paragraphs from the document; resolving all possible anaphors through a coreference resolution module; extracting entities from the sentences through entity recognition module for extracting entities from the sentence; resolving all ambiguous entities / noun phrases, acronyms and abbreviations; carrying out a generic relation extraction and a semantic relations extraction for extracting triples for a sentence based
on the extracted entities, wherein a weighted ranking score is computed for each triple
for a sentence for selecting and storing an most appropriate triple.

[0015] In one embodiment, the relation extraction further comprises computing a basic ranking score by matching the identified concepts to the respective verb
schemas; retrieving a popularity score for each of the schemas, computing a popularity
weighted score for each schema to obtain weighted ranking score for each schema;
selecting a schema with the highest weighted ranking score; generating the relation
triples based on the selected schema.

[0016] In another embodiment the generic relation extraction among entities in
an intra-sentential context may extract noun phrases and verb from linguistic resources
to match the sentence with predefined patterns, the linguistic resources and a Linked
Data is searched for possible properties to generate corresponding relation the sentence
refers to; and the semantic relations extraction for entities in an intra-sentential context
matches predefined patterns to extract and match schemas based on the identified verb,
and searches entities or noun phrases through a linked database.

[0017] In a further embodiment, the generic relation extraction among entities in
an inter-sentential context retrieves a list of retrieves a list of entities and noun phrases
from a repository and another list of entities, noun phrases that are not within the same
paragraph and the connecting verbs in order to extract and match schemas based on the
verb identified from the linguistic resources to identify relations from a knowledge
base, the linguistic resources is searched for possible triples based on the properties
found; and the semantic relations extraction for entities in an inter-sentential context
retrieves a list of entities and noun phrases from a repository and another list of entities,
noun phrases that are not within the same paragraph and the connecting verbs in order to extract and match schemas based on the verb identified from the linguistic resources.

**Brief Description of the Drawings**

[0018] Preferred embodiments according to the present invention will now be described with reference to the figures accompanied herein, in which like reference numerals denote like elements;

[0019] FIG. 1 illustrates a relation discovery system in accordance with one embodiment of the present invention;

[0020] FIG. 2 illustrates a process of relation extraction process in accordance with one embodiment of the present invention;

[0021] FIGs. 3A and 3B exemplify an example of a sentence that is being processed to extract generic relation;

[0022] FIG. 4 exemplifies a further example of generic relation extractions;

[0023] FIG. 5 exemplifies an example of semantic relation extractions;

[0024] FIG. 6 exemplifies another example of generic relation extractions;

[0025] FIG. 7 illustrates the ranking process with a popularity weight in accordance with one embodiment of the present invention; and

[0026] FIG. 8 illustrates a process of inter-sentential discovery in accordance with one embodiment of the present invention.
Detailed Description

[0027] Embodiments of the present invention shall now be described in detail, with reference to the attached drawings. It is to be understood that no limitation of the scope of the invention is thereby intended, such alterations and further modifications in the illustrated device, and such further applications of the principles of the invention as illustrated therein being contemplated as would normally occur to one skilled in the art to which the invention relates.

[0028] FIG. 1 illustrates a relation discovery system 100 in accordance with one embodiment of the present invention. The system 100 is adapted for automatically discovering relations within and across sentences from natural language texts. The system is able to process both structured and unstructured texts. It is further capable of resolving intra-sentential and inter-sentential contexts. Yet, it is capable of discovering generic and semantic relations of the intra-sentential and inter-sentential contexts. The system comprises a text preprocessor 101, a relation discovery module 102, a Linked Data 103, a pattern database 104, linguistic resources 105 and triples 106. The system 100 receives machine-readable texts 190, or simple texts for processing. The texts are to be first processed by the text preprocessor 101 and subsequently through the relation discovery module 102 to discover relations between the texts. The text preprocessor 101 includes a tokenizer 112, a lemmatization module 114 and paragraph identifier 116. The text preprocessor 101 is adapted for transforming sentence into tokenized lemma form.

[0029] The relation discovery module 102 comprises a coreference resolution module 122, an entity recognition module 124, an entity resolution and disambiguation
module 126 and a relation extractor 128. The coreference resolution module 122 is configured to resolve all possible anaphors into the according noun antecedents. The entity recognition module 124 is configured to extract all possible entities from a sentence. The texts are then processed by the entity resolution and disambiguation module 126 to resolve all ambiguous entities/noun phrases, acronyms and abbreviations. Once the entities are identified and resolved, the relation extractor 128 determines and extracts the relationship that exists between entities.

[0030] During the operations of the relation discovery, the relation discovery module 102 is assisted through the Linked Data 103, the pattern database 104 and the linguistic resources 105. Specifically, the correlation resolution 122, entity recognitions 124, entity resolutions and disambiguation 126 are processed with the use of Linked Data 103; the coreference resolution 122 and the relation extraction 128 are processed with the use of linguistic resources 105; and the relation extraction 128 is further processed with the use of the predefined pattern database 104.

[0031] Through the paragraph identification 116, coreference resolution 122, entity recognition, entity resolution and disambiguation, inter-sentential contexts can be resolved. On the other hand, generic relations among entities in the intra-sentential context can be resolved by applying noun phrase extraction, then identifying the verb through an aid of the linguistic resources 105, then matching the sentence with the pattern database 104, and searching the linguistic resources 105 and the Linked Data 103 to determine the corresponding relation it refers to. It is to be noted that the Linked Data 103 can be a proprietary database or a public database.
FIG. 2 illustrates a process of relation extraction process in accordance with one embodiment of the present invention. At step 201, the system 100 performs a coreference resolution to resolve all anaphors to the corresponding noun antecedents. For example, given a sentence: "I saw Scott yesterday. He was fishing by the lake,". It can be identified that "Scott" and "he" in that sentence are coreference and therefore the anaphor 'he' will be resolved to refer to 'Scott'.

At step 202, the texts are tokenized into sentences. For each sentence, the entities/noun phrases are identified at step 203, and at step 204, the verbs are identified. The identified entities/noun phrases and verbs are then stored onto the index repository 205. To identify and extract the verb(s), the patterns of the sentence under process is being matched with the pattern database 104 to trigger a specified rule for extracting an appropriate relation. The schemas of the verb are also extracted from integrated linguistic resources to trigger the specified rule for extracting the most appropriate relation.

At step 210, the system 100 matches for intra-sentential patterns to identify relation triples 212 for storing on a knowledge base 215. To do that, it performs dual granularity for discovering and extracting both generic and semantic relations from texts. Once the relation extractions are done at the step 210, the system 100 determines if more sentences are to be processed. If there are at step 211, the system 100 returns to the step 202 to identify the next sentence to be processed. If not further sentence is to be processed, at step 220, the extraction type shall be specified. Subsequently, the system 100 further determines if there exist any intra-sentential to be processed at step 222. If intra-sentential pattern is matched a ranking/filtering is carried out at step 230. If the no further intra-sentential context is to be processed, the system
100 retrieves the entity and verb indexes 205 at step 224. The index is utilised for matching inter-sentential patterns at step 225. Similarly, the inter-sentential patterns matching also includes a generic relations extraction and a semantic relations extraction. Through the inter-sentential patterns matching at step 225, triples relation will be generated at step 228 and stored on the knowledge base 215.

[0035] The candidate triples in the intra-sentential and/or inter-sentential context are being ranked and filtered in the step 230. During the ranking and filtering process, the system computes a basic ranking score by matching the identified entities with the verb schemas. It then retrieves a popularity weight table that provides the popularities scores for the respective schemas. Accordingly, popularity weighted scores is computed accordingly for each schema. The schema with the highest weighted score is then selected to generate the relation triples based on the selected schema.

[0036] In a generic relations extraction, all the possible entities and noun phrases are identified through the entity recognition module 124. Generic relation extraction comprises the steps of retrieving a list of entities/noun phrase from the paragraph to be processed to form a List A from index and iterating each of the entities of the List A; selecting a next entry $E_i$ from the List A and retrieving a list of entities/noun phrase which is not within the same paragraph (i.e. other paragraphs) to generate a List B; searching the intra-sentential results stored in the knowledge base if the relation exist between $E_i$ and List B, wherein the List B can be reuse if exist; search the linguistic resources for $E_i$ and obtaining all possible candidate relations; removing candidate triple if the subject/object do not match the List B; and repeating the above steps the last entries $E_i$ in List A.
Given an example showing in FIG. 3A, where a partial sentence is provided "... most European countries especially England, Germany and France ...", "England", "Germany" and "France" shall be identified as possible entities, and "European countries" as noun phrase. These identified nouns are then being lemmatized. In this case, "European countries" is being lemmatized as "European country" and the rest remains no change. Following that, the linguistic resources 105, such as VerbNet and/or FrameNet, are used for identifying verb(s). In this sentence, no verb can be identified. Thereafter, matching patterns are retrieved to trigger the corresponding rule to generate relation triples. In this example, three hyponym relation triples are generated: HYPONYM(Germany, European country); HYPONYM(Germany, European country); and HYPONYM(France, European country).

Following the above, as shown in FIG. 3B, the system 100 goes on searching the linguistic resources 105, such as WordNet, etc, and retrieves all possible properties related to the entities/noun phrases to generate holonym relation triples: Holonym(Germany, Europe); Holonym(Germany, European Union); and Holonym(Germany, European Economic Community). A similarity measure between the additional property labels with all the entities/noun phrases identified in the sentence is carried out, and respective scores are being assigned to each holonym. The candidate relation triples are being filtered based on similarity scores computed with a threshold specified. In this case, "Europe" and "European country" will obtain a similarity match score of 73.33%; whereas, "European Union" and "European country" will obtain the match score of 54.55%.
FIG. 4 exemplifies a further example of generic relation extractions based on a sentence "Samsung release Galaxy Mini, another smartphone for Android fans." in another embodiment of the present invention. Similarly, the system 100 first identifies the possible entities/noun phrases from the exemplified example: Samsung, Galaxy Mini, etc. and the noun phrase that can be identified is "smartphone". The identified entities are then lemmatized accordingly. The linguistic resources 105 are then being used to identify verb, if any. In this given example, "release" is identified as a verb. The identified verb(s), entities and nouns are marked up accordingly for matching with the pattern database 104 according to the structure of the sentence exemplified here.

The matching patterns, if one can be found, trigger the corresponding rule to generate the relation triples. The relation triples are well known in the field of Resource Description Framework (RDF). For example, hyponymy or is-a relation, metonymy or part-of relation, synonyms and etc. may be utilized for determining relations between the phrases. The rules and matching patterns are used to extract generic relations, which are preferably of taxonomic type derivable derived from a knowledge base e.g. hyonyms, metonyms, synonyms, hypernyms, holonyms, antonyms, etc. By way of illustrations, not limitation, in a sentence like "... such exotic fruit as kiwis, mangoes, pineapples or coconuts ...", it is possible to extract that "kiwi is-a exotic fruit, mango is-a exotic fruit, etc. In another sentence link "... basement of a building ...", it is possible to extract that basement part-of building. In yet another sentence " ... United Kingdom or Great Britain ...", it is possible to extract that United Kingdom same-as Great Britain. This can be achieved by identifying certain pattern, which may consist specific words or terms defining the patterns.
Still referring to FIG. 4, two relation triples can be matched: (Samsung, release, Galaxy Mini) and (Samsung, release, smartphone), based on a predefined pattern: NP1+VP1+(NP2,[NP3+PP3]), where NP1, NP2, NP3 are referring to noun phrases, VP1 is referring to verb phrase and PP3 is referring to preposition phase.

Following that, the linguistic resources is searched again to retrieve the possible properties related to the entities/noun phrases. A similarity measure between the additional property labels with all the entities/noun phrases identified in the sentence is carried out. The candidate relation triples based on similarity scores computed with a threshold specified is being filtered.

Returning to step 210, the semantic relation extraction comprising the steps of retrieving a list of entities/noun phrase from the paragraph/sentence to be processed to form a List A from index and iterating each of the entities thereof; selecting a next entry $E_i$ from the List A and retrieving a list of entities/noun phrase which is not within the same paragraph (i.e. other paragraphs) to generate a List B; searching the intra-sentential results stored in the knowledge base if the relation exist between $E_i$ and List B, wherein the List B can be reused if exist; search Linked Data for $E_i$ and obtaining all possible candidate relations; removing candidate triple if the subject/object does not match the List B; and repeating the above steps the last entries $E_i$ in List A. In general, generic relations extraction differs from the semantic relations extraction in that the Linked Data.

FIG. 5 exemplifies a sentence "The Bugatti Veyron EB 16.4 is a mid-engined grand touring car by the Volkswagen Group in France" which is to be processed by the system 100. In this sentence, the entities/noun phrases "Bugatti", ...
Veyron, etc. can be identified. No verb, on the other hand, can be identified. Similarly, no patterns can be matched too. Through the Linked Data 103, a class property label "Sports car" can be retrieved for the entity "Bugatti", and through a list of predefined property rules to denote a specific relation to be triggered, isType relation can be applied to generate a relation triple: isType(Bugatti, Veyron, Sports Car).

[0045] FIG. 6 exemplifies another sentence "John bought Mary a Ferrari" which is to be processed by the system 100. In this example, when determining semantic relations in the intra-sentential context, similar steps are followed as described in the previous examples. The only difference from the previous scenarios is that when the verb is identified while there are no available patterns for this verb found. All possible schemas related to this target verb are extracted from the Linguistic Resources 105.

[0046] VerbNet for example, offers a comprehensive resource for verbs. Through this Linguistic Resource, the use of syntax parsing of the sentences to identify verbs can be eliminated. Subsequently, FrameNet for example can be used to identify all schemas to match all the identified concepts. To match the concepts with the schemas, the selectional constrains for each of the semantic roles are to be met.

[0047] In this given example, the identified entities are "John", "Mary", "Ferrari", etc. The identified verb is "buy" being the lemmatized word of "bought". Through the linguistic resources, the schema "Agent v Beneficiary Theme" can be identified. A semantic role agnt has the selectional constraint (animate, action). In this case, John who is a person matched to the first constraint 'animate' (i.e. living person) as the reference of a concept hierarchy will be used to determine this.
FIG. 7 shows the process of ranking the schemas with popularity weight to select the schema to generate the corresponding relation triples with the example of FIG. 6. The sentence 701 is processed to identify an appropriate schema, Schema 1, Schema 2 and Schema 3, from a schemas list 702. As explained above, each of the schemas is computed with a basic ranking score based on matches of the identified concepts. The appropriate schema is selected based on a highest weighted score of those assigned to the schemas. Based on its current popularity as shown in table 704, the schemas are further weighted with a weighted ranking score as shown in table 703. When 2 schemas have a similar basic score, the popularity weight added to differentiate them. In which case, even Schema 2 has a higher weighted ranking score than Schema 1, because Schema 3 already had a higher basic score, the popularity score for Schema 3 is added with 1 and the table 704 is updated accordingly.

FIG. 8 shows steps involved for relation discovery in the inter-sentential context. At step 802, a list of entities / NP, List A, is retrieved from the index 222 and iterate each of the items on the list. It starts from the first entry. At step 804, select a next entry $E_i$ from List A and retrieve a list of entities/NP, which is not within the same paragraph as in List B. At step 806, the intra-sentential results stored in the knowledge base 215 if the relation exists between $E_i$ and List B (reuse if exist). The Linked Data 103 is searched for $E_i$ and obtains all possible candidate relations. In step 808, candidate triple is removed if the subject/object do not match to List A or it falls below a similarity threshold. The matched results are stored on the knowledge base. The processes are repeated until the last entry in List A is processed.

While specific embodiments have been described and illustrated, it is understood that many changes, modifications, variations, and combinations thereof
could be made to the present invention without departing from the scope of the invention.
Claims

1. A system (100) for carrying out a relation extraction from a machine-readable document (190) having sentences, the system (100) comprising:

   a text preprocessing module (101) for lemmatizing the sentences into tokenized text and identifying paragraphs from the document (190);

   a coreference resolution module (122) configured to resolve all possible anaphors;

   an entity recognition module (124) for extracting entities from the sentence;

   an entity resolution and disambiguation module (126) adapted for resolving all ambiguous entities/noun phrases, acronyms and abbreviations;

   a relation extraction module (128) configured to operably carrying out a generic relation extraction and a semantic relations extraction for extracting triples (106) for a sentence based on the extracted entities, wherein a weighted ranking score is computed for each triple for a sentence for selecting and storing an most appropriate triple.

2. The system (100) according to claim 1, wherein each triple (106) is computed with a basic ranking score by matching the identified concepts with the respective verb schemas, and weighted with their respective recorded popularity scores to obtain the weighted ranking score for each triple (106), the highest ranking score of which is selected as the most appropriate triple (106).

3. The system (100) according to claim 1, wherein the relation extraction module (128) carries out the generic relation extraction for resolving entities in an intra-
sentential context, the generic relation extraction operationally extracts noun phrases
and verb from the sentence, matches predefined patterns (104) from the linguistic
resources (105), and searches the linguistic resources again with a Linked Data (103)
for possible properties to generate corresponding relation of the noun phrases and verb
are referring to in that sentence.

4. The system (100) according to claim 1, wherein the relation extraction module
(128) carries out the generic relation extraction for resolving entities from an inter-
sentential context, the relation extraction module (128) operationally renders a list of
entities of noun phrases from a repository and another list of entities of noun phrases
that are not within the same paragraph and the connecting verbs, extracts and matches
both lists for schemas based on the verb identified from the linguistic resources (105),
identifies possible properties to generate corresponding relations from a knowledge base
(215), and searches the linguistic resources (105) for possible triples based on the
properties found.

5. The system (100) according to claim 1, wherein the semantic relations extraction
for entities in an intra-sentential context matches predefined patterns to extract and
match schemas based on the identified verb, and searches entities or noun phrases
through a linked database (103).

6. The system (100) according to claim 1, wherein the semantic relations extraction
for entities in an inter-sentential context retrieves a list of entities and noun phrases
from a repository and another list of entities, noun phrases that are not within the same paragraph and the connecting verbs in order to extract and match schemas based on the verb identified from the linguistic resources (105).

7. A method of carrying out a relation extraction from a machine-readable document (190) having sentences, the method comprising:
   - lemmatizing texts of the sentences into tokenized text and identifying paragraphs from the document (190);
   - resolving all possible anaphors through a coreference resolution module (122);
   - extracting (201) entities from the sentences through entity recognition module (124) for extracting entities from the sentence;
   - resolving (203) all ambiguous entities / noun phrases, acronyms and abbreviations;
   - extracting a generic relation and a semantic relations (210, 215) for extracting triples (106) for a sentence based on the extracted entities, wherein a weighted ranking score is computed for each triple (106) for a sentence for selecting and storing an most appropriate triple (106).

8. The method according to claim 7, wherein the relation extraction further comprising:
   - computing a basic ranking score by matching the identified concepts to the respective verb schemas;
   - retrieving a popularity score for each of the schemas;
computing a popularity weighted score for each schema to obtain weighted ranking score for each schema;
selecting a schema with the highest weighted ranking score;
generating the relation triples based on the selected schema.

9. The method according to claim 7, wherein
the extracting the generic relation among entities in an intra-sentential context further comprising:
extracting noun phrases and verb from linguistic resources to match the sentence with predefined patterns (104); and
searching the linguistic resources and a Linked Data for possible properties to generate corresponding relation the sentence refers to; and
the extracting the semantic relations extraction for entities in an intra-sentential context further comprising:
matching predefined patterns to extract and match schemas based on the identified verb, and
searching entities or noun phrases through a linked database (103).

10. The method according to claim 7, wherein
the generic relation extraction among entities in an inter-sentential context further comprising:
retrieving a list of entities and noun phrases from a repository and another list of entities for noun phrases that are not within the same paragraph and the connecting verbs;

extracting and matching schemas based on the verb identified from the linguistic resources (105) to identify relations from a knowledge base (215); and searching the linguistic resources (105) for possible triples based on the properties found; and

the semantic relations extraction for entities in an inter-sentential context further comprising:

retrieving a list of entities and noun phrases from a repository and another list of entities for noun phrases that are not within the same paragraph and the connecting verbs; and

extracting and matching schemas based on the verb identified from the linguistic resources.
Sentence: (…) most European countries especially England, Germany and France (…)

Step (1): Identify entity / noun phrase
E={England, Germany, France}

Step (2): Identify verb
no verb found

Step (3): Match patterns
yes & relations will be generated as
follows:
HYPONYM(England, European country); ← lemmatized form
HYPONYM(Germany, European country);
HYPONYM(France, European country)

Patterns:
NP, especially {NP,}*{or|and} NP

(Hyponym)

FIG. 3A
(... most European countries especially England, Germany and France

Patterns:
NP, especially \{NP, \}_j

Step (4): Search linguistic resources

Noun
- domain_length
- member
- related
- instance
- function
Sentence: Samsung released Galaxy Mini, another smartphone for Android fans

Step (1): Identify entity / noun phrase
E = {Samsung, Galaxy Mini, etc.}

Step (2): Identify verb
verb = release

Step (3): Match patterns
Found and trigger the corresponding rule
Return triples: (Samsung, release, Galaxy Mini)

Trigger rule:
(NP1, VP1, NP2)
(NP1, VP1, NP3)
Sentence:
The Bugatti Veyron EB 16.4 is a mid-engined grand touring car by the Volkswagen Group in France

**Step (1): Identify entity / noun phrase**
E = {Bugatti Veyron, etc.}

**Step (2): Identify verb**
no verb found

**Step (3): Match patterns**
no match found

**Step (4): Search “Bugatti” from LOD**
found: dbpedia-owl:class dbpedia:Sports_car → rdfs:label “Sports car”

**isType** (Bugatti Veyron, Sports car)

*Predefined property will refer to a specified relation*

FIG. 5
Sentence: John bought Mary a Ferrari.

Step (1): Identify entity / noun phrase
E = {John, Mary, Ferrari, etc.}

Step (2): Identify verb
verb = buy

Step (3): Extract & match schema
from linguistic resources
Schema 2 found

Step (4): Check selectional constraints

* V denotes the root verb

Agent V Theme (1)
Agent V Beneficiary Theme (2)
Agent V Theme Source (3)
**FIG. 7**

**Schema List**

<table>
<thead>
<tr>
<th>Schema 1</th>
<th>Schema 2</th>
<th>Schema 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>animate&lt;-&gt;(agent)-&gt;(travel)</td>
<td>animate&lt;-&gt;(agent)-&gt;(travel)</td>
<td>animate&lt;-&gt;(location)</td>
</tr>
<tr>
<td>(dest)-&gt;(location)</td>
<td>(inst)-&gt;(entity)</td>
<td></td>
</tr>
</tbody>
</table>

KB

707

S1: Schema 1
S2: Schema 2
S3: Schema 3

**Ranking Score**

<table>
<thead>
<tr>
<th>Schema</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.67</td>
</tr>
<tr>
<td>S2</td>
<td>0.67</td>
</tr>
<tr>
<td>S3</td>
<td>3.75</td>
</tr>
</tbody>
</table>

**Popularities**

<table>
<thead>
<tr>
<th>Schema</th>
<th>Popularities</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1</td>
</tr>
<tr>
<td>S2</td>
<td>3</td>
</tr>
<tr>
<td>S3</td>
<td>5</td>
</tr>
</tbody>
</table>

**Sentence**

John travels to London by bus.

**Extract triples:**

- (John, travel, London)

- This will assist when 2 schemas have the similar basic score - add the higher score.
### A. CLASSIFICATION OF SUBJECT MATTER

**INV.** G06F17/27

**ADD.** G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

### B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

### Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

### C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>WO 2010/132790 AI (COLLEXIS HOLDINGS INC [US]; SCHMIDT MARTIN [DE]; DIWERSY MARIO [DE]) 18 November 2010 (2010-11-18) abstract paragraphs [0002] - [0034] figure 2</td>
<td>1, 2, 7, 8</td>
</tr>
<tr>
<td>X</td>
<td>WO 2009/052277 AI (EVR INC [US]; LIANG JISHENG [US]; KOPERSKI KRZYSZTOF [US]; DHL LIONNA NA) 23 April 2009 (2009-04-23) abstract page 6, line 12 - page 16, line 9 figures 3-9</td>
<td>1, 7</td>
</tr>
</tbody>
</table>

Further documents are listed in the continuation of Box C.

See patent family annex.

* Special categories of cited documents:
  - "A" document defining the general state of the art which is not considered to be of particular relevance
  - "E" earlier application or patent but published on or after the international filing date
  - "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
  - "O" document referring to an oral disclosure, use, exhibition or other means
  - "P" document published prior to the international filing date but later than the priority date claimed
  - "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
  - "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
  - "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
  - "Z" document member of the same patent family

Date of the actual completion of the international search

26 February 2015

Date of mailing of the international search report

05/03/2015

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, 340-2040, 340-3016
Fax: (+31-70) 340-2040

Authorized officer

Woods, Justin
### DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>
**INTERNATIONAL SEARCH REPORT**

**Box No. II** Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. [ ] Claims Nos.:
   - because they relate to subject matter not required to be searched by this Authority, namely:

2. [X] Claims Nos.: 4, 6, 10
   - because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
     - see FURTHER INFORMATION sheet PCT/ISA/21Q

3. [ ] Claims Nos.:
   - because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

**Box No. III** Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. [ ] As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. [ ] As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. [ ] As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. [ ] No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.: 

**Remark on Protest**

- The additional search fees were accompanied by the applicant’s protest and, where applicable, the payment of a protest fee.

- The additional search fees were accompanied by the applicant’s protest but the applicable protest fee was not paid within the time limit specified in the invitation.

- No protest accompanied the payment of additional search fees.
Claim 10 begins by referring to an inter-sentential context, which is absent from claim 7. Even if it were assumed that such an inter-sentential context were referred to in claim 7, on which claim 10 depends, the remainder of the claim is so unclear as to render a search impossible.

A list of entities that are not within the same paragraph is retrieved, it is thus assumed that a separate list of all entities in the document apart from the current paragraph is retrieved, along with their connecting verbs. Then schemas are to be extracted based on the verb identified. Even if the person skilled in the art assumes that "verbs" in the plural is meant and such extracting can be carried out, there is no indication as to what purpose said extracting on and matching fulfill and how it relates to the further retrieval, extraction on and matching which is performed afterwards. As, contrary to the requirements of Article 6 PCT, the person skilled in the art cannot derive from claim, even in the light of the description, how the features interact with each other and what the purpose is of seemingly similar retrieval, extraction on and matching steps being repeated, no meaningful search is possible w.r.t. claim 10.

Similar reasoning applies to claims 4 and 6, which also claim rendering a list of entities not from the same paragraph, without providing an indication as to how such lists are to be used, or their relationship to claim 1, on which the claims depend or, indeed to the invention as a whole. That is, contrary to the requirements of Article 6 PCT, claims 4 and 6 do not define the matter for which protection is sought, even in the light of the description, rendering a meaningful search in their respect impossible.

The applicant's attention is drawn to the fact that claims referring to invention in respect of which no international search report has been established need not be the subject of an international preliminary examination. The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the applicant proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination on the EPO (see EPO Guidelines C-IV, 7.2), should the problems which led to the Article 17(2) declaration be overcome.

Claims Nos.: 4, 6, 10

Continuation of Box II.2
<table>
<thead>
<tr>
<th>Patent document cited in search report</th>
<th>Publication date</th>
<th>Patent family member(s)</th>
<th>Publication date</th>
</tr>
</thead>
<tbody>
<tr>
<td>WO 2010132790</td>
<td>18-11-2010</td>
<td>CN 102576355 A</td>
<td>11-07-2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP 2430568 A</td>
<td>21-03-2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 2012527058 A</td>
<td>01-11-2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2012158400 A</td>
<td>21-06-2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 2010132790 A</td>
<td>18-11-2010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 2703007 A</td>
<td>23-04-2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP 2206057 A</td>
<td>14-07-2010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2009144609 A</td>
<td>04-06-2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2014142922 A</td>
<td>22-05-2014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 2009052277 A</td>
<td>23-04-2009</td>
</tr>
</tbody>
</table>